A Multi-focus Image Fusion Network Deployed in Smart City Target Detection

Haojie Zhao¹, Shuang Guo¹, Gwanggil Jeon², and Xiaomin Yang¹

¹Sichuan University College of Electronics and Information Engineering
²Incheon National University

June 25, 2024

Abstract

In the global monitoring of smart cities, the demands of global object detection systems based on cloud and fog computing in intelligent systems can be satisfied by photographs with globally recognized properties. Nevertheless, conventional techniques are constrained by the imaging depth of field and can produce artifacts or indistinct borders, which can be disastrous for accurately detecting the object. In light of this, this paper proposes an artificial intelligence-based gradient learning network that gathers and enhances domain information at different sizes in order to produce globally focused fusion results. Gradient features, which provide a lot of boundary information, can eliminate the problem of border artifacts and blur in multi-focus fusion. The multiple-receptive module (MRM) facilitates effective information sharing and enables the capture of object properties at different scales. In addition, with the assistance of the global enhancement module (GEM), the network can effectively combine the scale features and gradient data from various receptive fields and reinforce the features to provide precise decision maps. Numerous experiments have demonstrated that our approach outperforms the seven most sophisticated algorithms currently in use.
A Multi-focus Image Fusion Network Deployed in Smart City Target Detection

Haojie Zhao | Shuang Guo | Gwanggil Jeon | Xiaomin Yang

1 College of Electronics and Information Engineering, Sichuan University, Chengdu, China
2 Embedded Systems Engineering, Incheon National University, Incheon 22012, Republic of Korea

Abstract

In the global monitoring of smart cities, the demands of global object detection systems based on cloud and fog computing in intelligent systems can be satisfied by photographs with globally recognized properties. Nevertheless, conventional techniques are constrained by the imaging depth of field and can produce artifacts or indistinct borders, which can be disastrous for accurately detecting the object. In light of this, this paper proposes an artificial intelligence-based gradient learning network that gathers and enhances domain information at different sizes in order to produce globally focused fusion results. Gradient features, which provide a lot of boundary information, can eliminate the problem of border artifacts and blur in multi-focus fusion. The multiple-receptive module (MRM) facilitates effective information sharing and enables the capture of object properties at different scales. In addition, with the assistance of the global enhancement module (GEM), the network can effectively combine the scale features and gradient data from various receptive fields and reinforce the features to provide precise decision maps. Numerous experiments have demonstrated that our approach outperforms the seven most sophisticated algorithms currently in use.

KEYWORDS

Multi-focus image fusion, Cloud and Fog computing, Smart city, Artificial intelligence, Gradient information

1 INTRODUCTION

Optical equipment often has a restricted depth of field, which means that only items that are visible within it will be sharply focused while anything outside of it will be blurry. The limitations of such hardware devices prompted the development of multi-focus fusion technology. When humans view areas of the source image that are out of focus, they become blurry. We wouldn’t be able to perceive a clear image unless all of the focus information from the source image is combined. A global focus attribute image is ultimately generated by extracting the focus characteristics from several images, assessing, and preserving the focused attributes. In previous decades, two types of traditional approaches for this research have been proposed: spatial domain methods and transform domain methods.

The spatial domain approach means that the original image is directly subjected to feature extraction in the spatial domain to determine the activity level measurements, and the original image is fused using specific fusion rules. The method initially constructs an initial decision map containing focus properties for each pixel. Some post-processing procedures, such as removing small holes, are typically utilized to fix some misclassification areas of the decision map. Finally, a decision map that has been optimized is obtained. The spatial domain approach can maintain object texture characteristics to an extensive extent, but it is prone to misclassifying the focus attributes of pixels at the focus and defocus area boundary regions.

The transform domain approach entails transforming the original image into a different specific domain for analysis operations, creating fusion rules to combine the decomposition coefficients that can represent the activity level measurements from various source images, and then reversing the fusion rules to obtain the fused image. Unsuitable fusion rules and decomposition coefficients can lead to unsatisfactory fusion results. Fused images frequently fail to effectively handle image detail information.
resulting in pixel intensity inaccuracies. Additionally, transform domain approaches typically necessitate extensive computation and processing.

Deep learning (DL)-based methods are frequently employed due to the advantages provided by convolutional neural networks (CNNs) over traditional methods for image feature extraction. According to the outcome of the network, the methods typically divide into two categories: regression-based networks and classification-based networks. Regression-based networks input source images with different focus attributes and then generate fused images end-to-end. The process can frequently be separated into three steps: feature extraction, feature fusion, and feature reconstruction. Classification-based networks, on the other hand, convert the fusion problems into classification or segmentation issues. This research applies the classification approach to the creation of a deep learning network. This method creates decision maps by analyzing the focus attributes of each image block. For the purpose of generating the fused image, the decision maps are employed to choose the relevant source image region. We train on three datasets including gradient information to ensure that reliable decision maps are generated. The multiple-receptive field block is utilized to capture features at various levels. Meanwhile, the global enhancement block is used to enrich specific learned features and extract global features, compensating for the fact that convolution is limited to extracting local features. The combination of the two blocks is capable of accurately identifying the focusing attributes of the input image block and providing output labels. This paper’s key contributions are comprised of the three points listed below:

1. Gradient information in multiple directions and source images are used to create three different sorts of datasets. The network is trained with the dataset, which can significantly increase the classification accuracy.
2. We design the multiple-receptive field global enhancement residual block to interactively aggregate the scale features.
3. We conducted extensive trials on the Lytro dataset, which demonstrates that our method could yield favorable results in both subjective and objective outcomes.

The remainder of this essay is structured as follows. The relevant work is presented in Section 2. The network structure and implementation details suggested in this article are described thoroughly in Section 3. The findings of the experiment will be discussed in Section 4. And finally, We conclude the article in section 5.

2 | RELATED WORKS

2.1 | Traditional methods

Conventional approaches primarily focus on executing transform domain fusion and spatial domain fusion while performing image fusion. Among the transform domain approaches, Burt et al. developed the pyramid-based method (Burt, P., Adelson, E., 1983) in 1983. The pyramid-based method constructs source and target images into image pyramids respectively. This means
that the original image is continuously downsampled to generate a sequence of images at various scales. The weight coefficients for each scale sub-image for fusion are then determined starting at the bottom of the pyramid. Subsequently, a classical algorithm called the wavelet transform-based method (H. Li, Manjunath, & Mitra, 1995; Hill, 2003) emerged. The wavelet transform-based method decomposes the original image at several scales using wavelet transform to extract the coefficients of different frequencies. A new sub-image is created by recombining the sub-images that were produced by decomposing at various scales. Ultimately, wavelet reconstruction is employed to create the fused image from the new sub-image. Yang and Li were the first to present SR-based techniques (B. Yang & Li, 2010) for multi-focus fusion, training an overcomplete dictionary to approximate the input signal. The input signal is subsequently divided into a variety of linear basis vector combinations, which are then fused in the sparse domain to form new basis vectors. The fused image is ultimately created by linearly combining the new basis vectors.

Spatial domain methods can be further refined into pixel-level-based fusion, region-level-based fusion, and image block-based fusion. Pixel-level-based fusion weights the pixels of two images with fusion one by one and adjusts them according to the position of the pixel point, gray value, gradient, and other factors to obtain a pixel decision map. The classical approaches include the GFF method (S. Li, Kang, & Hu, 2013) and the DSIFT method (Liu, Liu, & Wang, 2015). The GFF method obtains the detail and base layers by mean filtering the original image. The weights are then reconstructed by guided filtering. The DSIFT method generates a focus attribute characterization operator by localizing a window of pixel points. Region-level-based fusion splits the image and performs a weighted fusion operation for each segmented region. Image block-based fusion selects each patch from the source image and assigns specific weights for fusion. Numerous improved methods have been developed, including the ideal image block size-based method (Kurban, 2010) and the quadrilateral-based method (De & Chanda, 2013; Bai, Zhang, Zhou, & Xue, 2015), since the shape of the image block may alter the quality of the boundaries in the fusion output. The retrieved features are utilized to represent the source image activity level measurement to construct an initial decision map in these methods. For decision map adjustment, some algorithms are commonly employed. Contrary to transform domain methods, which introduce pixel errors, spatial domain methods select the pertinent pixels.

2.2 Deep Learning Methods

Convolutional neural networks are especially effective at feature extraction, which explains why DL techniques have gained popularity in the field of multi-focus image fusion. Similar to traditional methods, DL can be mainly categorized into regression-based networks and classification-based networks.

The most cutting-edge techniques for regression-based networks are listed below. A regression network that collects multilevel features was suggested by Zhao et al. (Zhao, Wang, & Lu, 2018). Zhang et al. (Y. Zhang et al., 2020) employed a two-way convolutional network with shared weights. They performed the convolution process after fusing the shallow 64-channel features.
Li et al. (J. Li, Guo, Lu, Zhang, & Zhang, 2020) generated the fused image directly from the mask by retaining the initial shallow features of large sensory fields through multilevel residual blocks and generating the initial mask through the network without further post-processing algorithms. Zhang et al. (A, A, B, A, & A, 2021) suggested a network that employs a densely linked generator with interactive information to generate fused images. To continuously refresh the network, they measure the difference between the maximum gradient information of the source picture and the Laplace features of the initial fused image.

Liu et al. (Liu, Chen, Peng, & Wang, 2017) employed convolutional networks to extract features, addressing issues such as the complexity of traditional approaches’ architecture. They implemented a direct transformation from images to decision maps via deep learning CNN models, which complete the tasks of activity level measurement and fusion rule design. Jump connections were employed by Yang et al. (Y. Yang, Nie, Huang, Lin, & Wu, 2019) to merge multi-layer features by fusing shallow features with up-sampled features. Ma et al. (B. Ma et al., 2021) employed an autoencoder with spatial compression, a channel-excited attention mechanism, paired with spatial frequency for encoding and decoding. Based on this, they improved the method and proposed a GACN network (B. Ma, Yin, Wu, & Ban, 2020) by adding channel compression and spatially motivated attention to the convolutional end of the decoder. Ma et al. (J. Ma, Le, Tian, & Jiang, 2021) created an initial binary mask using a densely connected encoder and employed supervised training to continuously modify the mask. Liu et al. (Liu, Wang, Cheng, & Chen, 2021) interactively fused multi-scale features by applying the coordinate attention to fuse different resolution features and train the network with the Wbce loss function. Gao et al. (Gao, Yu, Tan, & Yang, 2022) fused different scale features by inception modules for feature enhancement and adding gradient information to intermediate features. To increase algorithm robustness, Ma et al. (L. Ma et al., 2023) provide the source and initial focusing probability maps into a network for classification. Liu et al. (Liu, Wang, Li, & Chen, 2022) used multi-scale features combined with spatial and channel dual-attention models to generate an initial fused image. Notably, by running the original fused image through an encoder that creates a multi-channel feature map, the method incorporates the benefits and drawbacks of both regression-based and classification-based networks. The combining of elements at different scales can enrich the detailed information. The learned multi-scale features can be effectively enhanced through the compression and excitation module (H. Zhang, Zu, Lu, Zou, & Meng, 2021). The accuracy of network categorization can be improved through gradient features in several directions. Residual linking speeds up the training process and preserves

---

**Figure 3** The specific structure of the basic unit MRFGERB.

**Table 1** The specific convolution parameters of MRFGERB. K, P, G, S, I_C, and O_C are the size of the convolution, padding, groups, step size, input channels, and output channels respectively.

<table>
<thead>
<tr>
<th>Module</th>
<th>K</th>
<th>P</th>
<th>G</th>
<th>S</th>
<th>I_C</th>
<th>O_C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Conv2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>Conv3</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>Conv4</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>Conv5</td>
<td>9</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>Conv6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>Conv7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>32</td>
</tr>
</tbody>
</table>
shallow features. Based on the discussion above, this research suggests the multiple-receptive field global enhancement residual block as the fundamental constructional element of the deep learning network. We employed three distinct types of gradient datasets for network training.

3 | PROPOSED METHOD

We explicitly present the suggested approach in this part, emphasizing three elements: the network, the training specifics, and the fusion approach. The entire network flowchart is displayed in Fig 1. The initial stage is to send two images with different focus attributes into the network, and the network will output distinct labels, which we refer to as the focused attribute map. In the second stage, we binarize this map using the moving window approach. The smoothing procedure is utilized to handle jagged edges. Ultimately, pre-established rules are employed to create the fused image.

3.1 | Network structure

The overall network architecture for this approach is illustrated in Fig 2. For network feeding, the source picture is split into three branches. Initially, a gray-scale version of the source image is created in the first branch. The grayscale image blocks get fed into a 3 × 3 convolution to increase the number of channels to 32 and extract image features. The features are subsequently enhanced by three multiple-receptive field global enhancement residual blocks (MRFGERB), with the number of channels remaining constant throughout. At the end of the first branch, we modify the number of channels to 64 channels by a 3 × 3 convolution. In addition, we repeat the preceding method for gradient information. However, we do not perform a second convolution to modify the channels. Then, two feature maps with gradient information are combined. Finally, the feature map with 64 channels is obtained. The second and third branch output feature maps are concatenated with the preceding 64-channel feature map containing the source image information. At last, the classification is performed by a linear connection layer to determine the focusing attribute.

The specific structure of MRFGERB is depicted in Fig 3. It primarily comprises a multiple-receptive field block and a global enhancement block. The multiple-receptive field block can be utilized to extract the scale features. Convolution of various sensory fields can better catch items between focused and unfocused regions, allowing the model more robust. The global enhancement block then further improves the extracted features. Additionally, the original shallow features can be further preserved and the training process sped up by using residual concatenation. Initially, MRFGERB performs a 3×3 convolution. The multi-scale receptive field block then extracts multi-scale features by employing four convolutions of sizes 3, 5, 7, and 9, with groups 1, 2, 4, and 8, respectively. The specific parameters of all the convolutions are given in Table 1. The global enhancement block connects four different scale features and then applies varying channel attention weights to the respective scale features. The following features with channel attention are processed using a softmax activation function. Finally, it is multiplied by the multi-scale features which were previously connected. The global enhancement block can efficiently assist the multiple-receptive field block in the extraction of multi-scale information. The global enhancement block first compresses the global spatial data into the channel dimensions using global average pooling and then performs dimensionality reduction and dimensionality enhancement operations. Convolution, as is well known, can only extract local features. The global enhancement block that we utilize here is particularly effective at extracting global features. The combination of global and local features successfully overcomes the convolution weakness and increases model robustness. Fig 4 shows the detailed structure of the global enhancement block.

**FIGURE 4** The architecture of the SEblock. The GAP is a global average pooling operation.
3.2 Training details

This paper’s dataset is obtained from VOC2012 and contains 3200 randomly selected images, which are converted to grayscale versions. Among these, 2,000 images are utilized to form our training set, and 1,200 images are used to construct the test set. Four different types of Gaussian filters with various kernel sizes are used to create four images with various degrees of blurring in order to imitate the focussing and defocusing conditions in camera photos. In each blurred image, a square patch with a side length of 32 is subsequently selected. To build a new image patch of size $32 \times 64$, the original image patch is picked at the same spot and vertically attached to each of the four $32 \times 32$ image patches with varying blur values.

The generated new image patches are fed into the network to generate labels. We also extract the grayscale image’s horizontal and vertical gradient information. As previously explained, several image patches are created for the gradient images. Finally, we produced 300,000 image patches for testing and 1,000,000 for training. Our network is trained using stochastic gradient descent (SGD). The step size of the stepLR scheduler is 1, the gamma value is 0.9, and the starting value of the learning rate is set to 0.0002. The weight decay is 0.0005 and the momentum is 0.9. The cross-entropy loss function is utilized as an assessment criterion since this work tackles the multi-focus fusion problem as an issue of classification. To train the network, we use an NVIDIA GTX 2080 Ti. The training epoch is set to 50, while the training batch size is set to 256.

3.3 Fusion scheme

The initial two source images are designated as $I_A$ and $I_B$, and specific regions of $I_A$ are focused while those same regions in $I_B$ are defocused. When they are fed into the trained network as shown in fig[4] it will return two labels, 0 and 1, where 0 denotes a greater emphasis on $I_A$ and a lesser emphasis on $I_B$, and 1 denotes the exact opposite. This series of labels will subsequently
FIGURE 6  A collection of images titled "Lytro-06". (a) is the foreground-focused image. (b) is the background-focused image. (c)-(j) are the fusion results derived from GFF, NSCT-SR, CNN, MADCNN, SESF, DRLFPD, ECNN, and our algorithm.

FIGURE 7  A collection of images titled "Lytro-08". (a) is the foreground-focused image. (b) is the background-focused image. (c)-(j) are the fusion results derived from GFF, NSCT-SR, CNN, MADCNN, SESF, DRLFPD, ECNN, and our algorithm.

construct a focus attribute map, which will be binarized employing the moving window strategy, as represented by the formula in (1).

\[
M(x, y) = \begin{cases} 
M(x : x + k, y : y + k) + = 1, & \text{label} = 0 \\
M(x : x + k, y : y + k) + = -1, & \text{label} = 1
\end{cases}
\]  

(1)
A collection of images titled "Lytro-12". (a) is the foreground-focused image. (b) is the background-focused image. (c)-(j) are the fusion results derived from GFF, NSCT-SR, CNN, MADCNN, SESF, DRLFPD, ECNN, and our algorithm.

Table 2 Metrics results obtained by different algorithms on the "Lytro" dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>QMI</th>
<th>QNCIE</th>
<th>QY</th>
<th>QCB</th>
<th>QCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFF</td>
<td>1.0990(8)</td>
<td>0.8395(8)</td>
<td>0.9821(7)</td>
<td>0.7975(7)</td>
<td>16.0197(2)</td>
</tr>
<tr>
<td>NSCT-SR</td>
<td>1.1223(7)</td>
<td>0.8410(7)</td>
<td>0.9781(8)</td>
<td>0.7940(8)</td>
<td>16.0890(3)</td>
</tr>
<tr>
<td>CNN</td>
<td>1.1512(6)</td>
<td>0.8429(6)</td>
<td>0.9875(4)</td>
<td>0.8084(3)</td>
<td>16.2763(7)</td>
</tr>
<tr>
<td>MADCNN</td>
<td>1.1650(4)</td>
<td>0.8438(4)</td>
<td>0.9862(6)</td>
<td>0.8062(5)</td>
<td>16.2339(5)</td>
</tr>
<tr>
<td>SESF</td>
<td>1.1546(5)</td>
<td>0.8431(5)</td>
<td>0.9868(5)</td>
<td>0.8051(6)</td>
<td>16.3068(8)</td>
</tr>
<tr>
<td>DRLFPD</td>
<td>1.1730(3)</td>
<td>0.8444(3)</td>
<td>0.9877(3)</td>
<td>0.8083(4)</td>
<td><strong>15.9428(1)</strong></td>
</tr>
<tr>
<td>ECNN</td>
<td>1.1869(2)</td>
<td>0.8454(2)</td>
<td>0.9886(2)</td>
<td>0.8089(2)</td>
<td>16.2490(6)</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>1.1887(1)</strong></td>
<td><strong>0.8456(1)</strong></td>
<td><strong>0.9890(1)</strong></td>
<td><strong>0.8093(1)</strong></td>
<td>16.1771(4)</td>
</tr>
</tbody>
</table>

where x and y denote the rows and columns of the source image, M denotes the output focus attribute map, and k is a constant with a value of 32. We can obtain a decision map with focus attributes for each pixel based on the binarized focus attribute map in the following way.

\[
D(x, y) = \begin{cases} 
1, & M(x, y) > 0 \\
0, & \text{else}
\end{cases}
\]  

(2)

where D denotes the decision map. We compute the fused image in the following way defined as

\[
I(x, y) = I_A(x, y)D(x, y) + I_B(x, y)(1 - D(x, y))
\]  

(3)

Fig. illustrates the six types of images we were capable of capturing throughout the process.

4 EXPERIMENTS

4.1 Evaluation metrics

Aside from the subjective visual effect, professional indicators that assess image quality can also be utilized to determine an image's fusion effect. To assess the accuracy of the fused images quantitatively, we present the following five metrics. (1) \[Q_{MI}\] (Hossny, Nahavandi, & Creighton, 2008): utilized to evaluate how well the original information has been incorporated into the
FIGURE 9 Four different structures designed in the ablation experiment.

Fused image. (2) $Q_{NCIE}$ (Wang, Shen, & Jin, 2008): $Q_{NCIE}$ stands for nonlinear correlation information entropy, which describes the relationship between the fusion outcome and the initial images. (3) $Q_Y$ (C. Yang, Zhang, Wang, & Liu, 2008): utilized to determine how much of the structure-related details have been preserved. (4) $Q_{CB}$ (Y. Chen & Blum, 2007): utilized to measure the extent to which the fused image has retained information from the original image. (5) $Q_{CV}$ (H. Chen & Varshney, 2007): utilized to assess how much data from the two source images is incorporated.

4.2 Seven methods employed for comparative experiments

We compare the proposed method to seven advanced multi-focus fusion algorithms, including the NSCT-SR method (Q. Zhang & long Guo, 2009), GFF method (S. Li et al., 2013), MADCNN method (Lai, Li, Guan, & Xiong, 2019), CNN method (Liu et al., 2017), SESF method (B. Ma et al., 2021), DRLFPD method (Liu et al., 2022), and ECNN method (Amin-Naji, Aghagolzadeh & Ezoji, 2019), which confirm that it is indeed effective. We use the default parameters from each method paper to get the appropriate image fusion results to assure the fairness of the comparison. The "Lytro" dataset (Nejati, Samavi, & Shirani, 2015) was utilized for our test experiment, which includes 20 different pairs of color images.

4.2.1 Subjective visual comparison

We focus on several boundary-localized regions of the fusion result to assess the efficacy of the method suggested in this paper. Fig 6 displays two "locked" source images together with the fusion results obtained via multiple methods. We are able to observe that the NSCT-SR method fuses the image with the loss of the green background features at the boundary. The GFF, CNN, and SESF methods suffer from boundary blurring. The MADCNN method shows some undesired artifacts in the fusion. The DRLFPD method transforms the background color of the fusion result from green to yellowish. Additionally, the learned features of the ECNN approach contain considerable mistakes.

Fig 7 depicts the source photos of two "sculptures" as well as the fusion results obtained via seven distinct methods. To see the fusion effect of the fused images, we concentrate on the edges of the sculptures’ noses. The NSCT-SR method, DRLFPD method, and SESF method suffer from loss of details. The CNN method and the GFF approach undergo boundary blurring. The MADCNN method generates some artifacts. The ECNN method incorrectly produces jagged features.

Fig 8 displays the fusion results of two source photos from "Lytro-12" and other cutting-edge algorithms in addition to the approach used in this paper. In the labeled regions, the fusion results of the NSCT-SR, GFF, CNN, SESF, and ECNN methods
**FIGURE 10** A collection of images titled "Lytro-06". (a) is the foreground-focused image. (b) is the background-focused image. The results from four distinct modules, RCB, GERB, MRFGERB, and MRFGERB, are shown in (c)-(f). The differential images produced by four distinct results are displayed in (g)-(j).

**FIGURE 11** A collection of images titled "Lytro-07". (a) is the foreground-focused image. (b) is the background-focused image. The results from four distinct modules, RCB, GERB, MRFGERB, and MRFGERB, are shown in (c)-(f). The differential images produced by four distinct results are displayed in (g)-(j).

are all fuzzy at the signboard boundaries, while the MADCNN method exhibits some boundary artifacts and the DRLFDPD method exhibits color distortions.
Table 3: Objective indicators of fusion results obtained by different methods in ablation experiment.

<table>
<thead>
<tr>
<th>Methods</th>
<th>QMI</th>
<th>QNCIE</th>
<th>QY</th>
<th>QCB</th>
<th>QCv</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCB</td>
<td>1.1865</td>
<td>0.8454</td>
<td>0.9887</td>
<td>0.8085</td>
<td>16.8470</td>
</tr>
<tr>
<td>GERB</td>
<td>1.1885</td>
<td>0.8456</td>
<td>0.9889</td>
<td>0.8091</td>
<td>16.8283</td>
</tr>
<tr>
<td>MRFRB</td>
<td>1.1886</td>
<td>0.8456</td>
<td>0.9890</td>
<td>0.8093</td>
<td>16.6033</td>
</tr>
<tr>
<td>MRFGERB</td>
<td>1.1887</td>
<td>0.8456</td>
<td>0.9890</td>
<td>0.8093</td>
<td>16.1771</td>
</tr>
</tbody>
</table>

Table 4: Average time for different methods to process a 520 x 520 pixel image.

<table>
<thead>
<tr>
<th>Method</th>
<th>Times/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFF</td>
<td>0.67</td>
</tr>
<tr>
<td>NSCT-SR</td>
<td>159.93</td>
</tr>
<tr>
<td>CNN</td>
<td>139.47</td>
</tr>
<tr>
<td>MADCNN</td>
<td>6.02</td>
</tr>
<tr>
<td>SESF</td>
<td>0.74</td>
</tr>
<tr>
<td>DRLFPD</td>
<td>2.47</td>
</tr>
<tr>
<td>ECNN</td>
<td>163.02</td>
</tr>
<tr>
<td>Ours</td>
<td>434.16</td>
</tr>
</tbody>
</table>

4.2.2 Objective metrics comparison

We use the "Lytro" dataset to test several fusion methods. The average results for all fusion methods are displayed in Table 2. The top score for each index is denoted in bold, and the value in parenthesis reflects how the approach ranks overall. The table shows that, with the exception of the $Q_{cv}$ indicator, the strategy we proposed ranks top for all other indicators. In conclusion, the method used in this study yields results that are both qualitatively and numerically more competitive.

4.2.3 Average running time

On the "Lytro" dataset, we recorded the average running time of the various approaches. According to Table 4, there are significant differences in processing times for 520 x 520-pixel images when using various techniques. We employed an Intel Core i5-6400 CPU and Matlab software to create image fusion for the GFF, NSCT-SR, and CNN methods. We used the Nvidia GTX 2080 Ti GPU and PyTorch for image fusion when employing other deep-learning techniques. It is essential to emphasize that the long implementation time of the methods in this work and ECNN is due to the excessive quantity of convolutional computation and the exceptionally high number of input parameters in the linear connection layer. The network becomes even more complex when combined with the multiple-receptive field global enhancement residual block, but the fusion results are noticeably enhanced. Therefore, the method suggested in this study can deliver high-quality fusion results without taking the fusion time cost and storage cost into account.

4.3 Ablation experiment

We explore the function of the multiple-receptive field global enhancement residual block, which is composed primarily of the multiple-receptive field block and the global enhancement block. The ablation experiments in this section are structured as follows: removing the two blocks together (see Fig. 9a, residual convolution block, denoted as RCB). Substitute a normal convolution block for the multiple-receptive field block (see Fig. 9b, global enhancement residual block, GERB). Deleting the global enhancement block (see Fig. 9c, multiple-receptive field residual block, MRFRB). The complete multiple-receptive field global enhancement residual block (MRFGERB) is shown in Fig. 9d, which displays the average indicator scores following the removal of the separate modules, the greatest values are indicated in bold.

Fig. 10 displays the fusion results of the ablation experiment for the two source images from "Lytro-06". We chose to observe images of their respective differences from the background. We can observe from the different images that the algorithm in this study is better able to recover the slabs’ edge contours and that there is no feature loss for the corners of the square slabs.
Fig. 11 displays two "Lytro-07" source images together with the combined results of the ablation experiment. The background-focused source image is subtracted to create the differential image, and it is clear that the method described in this work is better able to recover the character's arm contour. In addition, the specific details of the dress in the green box have been preserved.

5 CONCLUSION

This research proposes a multi-focus image fusion network based on DL. Boundary artifacts and blurring issues can be resolved effectively through the implementation of text approaches on intelligent systems. To enhance classification performance, the network incorporates information about image gradients and interactively aggregates features at several scales using the multi-receptive field global enhancement residual blocks. The multi-receptive module improves boundary details by capturing various scale features to enhance information sharing. Better process fusion outcomes can be achieved by enhancing the learned multi-scale features utilizing global enhancement block. It has been demonstrated through numerous experiments that the suggested method can get satisfactory results on the "Lytro" dataset. We notice that the focus attribute remains difficult to discern at some pixel positions in the focus attribute map. It would be preferable to apply a more appropriate method to acquire the fused image for the pixel region with a value of 0. Future studies will be optimized in this manner.

ACKNOWLEDGMENTS

This work was partially supported by the funding from the Research Result of the Key Laboratory of Ultra HD Video Technology Application in Fusion Publishing, National Press and Publication Administration UHD-ZD-202306.

DATA AVAILABILITY STATEMENT

The data are available on request.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

References


