IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 21, 2019, accepted August 7, 2019, date of publication August 13, 2019, date of current version August 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2935006

Multi-Scale Visual Attention Deep Convolutional Neural Network for Multi-Focus Image Fusion

RUI LAI[1], YONGXUE LI[1], JUNTAO GUAN[1], AND AI XIONG[2]
[1]Department of Microelectronics, Xidian University, Xi'an 710071, China
[2]School of Control Engineering, Chengdu University of Information Technology, Chengdu 610103, China

Corresponding author: Rui Lai (rlai@ mail.xidian.edu.cn)

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 61674120, Grant 61571338, Grant U1709218, and Grant 61672131, in part by the Fundamental Research Funds for the Central Universities of China under Grant JBG161113 and Grant 300102328110, and in part by the Key Research and Development Plan of Shaanxi Province under Grant 2017ZDCXL-GY-05-01.

ABSTRACT To realize the multi-focus image fusion task, an end-to-end deep convolutional neural network (DCNN) model that produces the final fused image directly from the source images is presented in this paper. In order to promote the fusion precision, the innovative multi-focus fusion DCNN introduces a multi-scale feature extraction (MFE) unit to collect more complementary features from different spatial scales and fuse them to excavate more spatial information. Moreover, a visual attention unit is designed to help the network locate the focused region more accurately and pick more useful features for perfectly splicing the details in the fusion process. Experimental results illustrate that the proposed method is superior to several existing multi-focus image fusion methods in both of the subjective visual effects and objective quality metrics.

INDEX TERMS Image fusion, multi-focus, convolution neural network, multi-scale.

I. INTRODUCTION

As is known, it is often difficult to use a camera, such as a digital single-lens reflex camera, to get a full focus picture in which all objects or scenes are in focus at one shot. In general, only the objects in the focal distance of optical lens can retain sharp appearance information, and that beyond the focal distance will be blurred. The common method getting full focus images is the so called multi-focus image fusion that fuses multiple same scene observations taken with different focus distance into one full focus image, whose key thought is to preserve the information involved in each observation as complete as possible.

In recent years, a variety of image fusion methods were put forward in succession, which can be roughly divided into two categories, namely transform domain methods and spatial domain methods [1]. As for the transform domain methods, most of them are based on multi-scale transform (MST) theories and basically developed from the Laplacian pyramid method [2]–[5]. For example, the NSCT-based method decomposes the image using contourlet or shearlet transform without subsampling [6], which outperforms the traditional discrete wavelet transform (DWT) based image fusion methods. The single-scale transform (SST) methods transform the source image patches into the sparse domain according to advanced sparse representation (SR) theories [7]. Some hybrid methods such as NSCT-SR-based approach [8] combine the advantages of the MST and SST method under a unified framework. In addition, the Discrete Cosine Transform (DCT) based method achieves satisfactory results with low energy consumption [9].

In spatial domain, block based fusion strategy is often adopted [10], [11]. Since the size of the block has a great impact on the quality of the fusion, some improved methods are put forward to solve this problem [12]–[14]. Beyond that, image segmentation based fusion methods that rely on the accuracy of segmentation are presented in [15], [16]. In recent years, novel gradient information based methods [17], [18] have achieved impressive fusion results.

For either the transform domain or the spatial domain methods, the activity measurement and fusion rules are the two key factors that affect the fusion performance. In most existing algorithms, even these two factors are artificially well-designed and the intrinsic schemes become to be increasingly complex, the existing efforts are still insufficient to take all the influence impacts into account.

The associate editor coordinating the review of this article and approving it for publication was Martin Gonzalez-Rodriguez.

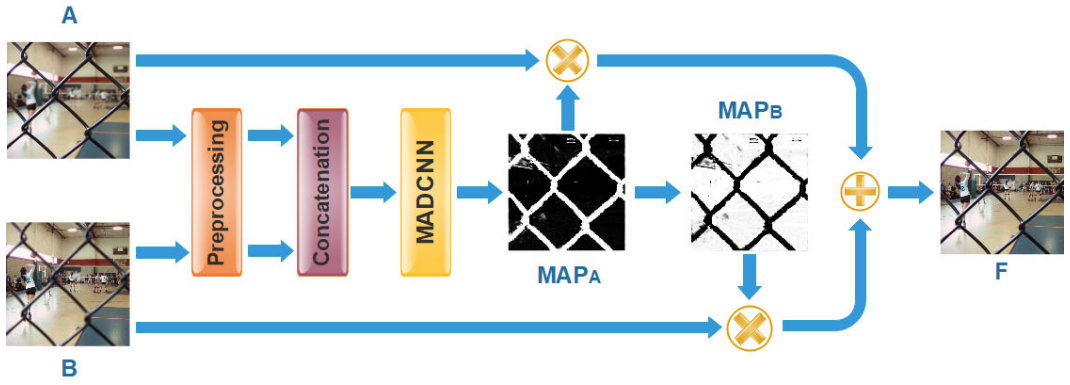

FIGURE 1. The schematic diagram of the proposed method.

Besides, the deep learning based methods achieved superior performance in several image processing and computer vision applications [19]–[21]. In view of this, Liu Yu generated the activity level measurement and fusion rule through deep learning with a convolutional neural network (CNN) model, which successfully gets rid of the limitation of manmade rules [22]. H. Tang employed pixel-wise convolution to recognize the focused and defocused pixels from neighborhoods [23]. M. Amin-Naji proposed the vote based methods with ensemble learning strategy for better fusion results [24], [25]. The fully-connected layer utilized in these works consumes large storage resources. Therefore, some works tend to employ the convolution layer instead the fully-connected operation to reduce the model complexity [26], [27]. This transformative method does not need to adjust the parameters under different input conditions, so that it has wider applicability for the changing scenarios. However, the patch-based strategy consumes more computation and is still struggling to process the boundary between the focused and defocused region more precisely.

In order to solve the abovementioned problems, this paper proposed a multi-focus image fusion method based on an innovative end-to-end multi-scale visual attention deep convolutional neural network (MADCNN) model. The main contributions of this paper can be summarized as follows:

(1) Propose an end-to-end model to generate the final fused image directly from multi-focus source images. Different from the existing methods based on classification CNN model, the proposed model avoids having to divide the source images into overlapping blocks, which processes the boundary more naturally as well as reduces the unnecessary computational load.

(2) Introduce a multi-scale feature extraction (MFE) unit into the proposed model to excavate more complementary features from different spatial scales, which conducts to preserve more source information in the fusion results.

(3) Present a visual attention unit to help the proposed model locate the focused region and pick out more useful features for high precision fusion.

The rest of this paper is organized as follows. Section 2 describes the proposed method based on an end-to-end MADCNN model in detail. Section 3 discusses the impact of different network structure on the fusion performance, and then proves the validity and superiority of the proposed method from the subjective visual effect and objective index. Finally, some concluding remarks are made in Section 4.

II. THE PROPOSED METHOD

A. METHOD OVERVIEW

In general, the fusion process of multi-focus images can be equivalent to successive fusion of dual-focus images. Therefore, the algorithm proposed in this paper only takes into account the case of two dual-focus images fusion. The schematic diagram of the presented method is shown in Figure 1. The process of the method is mainly divided into three key steps: data preparation, focus map generation and fusion operation.

Step 1: In the data preparation procedure, two source images A and B are firstly transformed into gray scale images and then normalized by mean and variance. This step makes the data be sparse and facilitate for training. Thereafter, the two preprocessed images are concatenated as a single input tensor and prepared to feed into MADCNN model.

Step 2: In focus map generation stage, different from the patch-based methods, the focus map will be directly estimated according to the input tensor prepared by step 1. This strategy avoids computing the patches recursively and saves most of computing resources. It is worthy to note that the MADCNN model is only designed to generate the focus map MAP_A. The focus map of source image B is directly calculated by

$$MAP_B = 1 - MAP_A \tag{1}$$

Step 3: In the fusion stage, the focus map produced by step 2 will be utilized to mark out the focus region and weight the two source images for generating the fused image F.

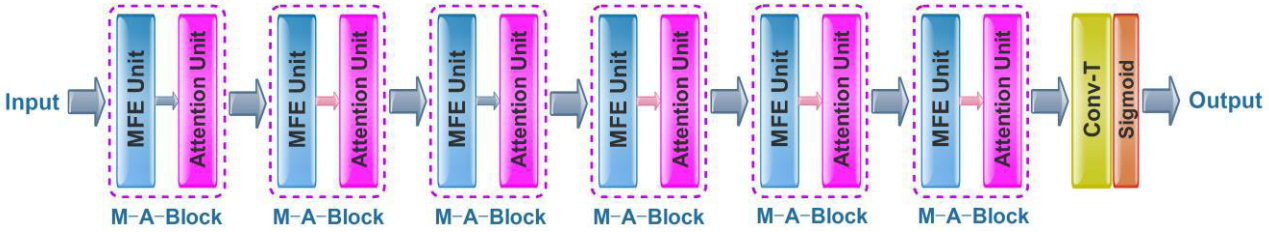

FIGURE 2. The structure of the proposed MADCNN model.

This stage can be mathematically presented by

$$F_c = A_c \times MAP_A + B_c \times MAP_B \tag{2}$$

where the subscript c denotes a certain color channel corresponding to R, G or B.

B. NETWORK DESIGN

Existing learning based multi-focus image fusion methods basically utilize the classification network framework that has been claimed to cause the inaccurate fusion result around the boundaries between focused and defocused region [28]–[32]. In view of this, an innovative MADCNN model only containing convolution layers and activation layers is proposed to directly produce the more precise fusion. Moreover, the proposed MADCNN model has fewer parameters and lower computation cost, which leads the whole fusion process to be more convenient and faster.

As shown in Figure 2, the proposed MADCNN model introduces 6 stacked MFE-Attention blocks (M-A-Blocks) to extract fusion related features, and then cascade a terminal convolution layer (Conv-T) consists of a single 3×3 sized convolution filter activated by sigmoidal function to fuse channel wise information and output the estimated focus map. From the structure diagram shown in Figure 3, we can see that the M-A-Block mainly composes of a multi-scale feature extraction (MFE) unit and an attention unit. Among them, the MFE unit employs standard convolution kernels and dilated convolution kernels to extract complementary features from different spatial scales. As for the dilated convolution, it can cheaply expand the receptive field without increasing the kernel size, which is especially effective while stacking multiple dilated convolutions one after another. Standard convolution operation and dilated convolution operation can be uniformly formulated as

$$O(j,k) = \sum_{m=-r}^{r} \sum_{n=-r}^{r} W(m,n) \cdot I(j + m \cdot d, k + n \cdot d) \tag{3}$$

where d stands for dilation rate. When d is set to 1, the formula represents a standard convolution operation. The structural parameters of all M-A-Blocks are listed in Table 1.

Although the MFE unit provides abundant multi-scale features, but not all of these features contribute to a high quality fusion results. To address this issue, we insert the attention unit to refine the output features of MFE unit and further locate the focused region. As shown in Figure 3, the attention unit involves base branch and mask branch, whose outputs are respectively marked as base feature and feature mask. Base branch is only responsible for transmitting feature maps provided by MEF units. The upper branch of the attention unit (named Mask Branch), whose cascaded convolution layers are configured as Table 1, is introduced to generate feature mask to hide the defocused background and preserve the focused area. The final focus map is generated by multiplying the base feature map with its corresponding feature mask. Furthermore, the focus attention mechanism can be mathematically represented as

$$x_M^c = B^c(x) \cdot M^c(x) \tag{4}$$

where $B(x)$, $M(x)$ and x_M respectively stand for the base feature map, feature mask and the output of attention unit. c denotes the index of the color channel.

TABLE 1. The structure parameters of Mfe-attention-block.

Unit	Layer	Filter size	Dilation rate	Output Channels
MFE-Unit	conv-1	3×3	1	8
	conv-2	3×3	2	24
Attention-Unit	conv-3	3×3	1	32
	conv-4	3×3	1	32

C. LOSS FUNCTION

At present, most literatures are concerning about the fine tuning of network structure in specific application fields. Although the loss layer plays an important role in generating images with desirable visual effects, the existing algorithms basically use the mean square error (MSE) function.

However, a widely accepted view is that the MSE function cannot reflect the human perception of image quality well. In other words, it still struggles to capture the complex characteristics of the human visual system. Therefore, many improved reference and non-reference based loss function have been put forward [33]. In this paper, we introduce a reference based structural similarity index (SSIM) to construct the loss function, which evaluates images according to the changes in local structure of image.

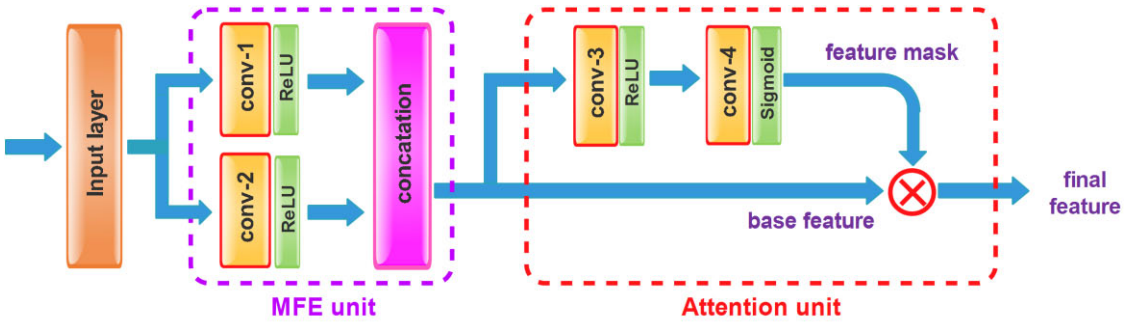

FIGURE 3. The structure of the MFE-Attention-Block (M-A-Block).

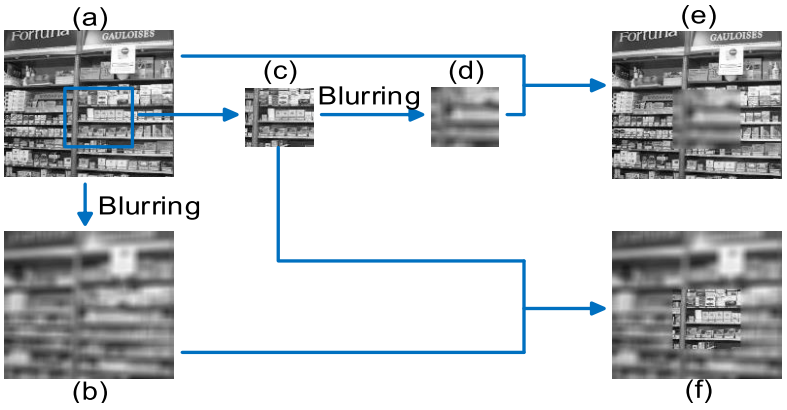

FIGURE 4. The sketch map of training sample generation.

The SSIM is calculated by the following formula

$$SSIM\ (x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{5}$$

where μ_x and μ_y stand for the mean value of x and y, σ_x^2 and σ_y^2 indicate the variance, σ_{xy} denotes the covariance. The constants c_1 and c_2 are used to stabilize the division with the weak denominator.

The range of SSIM is from 0 to 1. If the two images are exactly the same, SSIM approaches to 1. On the contrary, SSIM is close to 0. Thus, the loss function is defined as

$$\ell_{ssim} = 1 - SSIM(x, y) \tag{6}$$

D. TRAINING

In image fusion applications, the most common problem is that there are no ground-truth results, which makes the training of end-to-end network hardly to be achieved. In order to solve this problem, this paper proposes an artificial simulation method for generating approximate multi-focus images. We choose about 45000 images with relatively rich details from ILSVRC 2015 as the original dataset, which are then uniformly sliced into 128×128 sized image blocks. Followed that, the training data is generated from the abovementioned image blocks according to the steps in following example.

As for the preparation of the training data, we randomly select a sliced image block and convert it to a grayscale image shown in Fig.4 (a), from which a D × D sized patch shown in Fig.4 (c) is cropped. Thereafter, Gaussian filtering is performed on Fig.4 (c) and generate a defocused patch shown in Fig.4 (d), which is then used to substitute the cropped patch and produce the synthetic dual-focus image shown in Fig.4 (e). Meanwhile, the corresponding complementary dual-focus image shown in Fig.4 f) is produced by substituting the patch in Gaussian blurred Fig.4 (b) with the cropped patch shown in Fig4. (c). At this point, Fig.4 (e) and (f) form a pair of dual-focus images, and Fig.4 (a) is the corresponding ground-truth fusion result. Repeating the above process on all of the sliced samples, we can build up a complete training set.

It is worth noting that the size d and standard deviation σ of Gaussian blurring filter as well as the size D of the cropped patch mentioned above are randomly generated in a specific range. The generating formula for each parameter are defined as

$$\begin{aligned} d &= 11 + 2 \cdot i, \quad i = rand\{1, 2, 3\} \\ \sigma &= 4 + 2 \cdot j, \quad j = rand\{1, 2, 3, 4, 5\} \\ D &= 20 + 10 \cdot k, \quad k = rand\{1, 2, 3, 4, 5\} \end{aligned} \tag{7}$$

The training set used in this paper contains 300,000 samples, in which ten percent are used as validation data.

FIGURE 5. Exampleof the source image pairs from the "Lytro" multi-focus dataset.

The popular deep learning framework Keras is utilized to train our proposed MADCNN model. The Adam optimizer is applied to minimize the loss function, and the learning rate is set to 0.00005 during the 50 epochs training procedure. It should be noted that the learning rate is equal for all layers. Besides, the batch size is set to 64. The training phase only takes about 30 hours on the workstation with Intel Xeon E1620 CPU and single NVidia 1080Ti GPU. The fusion process take about 181ms for 520 × 520 sized test pairs.

III. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETTINGS

In order to verify the effectiveness and superiority of the proposed MADCNN based fusion method, we select 20 pairs of multi-focus images from a 520 × 520 sized publicly available set "Lytro" [34] as the testing data, part of which are shown in Figure 5. In the following experiments, the proposed method will be respectively compared with the transform domain methods NSCT [6], SR [7], the spatial domain methods MWG [18], DSIFT [22], and deep learning based CNN [33] and ECNN [24]. The parameters of each method are set strictly according to the recommended values mentioned in the relevant literatures for the optimal performance.

For a more comprehensive and objective comparison, we assess the performance of the fusion methods from various aspects with the following seven quantitative metrics:

(1) Nonlinear correlation information entropy Q_{NCIE} [35]. (2) Phase congruency Q_P [36]. (3) SSIM-based metric Q_Y [37]. (4) Human perception-based metric Q_{CB} [38]. (5) Mutual information Q_{MI} [39]. (6) Tsallis Entropy-based metric Q_{TE} [40]. (7) Ratio of spatial frequency error *rSFe* [41]. For all of the above fusion metrics, larger value means for higher performance.

B. DISCUSSION ON NETWORK STRUCTURE

As mentioned in Section II (B), specially designed MFE unit and Attention unit are introduced in the proposed MADCNN model. The roles of these two units will be discussed and verified thoroughly in this section. For ease of comparison, we show network models with different structure in Figure 6. By comparison, we can find that all the models share a same final convolutional layer called Conv-T, which is then cascaded with a sigmoid activation layer. Meanwhile, we can also observe the difference of these four models lies in the former convolution layers. The model (A) is simply stacked by the common convolution layer with 3×3 sized convolution kernels called Conv-C, in which the number of output channels for each convolution layer is 32. The model (B) is made up of 6 cascaded MFE units, which specializes in excavating the multi-scale spatial information. By inserting the Attention unit followed each convolution layer in model (A), we construct the model (C) known for salient feature separation. The model (D) is the proposed MADCNN, which combines MFE unit and Attention unit to build the M-A-block, which conducts to excavate abundant features in focus. The configuration of MFE unit and Attention unit is detailed fully in Figure 3 and Table 1. In order to verify the role of MFE unit and Attention unit, we train these four models with the same training set generated in Section II-(D), and then carry out the multi-focus fusion with each of the trained model and assess the fusion results with both of objective metrics and subjective visual effect.

To demonstrate the effectiveness of the attention unit, we firstly illustrate the visual attention mechanism for separating the features located in focus region intuitively. In Figure 7, the second column shows the base feature maps extracted from the source images, whose corresponding feature masks shown in the third column is generated from the mask branch of attention unit in Figure 3. By weighting the base feature map with the feature mask, the attention unit outputs the masked feature maps shown in the fourth column.

By observing the base feature maps in the second column of Figure 7, we can easily find that the features located in focused and unfocused region are mixed. The feature masks suppress the features located in the unfocused region by assigning smaller weights that seems to be darker. As a result, the masked feature maps mainly give prominence to the features located in the focused region. These four examples

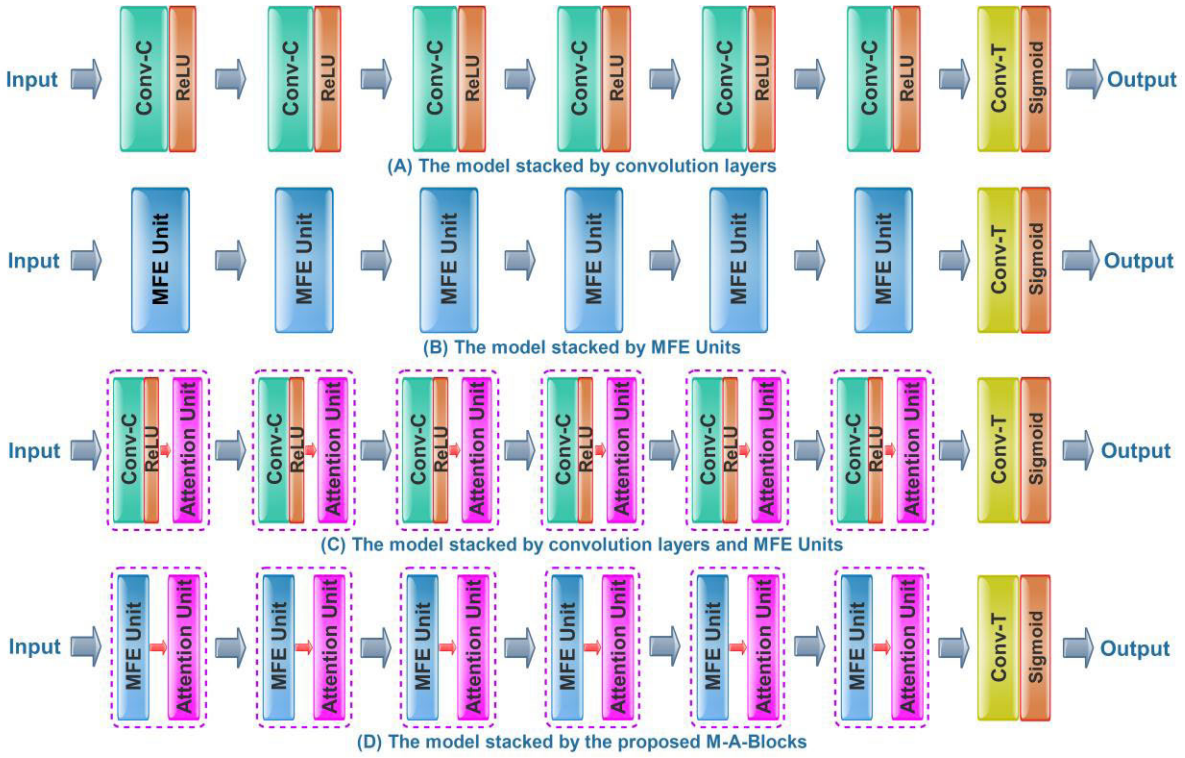

FIGURE 6. Comparisons for the architecture of four different network models.

TABLE 2. Objective assessment of different models.

Method	Q_{NCIE}	Q_P	Q_Y	Q_{CB}
Figure 6 (A)	0.8425	0.8450	0.9837	0.7985
Figure 6 (B)	0.8430	0.8441	0.9848	0.7998
Figure 6 (C)	0.8432	0.8462	**0.9866**	0.8054
Figure 6 (D)	**0.8437**	**0.8474**	0.9861	**0.8055**

thoroughly illustrate the role of the attention unit, which helps the network to concentrate on the sharp and clear contents in the focused region. How much dose the attention unit contribute to the performance promotion of the proposed MADCNN model? The answer lies in Table 2, which contains the average scores for the fusion results of 20 multi-focus image pairs assessed by four fusion metrics. From the objective assessment for different models, we can find that the additional attention unit helps model (c) achieve higher scores than model (a). This result further proves the effectiveness of the attention unit.

To further demonstrate the effectiveness of MFE unit, we continue to analyze the assessment results listed in Table 2. From the comparison between the second row and first row, we can easily find that the fusion performance of MFE unit is better than that of pure convolution structure. Although the attention unit and MFE unit are individually beneficial for high performance fusion, their combination so-called M-A-block can further improve the performance or not is still a problem need to be answered. By comparing the values in the last row with that in the former three rows, respectively, we can see that most scores of the combined model (D) are higher than that of model (B) or model (C). The reason lies in that the MFE unit provides much richer features and the attention unit further sieves out the useful features from the output of MFE unit to promote the fusion performance.

C. PERFORMANCE COMPARISONS AND ANALYSIS

In this section, the fusion results of three pairs of testing images are respectively exhibited in Figure 8, 9 and 10 to assess the performance of different fusion methods from the perceptive of visual quality. Furthermore, quantitative performance comparison will be implemented.

In Figure 8, we can see clearly that the fusion results of NSCT and SR make the focus and defocus scene be ambiguous. The difference image between the fusion results and source images show residues of varying degrees in either focus or defocus region. The rest five methods perform better in distinguishing the focus from defocus scene. However, the MWG leads to noticeable "halo" artifacts near the boundaries between the focused and defocused region, which can be observed clearly in the difference images. Similar phenomena also exist in the result of DSIFT, CNN and ECNN method. Moreover, CNN method fails to handle the richly detailed boundary marked out with red box. The DSIFT, CNN and ECNN fail to distinguish the focused fingers from defocused gap between fingers that is shown in the yellow box.

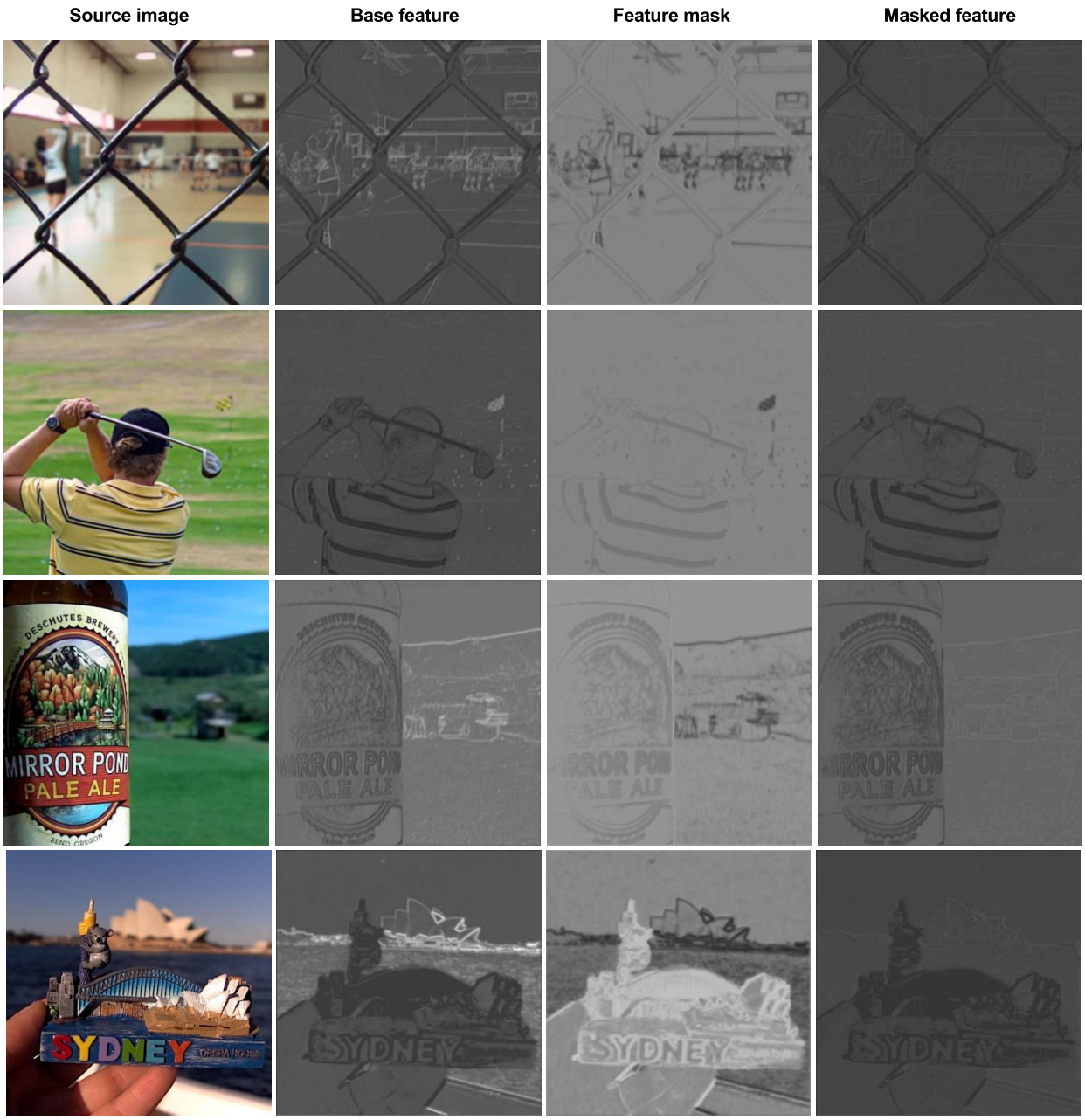

FIGURE 7. Illustration of the attention mechanism for the focus region.

In contrast, the proposed MADCNN method can precisely preserve the boundary between focused and defocused region as well as handle the richly detailed region efficiently.

In Figure 9, the fusion results of NSCT, SR and MWG are not satisfied, whose corresponding difference images presents plenty of misplaced residuals. The above drawbacks are obviously ameliorated in the following DSIFT, CNN, ECNN and MADCNN method. However, the CNN method processes the boundaries unsatisfactorily near the doll's hat, face and arm respectively marked out by yellow, red and blue box. The DSIFT and ECNN perform better than CNN method, but they fail near the hat marked out by yellow box. The proposed MADCNN provides outstanding fusion result and yields clear boundary without visible artifacts.

Figure 10 is employed to highlight the detail preservation capability of the different fusion methods. By comparing, we can observe significant residuals existed in the difference images belonging to the former two methods, which reveal their insufficient performance. The MWG, DSIFT and ECNN could not guarantee the boundary along the notebook to be continuous and complete, which can be observed in the red box. The fusion results of CNN and ECNN methods loss the detail of binding wire as well as produce residuals near the boundary of notebook, which are respectively indicated by

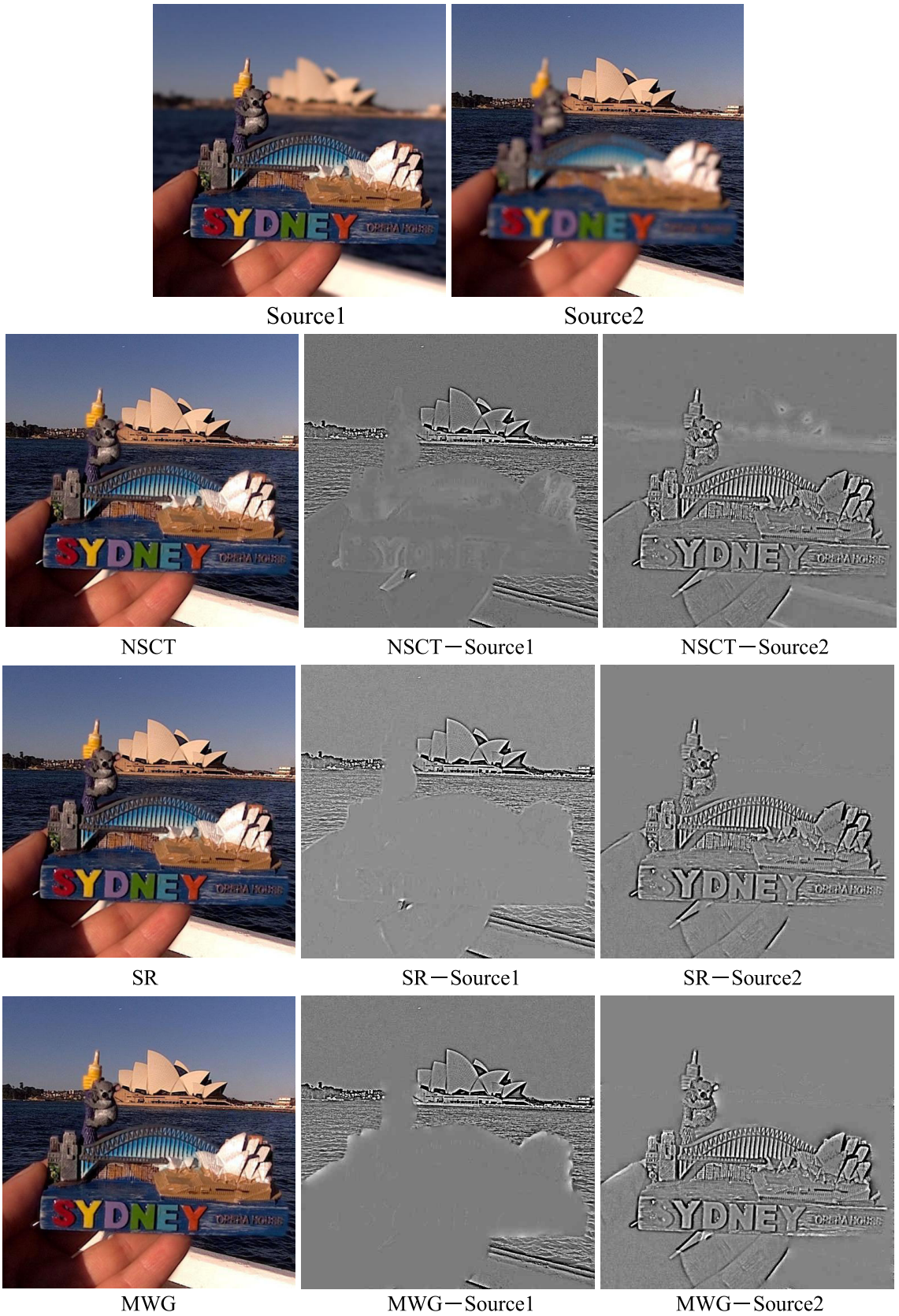

FIGURE 8. Fusion results of the 1st image pair in Fig. 5 using different methods. The first column is the fused images obtained by NSCT, SR, MWG, DSIFT, CNN, ECNN and MADCNN. The other two columns are respectively the difference between the fused image and two source images.

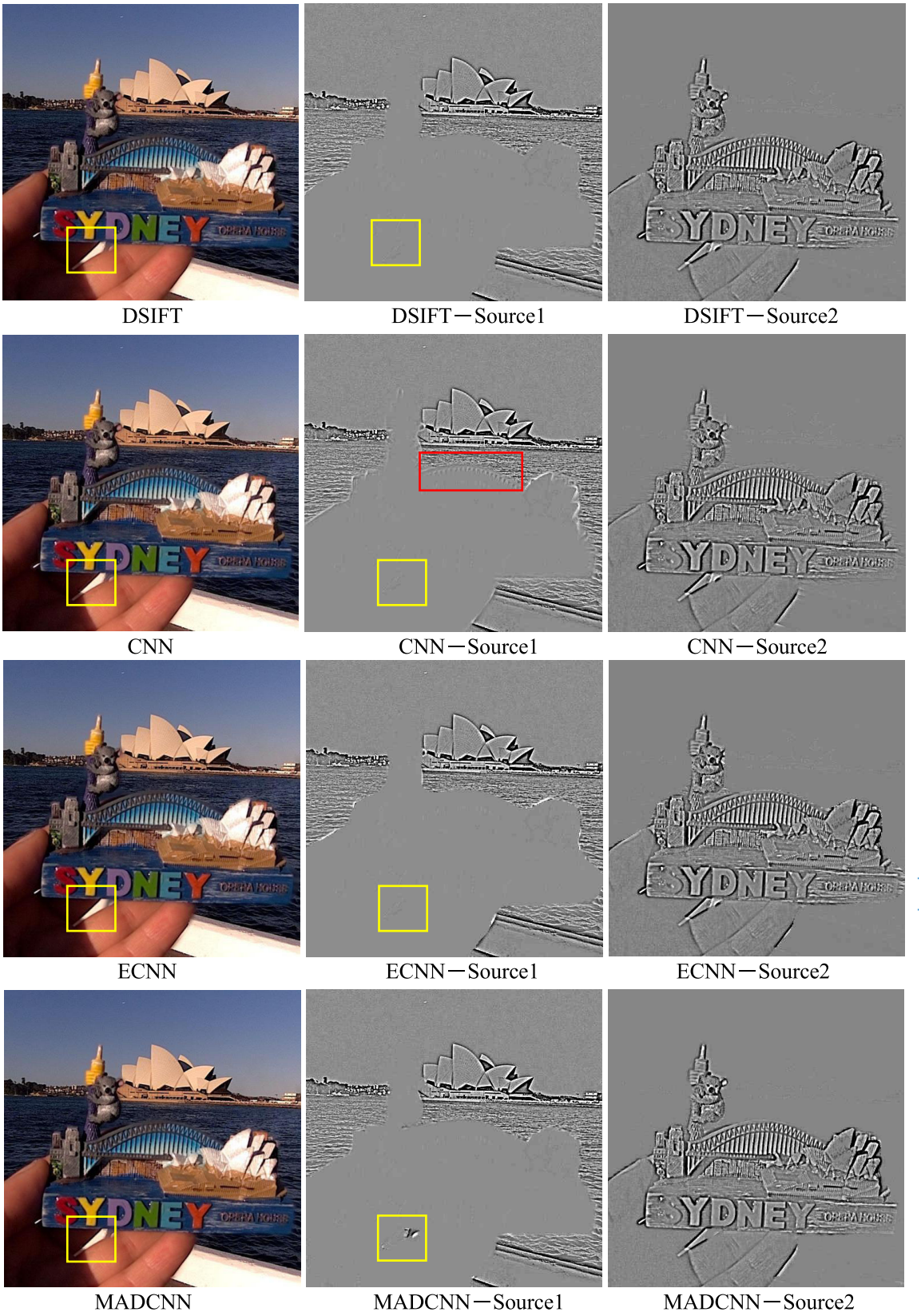

FIGURE 8. *(Continued.)* **Fusion results of the 1st image pair in Fig. 5 using different methods. The first column is the fused images obtained by NSCT, SR, MWG, DSIFT, CNN, ECNN and MADCNN. The other two columns are respectively the difference between the fused image and two source images.**

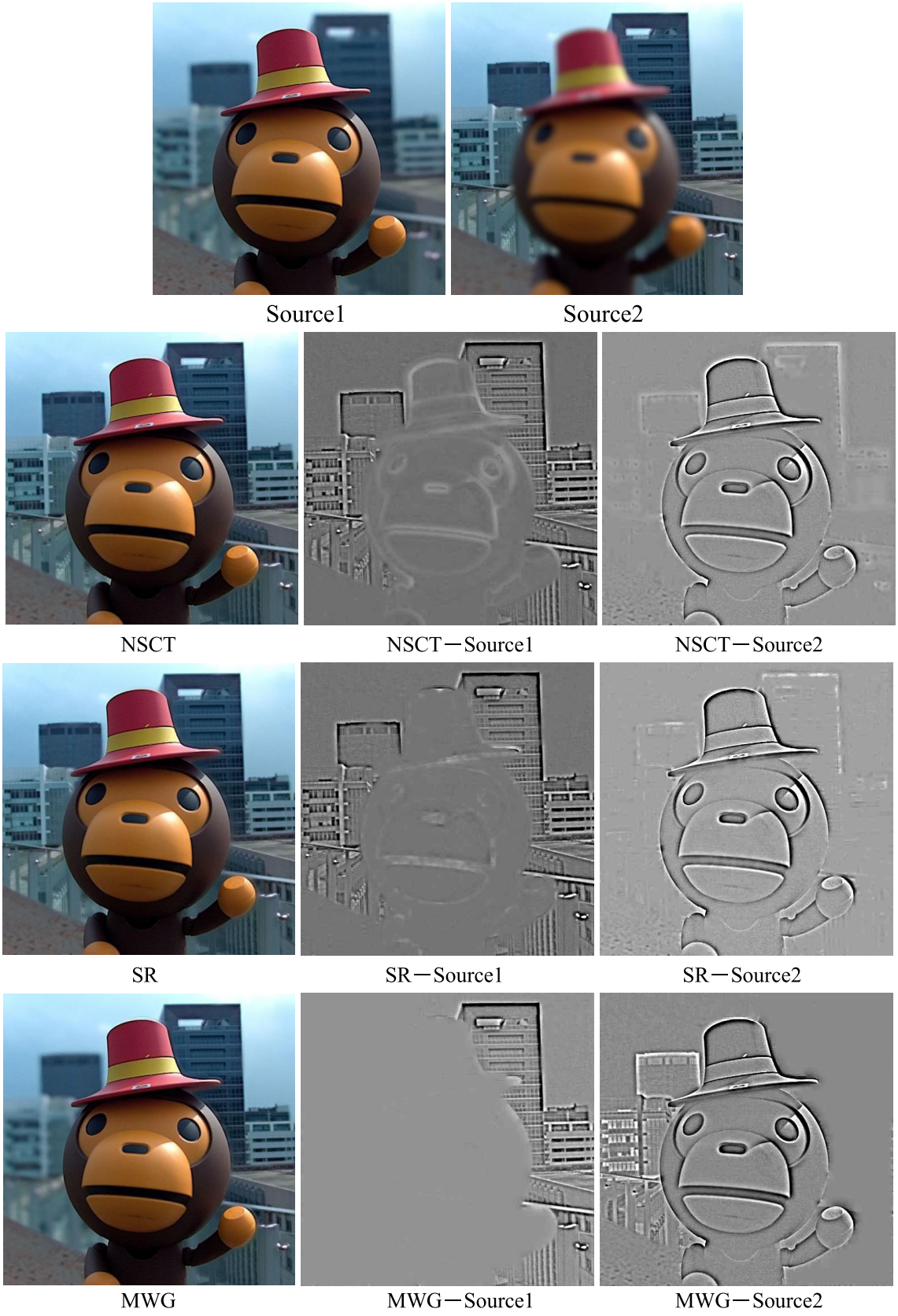

FIGURE 9. Fusion results of the 2nd image pair in Fig. 5 using different methods. The first column is the fused images obtained by NSCT, SR, MWG, DSIFT, CNN, ECNN and MADCNN. The other two columns are respectively the difference between the fused image and two source images.

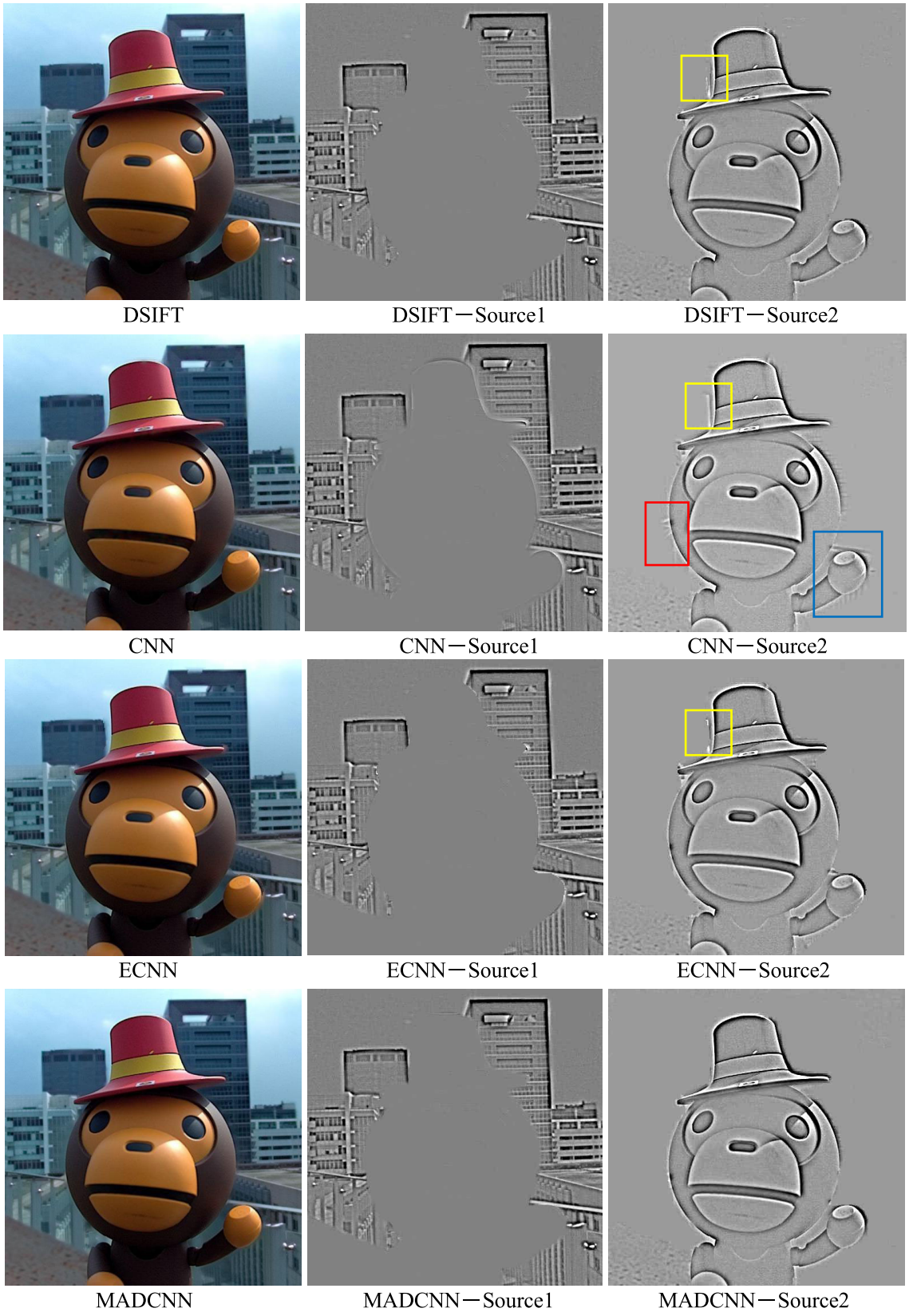

FIGURE 9. *(Continued.)* **Fusion results of the 2nd image pair in Fig. 5 using different methods. The first column is the fused images obtained by NSCT, SR, MWG, DSIFT, CNN, ECNN and MADCNN. The other two columns are respectively the difference between the fused image and two source images.**

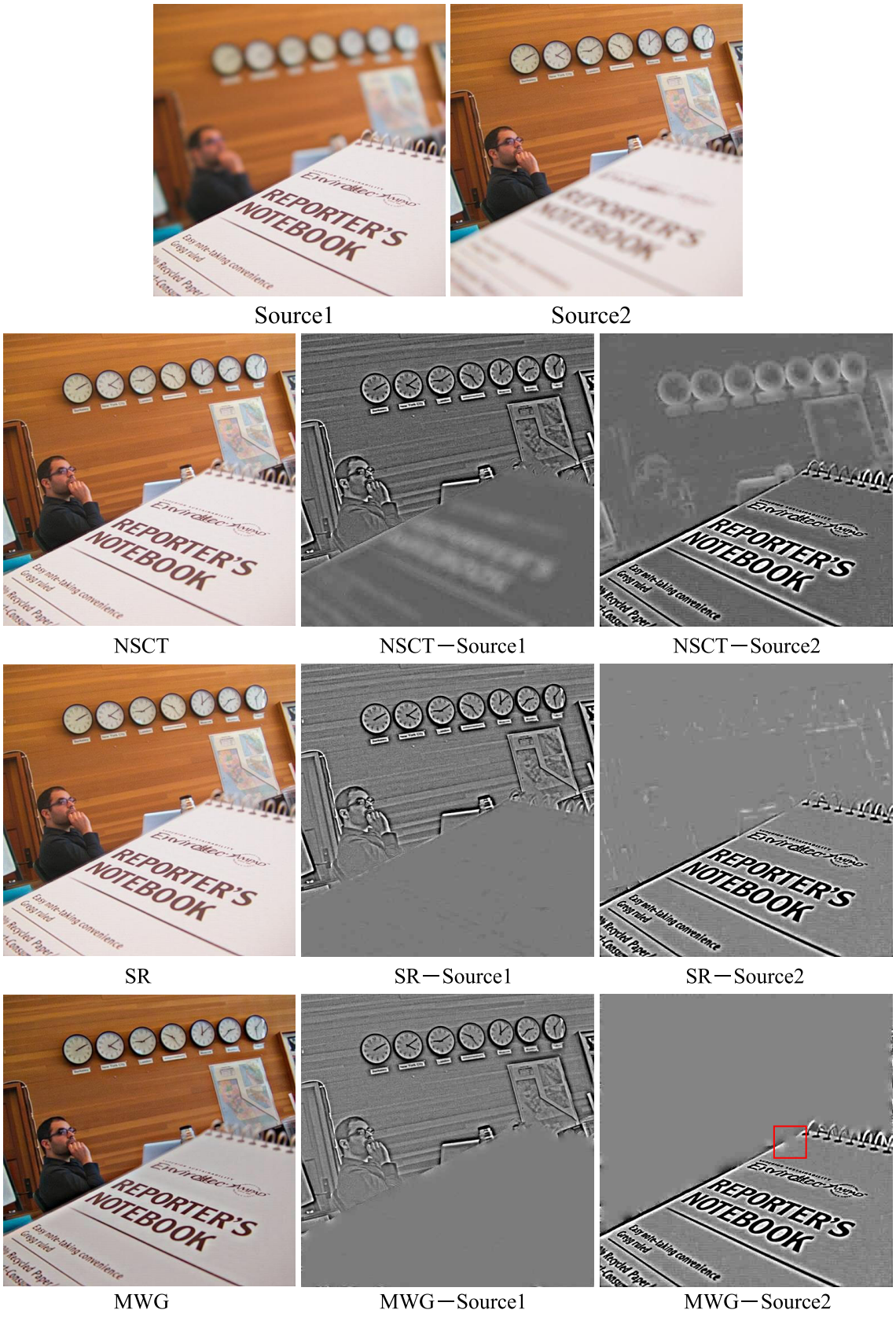

FIGURE 10. Fusion results of the 3rd image pair in Fig. 5 using different methods. The first column is the fused images obtained by NSCT, SR, MWG, DSIFT, CNN, ECNN and MADCNN. The other two columns are respectively the difference between the fused image and two source images.

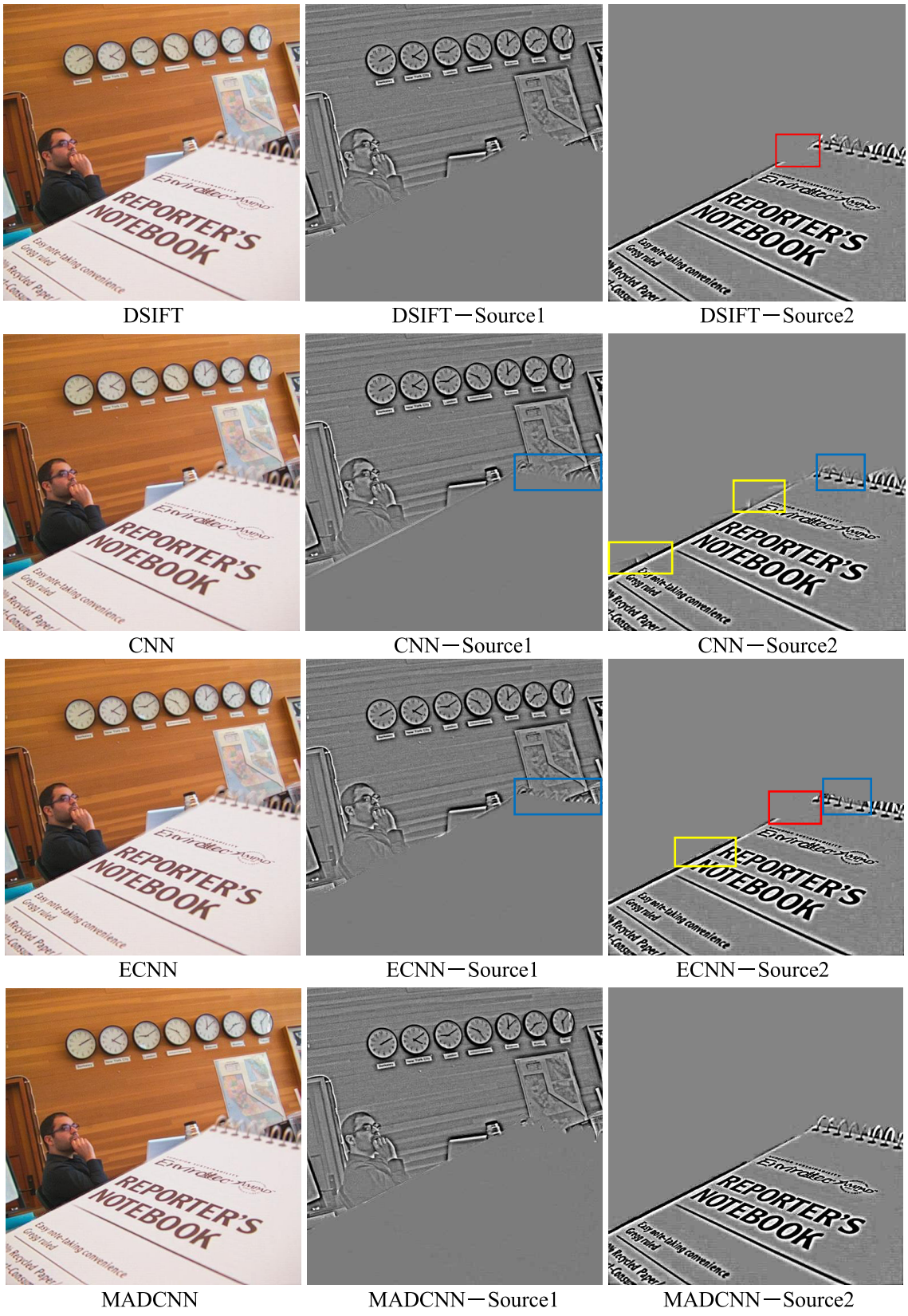

FIGURE 10. *(Continued.)* **Fusion results of the 3rd image pair in Fig. 5 using different methods. The first column is the fused images obtained by NSCT, SR, MWG, DSIFT, CNN, ECNN and MADCNN. The other two columns are respectively the difference between the fused image and two source images.**

TABLE 3. Objective assessment of different fusion methods.

Method	Q_{NCIE}	Q_P	Q_Y	Q_{CB}	Q_{MI}	Q_{TE}	rSFe
NSCT	0.8304	0.8178	0.9562	0.7461	0.9440	0.8100	1.5631
SR	0.8374	0.8168	0.9653	0.7756	1.0642	0.8133	2.0105
MWG	0.8396	0.8240	0.9786	0.7893	1.0918	0.8138	2.0557
DSIFT	0.8415	0.8320	0.9789	0.8014	1.1292	0.8162	2.1858
CNN	0.8402	0.8326	0.9781	0.7995	1.1095	0.8170	2.0407
ECNN	0.8423	0.8287	0.9797	0.8011	1.1345	0.8148	2.1782
MADCNN	**0.8437**	**0.8474**	**0.9861**	**0.8055**	**1.1641**	**0.8218**	**2.3852**

TABLE 4. Time complex of different models.

Method	CNN	ECNN	MADCNN
Parameters (M)	4.93	1.58	0.17
Computation (GOPs)	607.13	190310	42.59
Run Time (ms)	306	80000	181

blue and yellow boxes in its corresponding difference image. The proposed MADCNN method still outperforms the other competitive methods in visual effect.

For quantitatively assess the performance of different fusion methods, we carry out the fusion upon twenty pairs of source images with different scene and list the average scores belong to the seven metrics in Table 3. The optimal value of each metric is highlighted in bold. From the comparison of objective assessment, we can find that most performance scores of the former three methods comprehensively fall behind the later four methods. Specially, our proposed MADCNN method still achieves the overwhelmingly highest score in all of the seven fusion metrics.

To sum up briefly, the validity and superiority of the proposed MADCNN method are demonstrated by both of the subjective visual effects and objective performance assessment.

D. COMPLEXITY COMPARISION OF DEEP LARNING BASED METHODS

The deep learning based methods consume comparatively more computation to exchange for outstanding performance, which limits the popularization and application of these methods in real-time systems. To completely assess the complexity of the proposed method, we calculate the number of parameters (weight and bias) of different models. In addition, we perform the computation cost comparison on 520×520 sized dual-focus image fusion task and then list the results in Table 4. As can be seen that the patch-based CNN and ECNN methods need to calculate the patch recurrently and consume huge computation resources. In contrast, the MADCNN method is significantly outperforming the other deep learning based methods in computation and model parameters. As for the run time comparison, the MADCNN takes the shortest run time to fuse the dual-focus image which is beneficial from the lower model complexity.

IV. CONCLUSION

To solve the out-of-focus blur problem, this paper proposes a multi-focus image fusion method based on an end-to-end MADCNN model, which can get the final fused images directly from the source images. In this innovative MADCNN model, the MFE unit is designed to excavate enough feature information on different spatial scales. By introducing the attention unit, the network adaptively screens out the unwanted features and concerns only on the details in the focused region. The experimental results validated that the proposed method is superior to the existing algorithms in both of subjective and objective assessments.

In spite of the superior performance achieved by the proposed method, it still suffers from the drawbacks of relatively larger computation requirement than traditional methods. In future work, we will furtherly optimize our model and reduce the computational load by employing the recent proposed parallel depth wise separable convolution and apply our approach to hardware implementation for achieving better real-time performance with smaller power consumption.

ACKNOWLEDGMENT

The testing code of our MADCNN can be available at https://github.com/jtguan/MADCNN.

REFERENCES

[1] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COM-31, no. 4, pp. 532–540, Apr. 1983.
[2] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.
[3] A. Toet, "A morphological pyramidal image decomposition," *Pattern Recognit. Lett.*, vol. 9, no. 4, pp. 255–261, 1989.
[4] H. Li, B. S. Manjunath, and S. K. Mitra, "Multisensor image fusion using the wavelet transform," *Graph. Models Image Process.*, vol. 57, no. 3, pp. 235–245, 1995.
[5] 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.
[6] Q. Zhang and B.-L. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *Signal Process.*, vol. 89, no. 7, pp. 1334–1346, 2009.
[7] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 884–892, Apr. 2010.
[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] M. Amin-Naji and A. Aghagolzadeh, "Multi-focus image fusion in DCT domain using variance and energy of Laplacian and correlation coefficient for visual sensor networks," *J. AI Data Mining*, vol. 6, no. 2, pp. 233–250, 2018.

[10] 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, 2001.
[11] S. Li, J. T. Kwok, and Y. Wang, "Multifocus image fusion using artificial neural networks," *Pattern Recognit. Lett.*, vol. 23, no. 8, pp. 985–997, 2002.
[12] V. Aslantas and R. Kurban, "Fusion of multi-focus images using differential evolution algorithm," *Expert Syst. Appl.*, vol. 37, pp. 8861–8870, Dec. 2010.
[13] I. De and B. Chanda, "Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure," *Inf. Fusion*, vol. 14, no. 2, pp. 136–146, 2013.
[14] 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.
[15] M. Li, W. Cai, and Z. Tan, "A region-based multi-sensor image fusion scheme using pulse-coupled neural network," *Pattern Recognit. Lett.*, vol. 27, no. 16, pp. 1948–1956, 2006.
[16] S. Li and B. Yang, "Multifocus image fusion using region segmentation and spatial frequency," *Image Vis. Comput.*, vol. 26, no. 7, pp. 971–979, 2008.
[17] 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.
[18] 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.
[19] R. Lai, J. Guan, Y. Yang, and A. Xiong, "Spatiotemporal adaptive nonuniformity correction based on BTV regularization," *IEEE Access*, vol. 7, pp. 753–762, 2019.
[20] Y. Zhang, L. Sun, C. Yan, X. Ji, and Q. Dai, "Adaptive residual networks for high-quality image restoration," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3150–3163, Jul. 2018.
[21] J. Guan, R. Lai, and A. Xiong, "Wavelet deep neural network for stripe noise removal," *IEEE Access*, vol. 7, pp. 44544–44554, 2019.
[22] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense SIFT," *Inf. Fusion*, vol. 23, pp. 139–155, May 2015.
[23] 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.
[24] M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, "Ensemble of CNN for multi-focus image fusion," *Inf. Fusion*, vol. 51, pp. 201–214, Nov. 2019.
[25] M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, "CNNs hard voting for multi-focus image fusion," *J. Ambient Intell. Humanized Comput.*, to be published. doi: 10.1007/s12652-019-01199-0.
[26] X. Guo, R. Nie, J. Cao, D. Zhou, and W. Qian, "Fully convolutional network-based multifocus image fusion," *Neural Comput.*, vol. 30, no. 7, pp. 1775–1800, Jul. 2018.
[27] M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, "Fully convolutional networks for multi-focus image fusion," in *Proc. 9th Int. Symp. Telecommun. (IST)*, Dec. 2018, pp. 553–558.
[28] 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, 2013.
[29] 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.
[30] 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.
[31] X. Bai, M. Liu, Z. Chen, P. Wang, and Y. Zhang, "Multi-focus image fusion through gradient- based decision map construction and mathematical morphology," *IEEE Access*, vol. 4, pp. 4749–4760, Aug. 2016.
[32] 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.
[33] Y. Liu, X. Chen, H. Peng, and Z. F. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36 pp. 191–207, Jul. 2017.
[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[35] T. Stathaki, *Image Fusion: Algorithms and Applications*. Orlando, FL, USA: Academic, 2008, pp. 469–492.
[36] J. Zhao, R. Laganiere, and Z. Liu, "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement," *Int. J. Innov. Comput. Inf. Control*, vol. 3, no. 6, pp. 1433–1447, 2007.
[37] 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, 2008.
[38] Y. Chen and R. S. Blum, "A new automated quality assessment algorithm for image fusion," *Image Vis. Comput.*, vol. 27, no. 10, pp. 1421–1432, Sep. 2009.
[39] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on 'information measure for performance of image fusion,'" *Electron. Lett.*, vol. 44, no. 18, pp. 1066–1067, Aug. 2008.
[40] N. Cvejic, C. N. Canagarajah, and D. R. Bull, "Image fusion metric based on mutual information and Tsallis entropy," *Electron. Lett.*, vol. 42, no. 11, pp. 626–627, May 2006.
[41] Y. Zheng, E. A. Essock, B. C. Hansen, and A. M. Haun, "A new metric based on extended spatial frequency and its application to DWT based fusion algorithms," *Inf. Fusion*, vol. 8, no. 2, pp. 177–192, 2007.

RUI LAI received the B.S. degree in information engineering and the M.S. and Ph.D. degrees in electronic science and technology from Xidian University in 2002, 2005, and 2007, respectively.

From 2009 to 2010, he was a Postdoctoral Researcher with the Optoelectronic Science and Engineering Department, Huazhong University of Science and Technology. Since 2010, he has been an Associate Professor with the Microelectronics Department, Xidian University. He is the author of over 60 journals and conference papers. His research interests include computational photography, image sensors, image processing, and deep learning.

YONGXUE LI received the B.S. degree in microelectronics from Xidian University, Xi'an, China, in 2016, where he is currently pursuing the M.S. degree in microelectronics and solid-state electronics.

His research interest includes the applications of deep learning for image fusion.

JUNTAO GUAN received the B.S. degree in integrated circuit design and integration systems from the Harbin University of Science and Technology, in 2016. He is currently pursuing the Ph.D. degree in electronics science and technology with Xidian University, China.

His research interests include deep learning and image processing.

AI XIONG received the B.S. and M.S. degrees in automation and control theory and engineering from Chongqing University, Chongqing, in 2002 and 2005, respectively, and the Ph.D. degree in signal and information processing from the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China, in 2008, where he was a Research Assistant, from 2008 to 2016.

Since 2016, he has been an Associate Professor with the Control engineering College, Chengdu University of Information Technology, Chengdu, China. His research interests include optical sensors, high precision servo control, and adaptive control.