Hindawi
Scientific Programming
Volume 2021, Article ID 6614873, 12 pages
https://doi.org/10.1155/2021/6614873

Research Article

Suspect Multifocus Image Fusion Based on Sparse Denoising Autoencoder Neural Network for Police Multimodal Big Data Analysis

Jin Wang and Yanfei Gao

Department of Public Security, Railway Police College, Zhengzhou 450000, China

Correspondence should be addressed to Jin Wang; wangjj815@163.com

Received 4 November 2020; Revised 11 December 2020; Accepted 28 December 2020; Published 7 January 2021

Academic Editor: Liang Zhao

In recent years, the success rate of solving major criminal cases through big data has been greatly improved. The analysis of multimodal big data plays a key role in the detection of suspects. However, the traditional multiexposure image fusion methods have low efficiency and are largely time-consuming due to the artifact effect in the image edge and other sensitive factors. Therefore, this paper focuses on the suspect multiexposure image fusion. The self-coding neural network based on deep learning has become a hotspot in the research of data dimension reduction, which can effectively eliminate the irrelevant and redundant learning data. In the case of limited field depth, due to the limited focusing depth of the camera, the focusing plane cannot obtain the global clear image of the target in the depth scene, which is prone to defocusing and blurring phenomena. Therefore, this paper proposes a multifocus image fusion based on a sparse denoising autoencoder neural network. To realize an unsupervised end-to-end fusion network, the sparse denoising autoencoder neural network is adopted to extract features and learn fusion rules and reconstruction rules simultaneously. The initial decision graph of the multifocus image is taken as a prior input to learn the rich detailed information of the image. The local strategy is added to the loss function to ensure that the image is restored accurately. The results show that this method is superior to the state-of-the-art fusion methods.

1. Introduction

Image fusion refers to the comprehensive processing of two or more complementary source images obtained from different sensors to obtain a new fused image, which enables the fused image to have higher credibility [1–4], clarity, and better understandability. In the case of limited field depth, due to the limited focusing depth of the camera, the focusing plane cannot obtain the global clear image of the target in the depth scene, which is prone to defocusing and blurring phenomena. Multifocus image fusion technology is to fuse multiple images with different focus positions in the same scene into a fully focused image with more information [5]. At present, multifocus image fusion algorithms can be divided into transform domain-based fusion method, space domain-based fusion method, and deep learning-based fusion method according to the fusion strategy.

The fusion method based on the transform domain generally uses a variety of decomposition tools to decompose the source image into multilevel coefficients and then designs different fusion rules according to the characteristics of each level coefficient [6, 7]. Finally, it performs the inverse multiscale transformation on the fused coefficients of each level to obtain the fused image. The design of transformation tools and the design of fusion rules play an important role in the fusion performance of transformation domain-based fusion methods.

Common transformation tools include curvelet transform (CVT) [8], nonsubsampled contourlet transform (NSCT) [9], Laplacian pyramid (LP) [10], low-pass pyramid, and gradient pyramid (GP) [11]. The fusion rules include maximization, weighted average, saliency, and active contour. The sparse representation (SR), higher-order singular value decomposition (HOSVD) [12], and other sparse

principal component analysis- (RPCA-) based multifocus image fusion methods [13] have attracted more attentions.

The fusion method based on the spatial domain can be divided into three types according to the different focus measurement objects: pixel-based, block-based, and region-based. The pixel-based multifocus image fusion method can extract the feature information from the source image and retain the original information to the greatest extent. It has the characteristics of high accuracy and strong robustness, which includes dense scale-invariant feature transform (DSIFT), guided filtering (GF), and image matting (IM). The multifocus image fusion method based on blocks and regions adopts some segmentation strategies to divide the source image into different blocks or regions and then selects more focus blocks or regions as part of the fused image by focus measurement [14]. The common focus measurement methods include image gradient and spatial frequency. The block size and segmentation algorithm can directly affect the visual effect of the fused image, which is prone to "block effect." Both transform domain-based fusion methods and spatial domain-based fusion methods require to manually design the fusion rules. However, complex image scenes limit the expressive ability of features and the robustness of fusion rules.

In order to improve the feature expression ability and the robustness of fusion rules, deep learning technology has been introduced into multifocus image fusion research [15–17]. Karim et al. [18] proposed a drone plane for monitoring and targeting street crime criminals based on real time image processing techniques. Liu et al. [19] used the multiscale Gaussian filter with different standard deviations to fuzzy process the random region on the gray image to simulate the multifocus image. By using supervised learning, the image was classified into focusing pixels and defocusing pixels, and the focus map with the same size as the input image was obtained. Then, the focus decision graph was generated by verifying the size and consistency of the focus map. Finally, based on the judging criteria, the weighted average strategy was used to obtain the fused images in the spatial domain. Tang et al. [20] proposed a multifocus image fusion method based on a pixel-wise convolutional neural network (P-CNN). This model used Cifar10 as the training set, and three kinds of pixels could be learned from adjacent pixel information: focusing pixel, defocusing pixel, and unknown pixel. After the source image was scored by PCNN, a scored matrix representing the focusing level of the pixel was formed. Then, by comparing the scores matrix of the two source images, then it obtained the decision graph. Finally, the weighted average value of the two input images was obtained according to the final decision graph filtered by a threshold. The model had excellent performance in real-time performance and fusion effect, but the limitation of supervised learning was that accurate label data could not be obtained for image fusion.

To further distinguish the private and public features in multifocus images, Luo et al. [21] proposed a joint convolution self-encoding network, which obtained the focus map based on the image features learned by the private branch and used the pixel-level weighted average rule to obtain the fully focused fused image. This method adopted unsupervised learning and did not need manually designed label and achieved ideal results on subjective evaluations and multiple objective evaluation. However, these methods only take advantage of CNN feature extraction and classification capability and still use the manually designed fusion rules, which makes the model unable to adjust the fusion strategy according to the application scenarios.

To further realize the self-learning of fusion rules and make full use of the feature extraction of CNN, combined with the prior knowledge of manual features, in this paper, a multifocus image fusion network with self-learning fusion rules is designed. The multifocus image and its initial decision graph are taken as the input of the network, so that the network can learn more accurate detailed information. The structural similarity index measure (SSIM) and local mean squared error (MSE) are used as loss functions to drive fusion rules.

The rest of this paper is organized as follows. Section 2 designs the proposed approach and, after that, Section 3 describes experimental results. Finally, Section 4 concludes the paper.

2. Proposed Multifocus Image Fusion

This paper first introduces the network structure of multifocus image fusion, then discusses the network fusion in detail, and finally discusses the loss function design.

2.1. Feature Extraction Network Based on Sparse Denoising Autoencoder Neural Network. Figure 1 shows the sparse denoising autoencoder neural network (SDNA-ENN).

The whole network is divided into the input layer, coding layer, fusion layer, decoding layer, and output layer. The input layer includes the initial decision graph of multifocus image A, multifocus image B, and multifocus image A. The coding layer includes 9 trainable convolutional layers with a convolution kernel size of 3×3, and each convolutional layer is followed by a ReLU layer. The coding layer can be divided into the private branch PriA, public branch ComA of multifocus image A, and the private branch PriB, and public branch ComB of multifocus image B, where PriA and PriB are used to extract the private features of the input images, respectively. ComA and ComB share weights to extract the common features from multiple input images. The fusion layer cascades the feature map output by PriA and PriB along the channel and then connects the cascaded feature map to the next trainable convolution layer with a convolution kernel size of 1×1. The output feature map of ComA and ComB is treated in the same way as PriA and PriB. The decoding layer consists of four trainable convolution layers with a convolution kernel size of 3×3, and the last convolutional layer is used to reconstruct the fully focused image. In this paper, a short connection is added to the public branch to solve the problem of gradient disappearance during the training process. Compared with the previous networks, this new network adds fusion units and uses short connections to improve the robustness of feature learning.

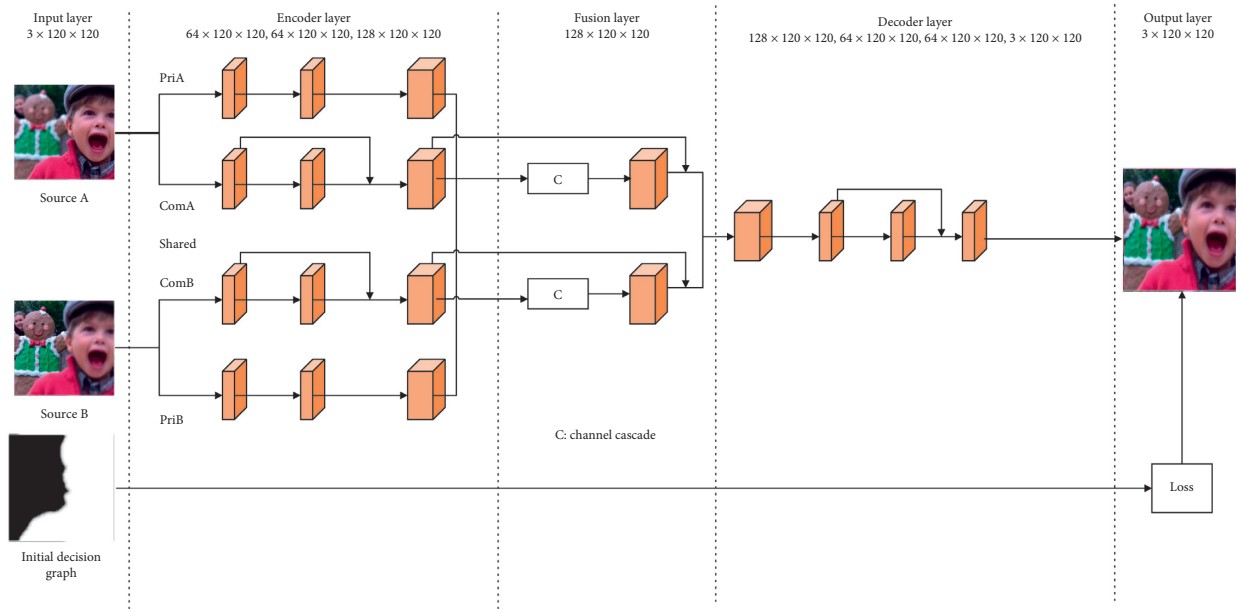


Figure 1: Structure of SDNA-ENN.

2.2. Fusion Layer Design. In the study of multifocus image fusion based on deep learning, the network fusion layer usually contains two methods that can be used to fuse the convolution features of multiple inputs:

(1) Cascade the convolution features of multiple inputs along the channel, and then fuse them with the next convolutional layer

(2) The multiple input convolution features are fused by the pixel-level fusion rule

The cascade fusion method stacks multiple inputs, so that the network can learn sufficient feature information.

The pixel-level fusion rule includes summation, taking large and mean value [22]. The fusion strategy can be selected according to the features of the data set. In multifocus images, because the pixel value of the image represents the information saliency, the proposed method in this paper introduces the mean rule on the basis of cascading fusion to ensure the diversity and accuracy of feature learning. The concrete realization of the fusion layer design includes weight initialization and weight constraint.

2.2.1. Weight Initialization. The weight initialization is to simulate the weighted average fusion rule, and the features extracted from the coding layer can be accurately fused by the reasonable weight assignment in the fusion layer. The output feature graphs of PriA, PriB, ComA, and ComB coding layers are splicing along the channel, followed by a trainable convolutional layer of 1×1. The first and $1 + p$ weight value of the k-th channel in the 1×1 convolutional layer is initialized to 0.5; that is,

$$W_k^l = W_k^{I+p} = 0.5, \quad I = 1, \ldots, 127; k = 1, 2 \ldots, 127, \quad (1)$$

where k is the channel number after the convolution operation. I is the filter number of the k-th channel. $P = 128$, which can be adjusted according to actual requirements. W_k^l is the I-th weight value of the k-th channel.

2.2.2. Weight Constraint. Because the weight value may appear numerical over-bounds phenomenon in the process of network iteration, the constraints are added to each weight value to realize the weight value fluctuation in the effective range. According to the mean value rule in the image fusion method, the sum of fusion coefficients of the two images is 1. However, the activation function of the training network adopts ReLU, for the k-th channel, $\sum_{I=0}^{p-1} W_k^I + \sum_{I=p}^{2p-1} W_k^{I+p} > 1$. Therefore, we make two improvements in this process. One is to improve the activation cost function, the second is to apply the minimum/maximum norm weight constraint to the $2p$ weights of the k-th channel in the fusion layer.

In order to make the activation units with fewer hidden layers represent the most effective features, through the traditional autoencoder neural network research, this paper proposes to add sparse restriction to the hidden neurons in the denoising autoencoder neural network (DAE), which can suppress most of the output neurons and use fewer activation units to represent features.

The sparse denoising autoencoder network structure consists of a sparse denoising autoencoder and a softmax classifier as shown in Figure 2. X represents the original data layer, $\tilde{X}$ represents the data layer with disturbing noise, and $\tilde{H}$ represents the hidden layer.

Specifically, assuming that the number of input samples is m. x represents the input. y represents the output. l represents the layer number of the neural network. s^l represents the neuron number in hidden layer l. Then the

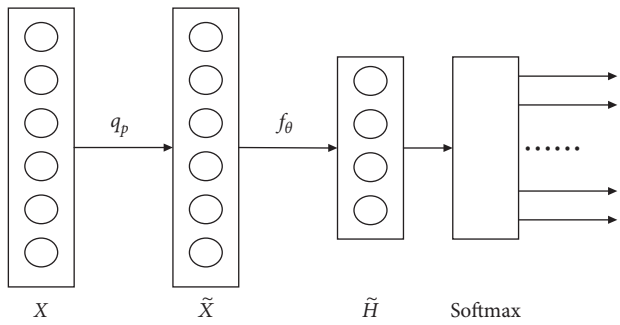


Figure 2: Structure of sparse denoising autoencoder neural network.

activation cost function of the sparse denoising autoencoder neural network is defined as follows:

$$J_{\mathrm{SDAE}}(w,b) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{w,b}(\widetilde{x})^{(i)} - y^{(i)}\right\|^2\right) + \beta\sum_{j=1}^{s^l}\mathrm{KL}\left(\rho\|\widehat{\rho}_j\right). \tag{2}$$

The residual of each neuron in the hidden layer is

$$\begin{cases} \delta_i^l = -(y_i - a_i)f'(z_i), & \text{when } l \text{ is output layer,} \\ \delta_i^l = \left[\left(\sum_{j=1}^{s_2} w_{ji}\delta_i^{l+1}\right) + \beta\left(-\dfrac{\rho}{\widehat{\rho}_j} + \dfrac{1-\rho}{1-\widehat{\rho}_j}\right)\right]f'(z_i), & \text{when } l \text{ is hidden layer.} \end{cases} \tag{3}$$

Then, the partial derivatives of weight and bias items are calculated as follows:

$$\begin{aligned} \nabla_{W^l}J(W,b) &= \frac{\partial}{\partial W_i^l}J_{\mathrm{SDAE}}(w,b) = a_j^l\delta_i^{l+1}, \\ \nabla_{b^l}J(W,b) &= \frac{\partial}{\partial b_i^l}J_{\mathrm{SDAE}}(w,b) = \delta_i^{l+1}. \end{aligned} \tag{4}$$

Then, we calculate the L2-norm of $2p$ weights in the k-th channel.

$$S_k = \sqrt{\sum_{I=0}^{p-1}\left(W_k^I\right)^2 + \sum_{I=p}^{2p-1}\left(W_k^{I+p}\right)^2}. \tag{5}$$

S_k is truncated in the range $(S_{\min}, S_{\max})$; that is,

$$S_t = \begin{cases} S_{\min}, & S_k < S_{\min}, \\ S_k, & S_{\min} < S_k < S_{\max}, \\ S_{\max}, & S_k > S_{\max}, \end{cases} \tag{6}$$

where $S_{\min}$ is the minimum L2-norm of input weight value. $S_{\max}$ is the maximum L2-norm of input weight value.

Finally, each weight value of the k-th channel is readjusted.

$$\begin{aligned} W_k^m &= W_k^m \times Z_k, \quad m = 0, 1, 2, \ldots, 2p-1, \\ Z_k &= \frac{\alpha \times S_t + (1-\alpha)\times S_k}{\gamma + S_k}, \end{aligned} \tag{7}$$

where W_k^m is the m-th weight value of the k-th channel and Z_k is the constraint range of the weight value. α is the proportion of constraint; when $\alpha = 1$, the constraint is strictly enforced, and when $\alpha < 1$, the weight must be adjusted for each step. In order to avoid gradient explosion, $\gamma = e^{-3}$. After weight initialization and constraint, the rules of the fusion layer are finally converted to

$$\widehat{f}_k(x,y) = W_k^I f_I(x,y) + W_k^{I+p} f_{I+p}(x,y). \tag{8}$$

2.3. The Design of Loss Function. In order to ensure that the network can learn the features of the input image accurately and effectively, the local strategy is added into the loss function, including local structure similarity and local mean square error.

2.3.1. Local Structure Similarity. Human visual system is more sensitive to structural loss and deformation. Therefore, the structural similarity index measure (SSIM) [23] can be used to intuitively compare the structural information of distorted images and original images. SSIM is mainly composed of three parts: relevancy, brightness, and contrast as shown in the following:

$$\mathrm{SSIM}(X,F) = \sum_{x,f}\frac{\left(2\mu_x\mu_f + C_1\right)\left(2\mu_x\mu_f + C_1\right)\left(2\mu_x\mu_f + C_1\right)}{\left(\mu_x^2 + \mu_f^2 + C_1\right)\left(\sigma_x^2 + \sigma_f^2 + C_2\right)\left(\sigma_x\sigma_f + C_3\right)}, \tag{9}$$

where $\mathrm{SSIM}(X, F)$ represents the structural similarity of source image X and fused image F. x and f represent the image blocks in the source image and the fused image, respectively. μ_x and σ_x represent the mean and standard deviation of the image X, respectively. μ_f and σ_f represent the mean and standard deviation of fused image F respectively. σ_{xf} represents the covariance of the source image and the fused image. C_1, C_2, and C_3 are the parameters used to stabilize the algorithm.

On the basis of SSIM, the corresponded region of image X is extracted by combining the initial decision graph X_m of the input image X.

$$\overline{X} = \min\left(X_m, X\right). \tag{10}$$

The initial decision graph corresponding to the input images A and B are X_A and X_B, respectively. According to (10), corresponding regions $\overline{A}$, $\overline{B}$, and $\overline{F}$ of images A and B and fused image F can be obtained, respectively. According to (9), $\mathrm{SSIM}(\overline{A}, t\overline{F})$ and $\mathrm{SSIM}(\overline{B}, t\overline{F})$ can be calculated.

2.3.2. Local Mean Square Error. Mean square error is used to measure the difference degree between the source image and the fused image. The mean square error is inversely proportional to the quality of the fused image. The smaller value denotes higher fusion quality. Its calculation formula is

$$\mathrm{MSE}(X, F) = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \left(X(i, j) - F(i, j)\right)^2, \tag{11}$$

where $\mathrm{MSE}(X, F)$ represents the difference between the input image X and the fused image F.

According to (11), $\mathrm{MSE}(\overline{A}, t\overline{F})$ and $\mathrm{MSE}(\overline{B}, t\overline{F})$ can be obtained.

The final loss function of the proposed network is

$$L = \lambda_1 \left(\mathrm{SSIM}(\overline{A}, \overline{F}) + \mathrm{SSIM}(\overline{B}, \overline{F})\right) + \lambda_2 \left(\mathrm{MSE}(\overline{A}, \overline{F}) + \mathrm{MSE}(\overline{B}, \overline{F})\right), \tag{12}$$

where λ_1 and λ_2 represent the weights of local structure similarity and local mean square error, respectively. In this paper, λ_1 is used to adjust the similarity between the fused image and the source image. The larger λ_1 denotes the higher similarity between the fused image and the source image. λ_2 is used to enhance the focus area of the source image in the fused image. The larger λ_2 denotes the significant focus area of the source image. Based on the extensive experiments, this paper sets $\lambda_1 = 5$, $\lambda_2 = 5$, respectively.

3. Experiment and Analysis

In order to verify the performance of the proposed fusion method, we conduct comparison experiments with seven state-of-the-art fusion methods, namely, DE [24], NFBD [25], GDMC [26], LRRW [27], NNSR [28], CFM [29], and FRL-PCNN [30]. The experiment environment is MATLAB7a, Windows10, GPU TX1060, Memory 16G, and Intel(R) Core(TM) i7-67001. The Keras framework of Tensorflow is used for network training in this paper. All the comparison methods use the same parameters [31, 32]. Then, the detailed subjective and objective comparison and analysis are carried out on multiple multifocus images.

Because suspects are classified as the country secret data, this paper tests suspects and open datasets in the laboratory. The results are only from the public datasets. This paper conducts experiments on 60 pairs of multifocus images. 20 pairs are from the open-source dataset Lytro [33], the other 20 pairs have been widely used in the study of multifocus image fusion, and another 20 pairs are from actual suspect images. The sliding window method is adopted to take blocks with a stride length of 14. Each image in the dataset is divided into M image blocks with 224×224pixel. The initial decision graph acquisition in this paper consists of three parts: segmentation, mapping, and reprocessing. First, each image in the dataset is segmented into blocks with 4×4 pixel, and the spatial frequency is calculated. Then, the spatial frequency matrix is mapped to the original size of the source image, and the overlap part is processed with the mean value to obtain the spatial frequency map. The binary map is obtained by comparing the size. Finally, the initial decision graph of the network is obtained through consistency verification and guided filtering. The fusion results with different methods are shown in Figures 3–8.

To compare the fusion methods more intuitively, this paper selects a smaller region at a certain contour in each fused image, marks it with rectangular box, and gives an enlarged region. We give an analysis for image "disk." It can be seen from Figure 7 that the above methods can obtain fully focused images with good subjective vision. DE and NFBD present false information such as "artifact" in the edge of alarm clock. The fusion effect of IM is good, but there is a certain "Gibbs" phenomenon in the disk area, and some details are lost. GDMC shows fuzzy distortion in the local amplification region due to the emphasis on looking for boundaries and the focus metric is performed within a single block. The fusion results from LRRW, NNSR, CFM, and FRL-PCNN are good, but there is a slight "sag" on the left edge of the alarm clock.

Comparatively, the visual effect of the proposed method in this paper is similar to the subjective visual effect of other methods. It can be seen from the enlarged area in Figure 7 that the proposed method in this paper handles the details well, especially the edge area of the alarm clock is smooth and natural. A better fusion result is obtained. Since the initial decision graph of the focused image and the local strategy of the loss function are added into the network, the obtained fused image by the proposed method in this paper performs well in the retention of key information and is suitable for human visual perception. Figures 3–6 and 8 show the fusion results of the other 5 pairs of multifocus images in various fusion methods. As can be seen from the figures, all the methods can better fuse the multifocus image to some extent. Compared with other methods, the proposed method achieves better fusion results.

To objectively evaluate the results of each fusion method, this paper uses the evaluation index: entropy (EN), Q_W proposed by Piella and Heijmans, correlation-coefficient (CC), and Visual Information Fidelity (VIFF) to verify the

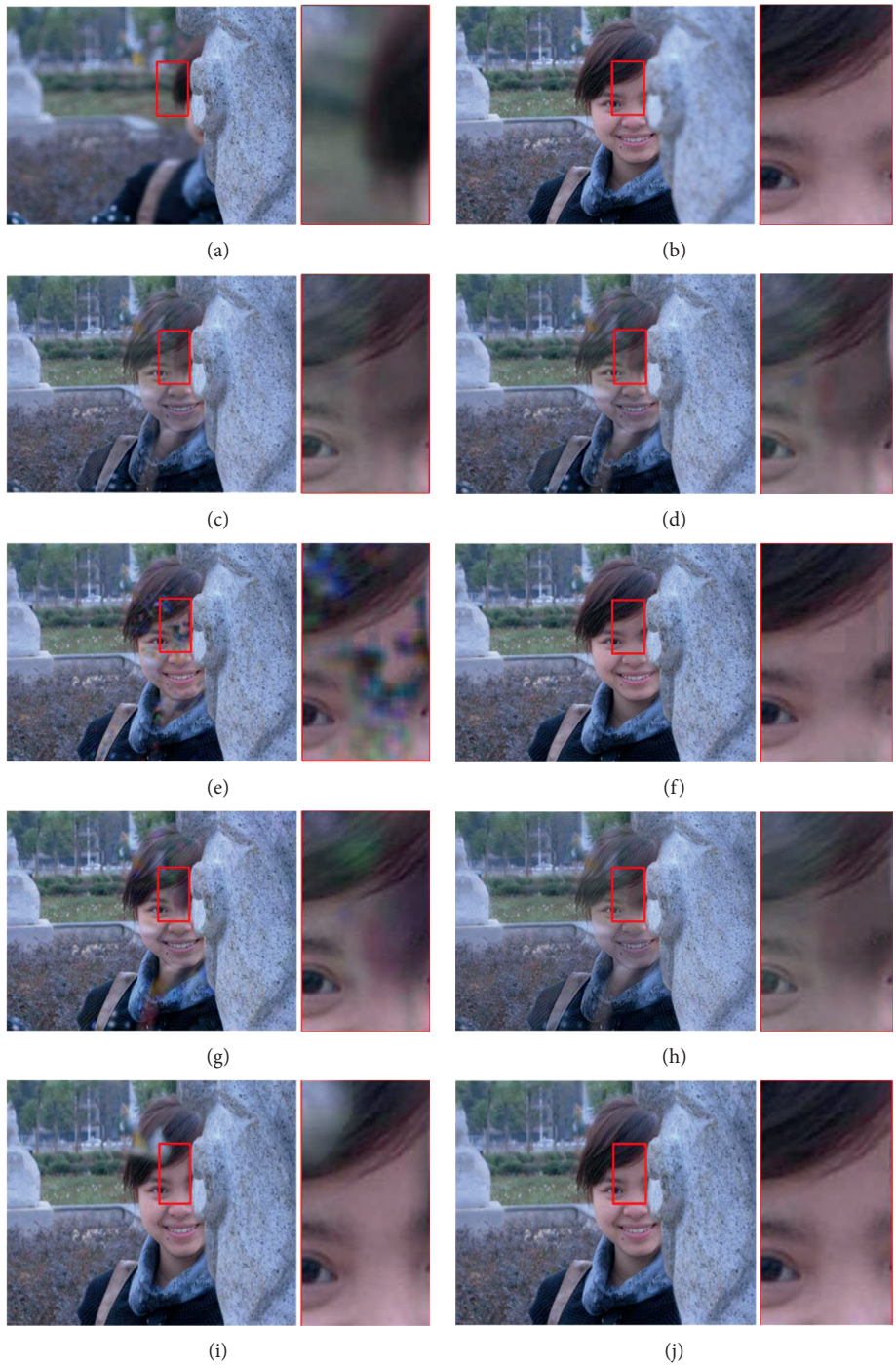


Figure 3: The "girl" source images and result images with different algorithms. (a, b) Source images; (c–j) the fusion images of DE, NFBD, GDMC, LRRW, NNSR, CFM, FRL-PCNN, and proposed method.

effectiveness of the proposed method. Entropy is an index based on information theory, which is used to reflect the amount of information in an image. If the entropy value is relatively large, it indicates that the fused image contains relatively more information. Q_W is a variant of the universal image quality index, which explores the position and size of distorted pixels by assigning high weights to visual saliency areas. The greater Q_W denotes the better fusion effect. The correlation coefficient measures the correlation between the source image and the fused image. The correlation value is positively correlated with the fusion effect. The VIFF is an index that simulates the subjective vision of human eyes to measure the fidelity of fused image. It includes four steps: partitioning, evaluation, calculating the fidelity of subband, and calculating the total fidelity. The higher VIFF presents the lower the distortion between the fused image and the source image. In order to ensure the fairness of objective evaluation, all indexes use the same parameters.

Figure 4: The "tree" source images and result images with different algorithms. (a, b) Source images; (c–j) the fusion images of DE, NFBD, GDMC, LRRW, NNSR, CFM, FRL-PCNN, and proposed method.

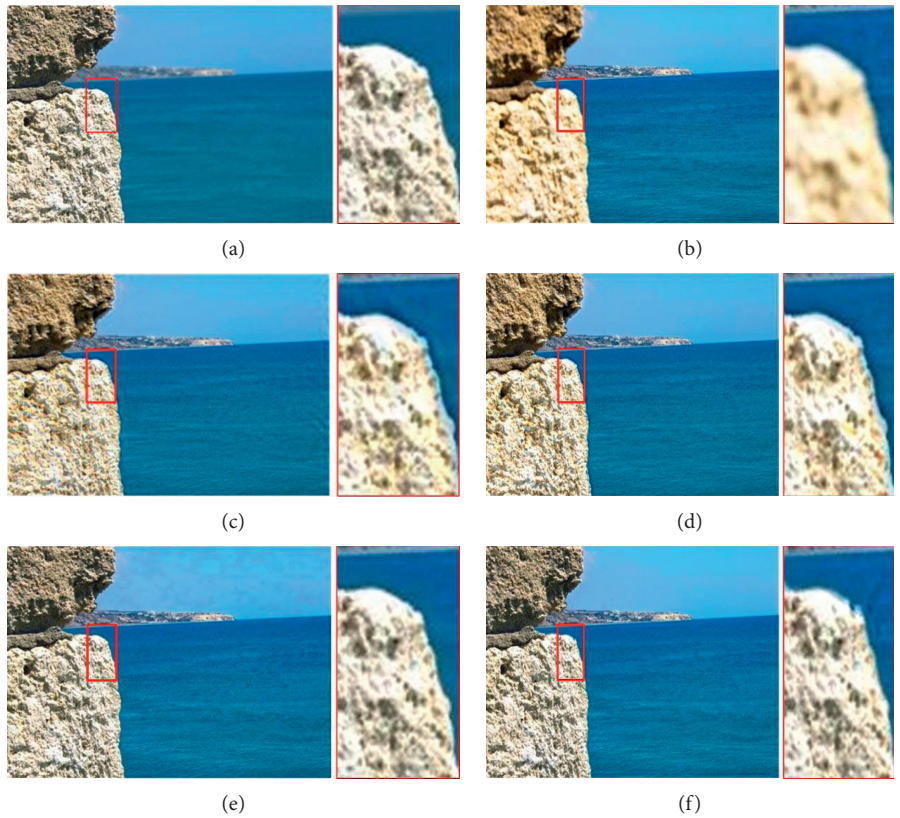
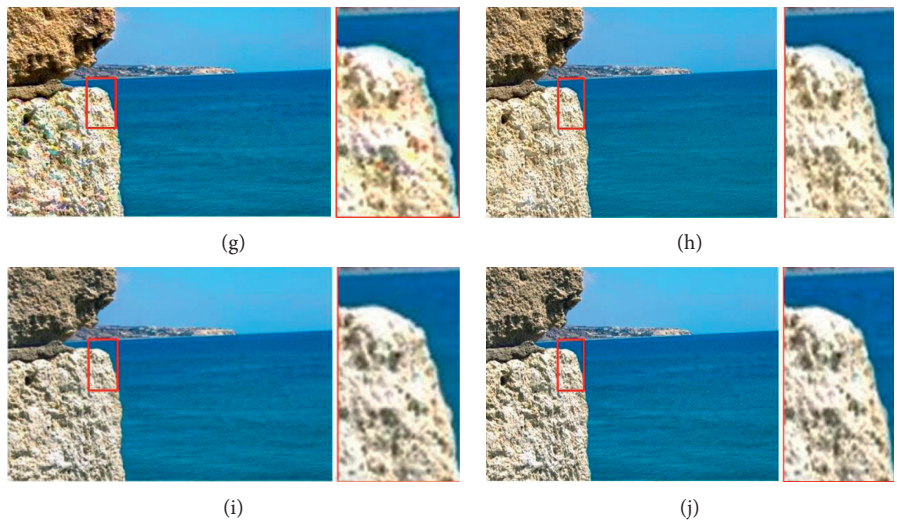


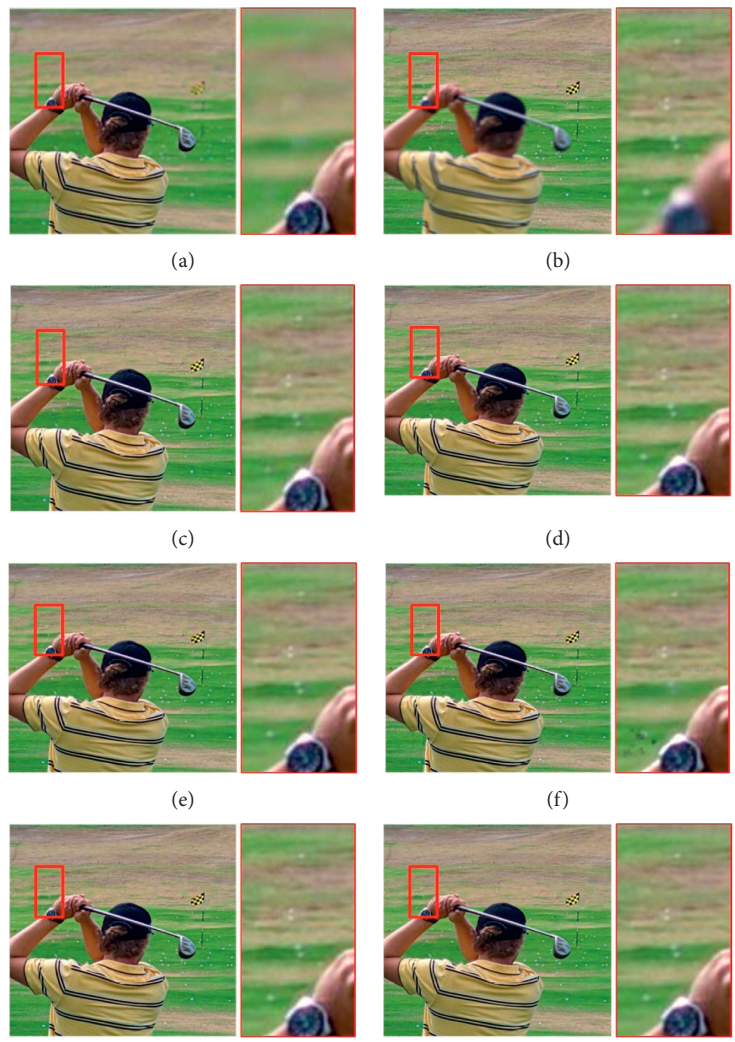
Figure 5: Continued.



(g) (h)

(i) (j)

Figure 5: The "sea" source images and result images with different algorithms. (a, b) Source images; (c–j) the fusion images of DE, NFBD, GDMC, LRRW, NNSR, CFM, FRL-PCNN, and proposed method.



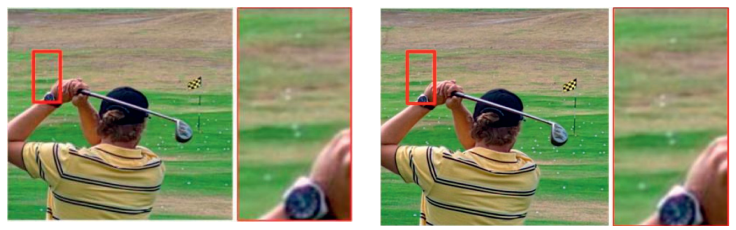
(a) (b)

(c) (d)

(e) (f)

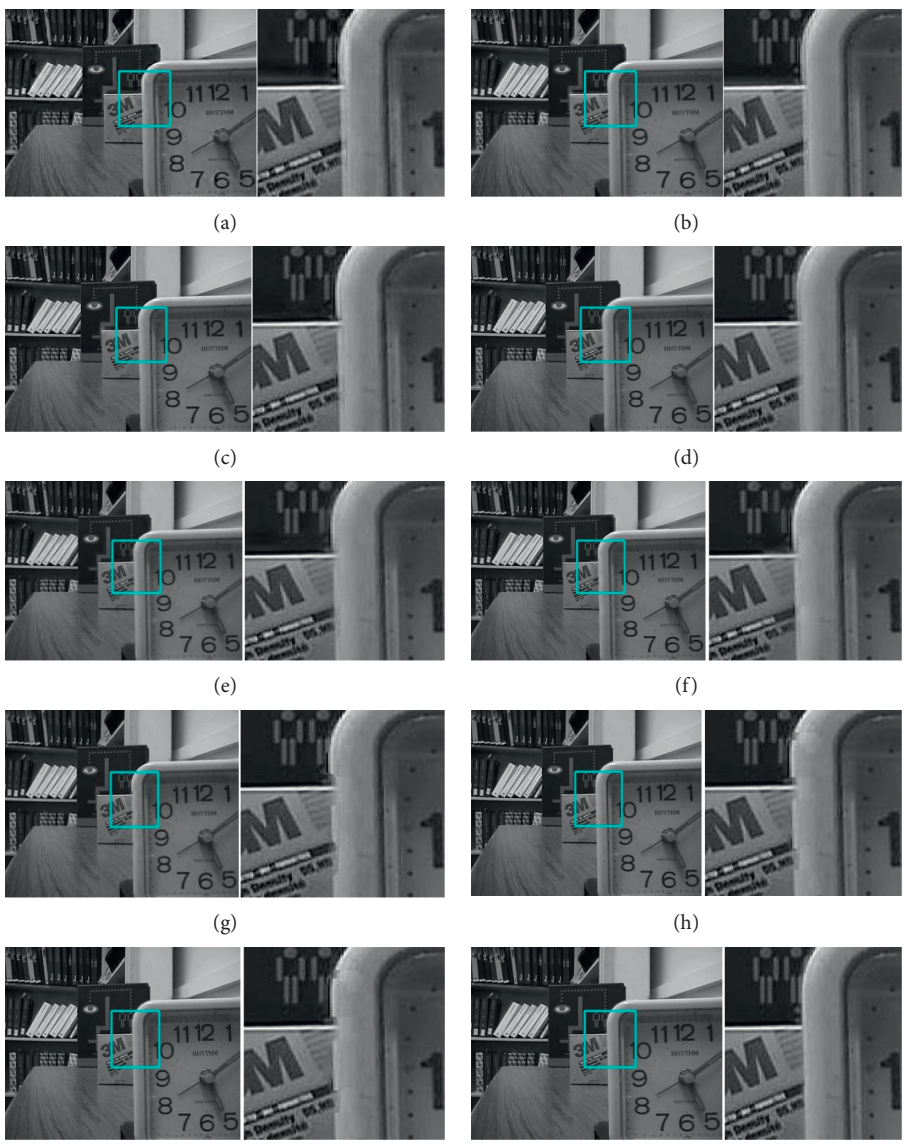
(g) (h)

Figure 6: Continued.



(i) (j)

Figure 6: The "golf" source images and result images with different algorithms. (a , b) source images; (c–j) the fusion images of DE, NFBD, GDMC, LRRW, NNSR, CFM, FRL-PCNN, and proposed method.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

(i) (j)

Figure 7: The "disk" source images and result images with different algorithms. (a, b) Source images; (c–j) the fusion images of DE, NFBD, GDMC, LRRW, NNSR, CFM, FRL-PCNN, and proposed method.

Figure 8: The "temple" source images and result images with different algorithms. (a, b) Source images; (c–j) the fusion images of DE, NFBD, GDMC, LRRW, NNSR, CFM, FRL-PCNN, and proposed method.

Table 1: The objective metrics of different fusion approaches for image "girl".

Method	EN	Q_W	CC	VIFF
DE	7.8668	0.8654	0.9811	0.7298
NFBD	7.8651	0.8552	0.9824	0.7398
GDMC	7.8662	0.8741	0.9825	0.7399
LRRW	7.8674	0.8742	0.9822	0.7435
NNSR	7.8662	0.8746	0.9827	0.7471
CFM	7.8673	0.8755	0.9836	0.7488
FRL-PCNN	7.8695	0.8768	0.9839	0.7527
Proposed	**7.8698**	**0.8879**	**0.9857**	**0.7879**

Table 2: The objective metrics of different fusion approaches for image "tree".

Method	EN	Q_W	CC	VIFF
DE	7.4493	0.8894	0.9721	0.8625
NFBD	7.4512	0.8942	0.9738	0.8756
GDMC	7.4528	0.8953	0.9742	0.8747
LRRW	7.4538	0.8974	0.9758	0.8827
NNSR	7.4688	0.8995	0.9763	0.8875
CFM	7.4848	0.9027	0.9778	0.8957
FRL-PCNN	7.5128	0.9037	0.9781	0.8998
Proposed	**7.6456**	**0.9122**	**0.9829**	0.9025

Table 3: The objective metrics of different fusion approaches for image "sea".

Method	EN	Q_W	CC	VIFF
DE	7.1803	0.9237	0.9617	0.9187
NFBD	7.1809	0.9242	0.9618	0.9238
GDMC	7.1814	0.9248	0.9624	0.9247
LRRW	7.1816	0.9257	0.9627	0.9355
NNSR	7.1825	0.9259	0.9633	0.9371
CFM	7.1897	0.9265	0.9638	0.9382
FRL-PCNN	7.1907	0.9277	0.9688	0.9407
Proposed	**7.1938**	**0.9359**	**0.9729**	**0.9477**

Table 4: The objective metrics of different fusion approaches for image "golf".

Method	EN	Q_W	CC	VIFF
DE	7.2714	0.9286	0.9824	0.9367
NFBD	7.2728	0.9288	0.9827	0.9382
GDMC	7.2734	0.9297	0.9828	0.9341
LRRW	7.2741	0.9314	0.9831	0.9345
NNSR	7.2832	0.9319	0.9833	0.9346
CFM	7.2835	0.9324	0.9837	0.9452
FRL-PCNN	7.2839	0.9328	0.9841	0.9557
Proposed	**7.2847**	**0.9342**	**0.9852**	**0.9722**

Table 5: The objective metrics of different fusion approaches for image "disk".

Method	EN	Q_W	CC	VIFF
DE	7.2654	0.9271	0.9762	0.8692
NFBD	7.2693	0.9317	0.9768	0.8714
GDMC	7.2754	0.9345	0.9769	0.8725
LRRW	7.2768	0.9368	0.9701	0.8736
NNSR	7.2791	0.9372	0.9715	0.8745
CFM	7.2836	0.9408	0.9718	0.8767
FRL-PCNN	7.2914	0.9412	0.9724	0.8771
Proposed	**7.2988**	**0.9477**	**0.9783**	**0.8859**

Table 6: The objective metrics of different fusion approaches for image "temple".

Method	EN	Q_W	CC	VIFF
DE	7.2563	0.9157	0.9687	0.8774
NFBD	7.2566	0.9188	0.9688	0.8793
GDMC	7.2569	0.9236	0.9702	0.8817
LRRW	7.2572	0.9237	0.9711	0.8824
NNSR	7.2574	0.9239	0.9714	0.8829
CFM	7.2579	0.9385	0.9718	0.8836
FRL-PCNN	7.2583	0.9427	0.9724	0.8867
Proposed	**7.2594**	**0.9507**	**0.9791**	**0.8958**

Tables 1–6 display the fusion objective evaluation results on 6 pairs of multifocus images with the eight fusion methods. As can be seen from the tables, the proposed fusion method has obvious advantages over other fusion methods in terms of the fusion indexes. In general, the proposed method achieves the best results in terms of Q_W, CC, EN, VIFF, and average accuracy index, indicating that this new algorithm is an effective fusion method.

4. Conclusions

In this paper, an end-to-end unsupervised multifocus image fusion algorithm based on sparse denoising autoencoder neural network is proposed. Combined with the prior knowledge of multifocus image, the network can learn accurate image details. Reasonable weight initialization and weight constraint are designed in the fusion layer. Local structure similarity and local mean square error strategies are used in the loss function to drive the fusion unit to learn the fusion rules effectively. Experimental results show that the proposed method not only can realize the fusion rules in the fusion process of self-learning. In addition, good results can be obtained in subjective vision and objective evaluation. It is of great significance to further understand the multifocus image fusion mechanism based on deep learning and to study the general multi-modal image fusion framework. In the future, more newest deep learning methods will be utilized to analyze the multifocus image fusion.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] W. Zhao, H. Lu, and D. Wang, "Multisensor image fusion and enhancement in spectral total variation domain," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 866–879, 2018.

[2] S. Yin and Y. Zhang, "Singular value decomposition-based anisotropic diffusion for fusion of infrared and visible images," *International Journal of Image and Data Fusion*, vol. 10, no. 2, pp. 146–163, 2019.

[3] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multimodality image fusion method based on image decomposition and sparse representation," *Information Sciences*, vol. 432, pp. 516–529, 2018.

[4] L. Bungert, D. Coomes, M. Ehrhardt et al., "Blind image fusion for hyperspectral imaging with the directional total variation," *Inverse Problems*, vol. 34, no. 4, 2018.

[5] L. Zhao, T. Yang, J. Zhang, Z. Chen, Y. Yang, and Z. J. Wang, "Co-learning non-negative correlated and uncorrelated features for multi-view data," *IEEE Transactions on Neural Networks and Learning Systems*, p. 1, 2020.

[6] J. Li, B. Li, and Y. Jiang, "An infrared and visible image fusion algorithm based on LSWT-NSST," *IEEE Access*, vol. 8, pp. 179857–179880, 2020.

[7] J. Wang, C. Qin, and X. Zhang, "A multi-source image fusion algorithm based on gradient regularized convolution sparse representation," *Systems Engineering and Electronics*, vol. 31, no. 3, pp. 447–459, 2020.

[8] L. Dong, Q. Yang, H. Wu, H. Xiao, and M. Xu, "High quality multi-spectral and panchromatic image fusion technologies based on curvelet transform," *Neurocomputing*, vol. 159, pp. 268–274, 2015.

[9] M. Nazrudeen and M. Rajalakshmi, "CT and MRI image fusion using non-subsampled contourlet transform," in *Proceedings of the International Conference on Communication and Computer Networks of the Future*, Tokyo, Japan, June 2014.

[10] X. Wang, S. Yin, K. Sun et al., "KFC-CNN: modified Gaussian kernel fuzzy C-means and convolutional neural network for apple segmentation and recognition," *Journal of Applied Science and Engineering*, vol. 23, no. 3, pp. 555–561, 2020.

[11] M. J. Li, Y. B. Dong, and X. L. Wang, "Image fusion algorithm based on gradient pyramid and its performance evaluation," *Applied Mechanics and Materials*, vol. 525, pp. 715–718, 2014.

[12] J. Liang, Y. He, D. Liu, and X. Zeng, "Image fusion using higher order singular value decomposition," *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, vol. 21, no. 5, pp. 2898–2909, 2012.

[13] T. Wan, C. Zhu, and Z. Qin, "Multifocus image fusion based on robust principal component analysis," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1001–1008, 2013.

[14] L. Zhao, T. Zhao, T. Sun, Z. Liu, and Z. Chen, "Multi-view robust feature learning for data clustering," *IEEE Signal Processing Letters*, vol. 27, pp. 1750–1754, 2020.

[15] S. Yin, H. Li, and L. Teng, "Airport detection based on improved faster RCNN in large scale remote sensing images," *Sensing and Imaging*, vol. 21, no. 1, 2020.

[16] S. Yin and H. Li, "Hot region selection based on selective search and modified fuzzy C-means in remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5862–5871, 2020.

[17] P. Li, Z. Chen, L. T. Yang, Q. Zhang, and M. J. Deen, "Deep convolutional computation model for feature learning on big data in internet of things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 790–798, 2018.

[18] S. Karim, Y. Zhang, A. A. Laghari, and M. R. Asif, "Image processing based proposed drone for detecting and controlling street crimes," in *Proceedings of the 2017 IEEE 17th International Conference on Communication Technology (ICCT)*, pp. 1725–1730, Chengdu, China, October 2017.

[19] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[20] H. Tang, B. Xiao, W. Li, and G. Wang, "Pixel convolutional neural network for multi-focus image fusion," *Information Sciences*, vol. 433-434, pp. 125–141, 2018.

[21] X. Luo, J. Zhang, and Q. Dai, "A regional image fusion based on similarity characteristics," *Signal Processing*, vol. 92, no. 5, pp. 1268–1280, 2012.

[22] A. A. Laghari, H. He, M. Shafiq, and A. Khan, "Assessment of quality of experience (QoE) of image compression in social cloud computing," *Multiagent and Grid Systems*, vol. 14, no. 2, pp. 125–143, 2018.

[23] L. Zhao, C. Mo, T. Sun, and W. Huang, "Aero engine gas-path fault diagnose based on multimodal deep neural networks," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8891595, 10 pages, 2020.

[24] L. Zhang, G. Zeng, and J. Wei, "Adaptive region-segmentation multi-focus image fusion based on differential evolution," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 3, 2019.

[25] Y. Yang, Y. Zhang, J. Wu, L. Li, and S. Huang, "Multi-focus image fusion based on a non-fixed-base dictionary and multi-measure optimization," *IEEE Access*, vol. 7, pp. 46376–46388, 2019.

[26] 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, 2016.

[27] Z. Ji, X. Kang, K. Zhang, P. Duan, and Q. Hao, "A two-stage multi-focus image fusion framework robust to image misregistration," *IEEE Access*, vol. 7, pp. 123231–123243, 2019.

[28] Q. Zhang, G. Li, Y. Cao et al., "Multi-focus image fusion based on non-negative sparse representation and patch-level consistency rectification," *Pattern Recognition*, vol. 104, Article ID 107325, 2020.

[29] L. He, X. Yang, L. Lu et al., "A novel multi-focus image fusion method for improving imaging systems by using cascade-forest model," *EURASIP Journal on Image and Video Processing*, vol. 5, no. 1, p. 2020, 2020.

[30] K. He, D. Zhou, X. Zhang et al., "Multi-focus image fusion combining focus-region-level partition and pulse-coupled neural network," *Soft Computing—A Fusion of Foundations, Methodologies and Applications*, vol. 23, no. 13, 2019.

[31] A. Laghari, H. He, A. Khan, and S. Karim, "Impact of video file format on quality of experience (QoE) of multimedia content," *3D Research*, vol. 9, no. 3, p. 39, 2018.

[32] K. Shahid, Y. Zhang, S. Yin, A. Laghari, and A. Brohi, "Impact of compressed and down-scaled training images on vehicle detection in remote sensing imagery," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 32565–32583, 2019.

[33] H. T. Mustafa, J. Yang, and M. Zareapoor, "Multi-scale convolutional neural network for multi-focus image fusion," *Image and Vision Computing*, vol. 85, pp. 26–35, 2019.