Neural Computing and Applications
https://doi.org/10.1007/s00521-020-05358-9

ORIGINAL ARTICLE

SESF-Fuse: an unsupervised deep model for multi-focus image fusion

Boyuan Ma[1,2,3] · Yu Zhu[1,2,3] · Xiang Yin[1,2,3] · Xiaojuan Ban[1,2,3] · Haiyou Huang[1,4] · Michele Mukeshimana[5]

Received: 12 May 2020 / Accepted: 8 September 2020

Abstract
Muti-focus image fusion is the extraction of focused regions from different images to create one all-in-focus fused image. The key point is that only objects within the depth-of-field have a sharp appearance in the photograph, while other objects are likely to be blurred. We propose an unsupervised deep learning model for multi-focus image fusion. We train an encoder–decoder network in an unsupervised manner to acquire deep features of input images. Then, we utilize spatial frequency, a gradient-based method to measure sharp variation from these deep features, to reflect activity levels. We apply some consistency verification methods to adjust the decision map and draw out the fused result. Our method analyzes sharp appearances in deep features instead of original images, which can be seen as another success story of unsupervised learning in image processing. Experimental results demonstrate that the proposed method achieves state-of-the-art fusion performance compared to 16 fusion methods in objective and subjective assessments, especially in gradient-based fusion metrics.

Keywords Multi-focus image fusion · Unsupervised deep learning · Spatial frequency

1 Introduction

Multi-focus image fusion is an important issue in image processing. Optical lenses have the limitation that only objects within the depth-of-field (DOF) have a sharp appearance in a photograph, while other objects are likely to be blurred. Hence, it is difficult for objects at varying distances to all be in focus in one camera shot [20]. Many algorithms have been designed to create an all-in-focus image by fusing multiple images that capture the same scene with different focus points. The fused image can be used for human visualization or computer processing, such as feature extraction, segmentation, or object recognition.

Deep learning has had great success in image processing, and some multi-fusion methods based on a convolutional neural network (CNN) have been proposed. A supervised CNN-based multi-focus image fusion method used a Gaussian filter to generate synthetic images with different blurred levels to train a two-class image classification network [26]. With this supervised learning strategy, the network could distinguish whether a patch was in focus. DeepFuse [32] was developed in an unsupervised manner to fuse multi-exposure images. DenseFuse [15] fuses infrared and visible images using an unsupervised encoder–decoder network to extract deep features. An l1-norm fusion strategy fuses two feature maps, and the decoder uses fused features to obtain a fused image. The basic assumption is that the l1-norm of the feature vector for each node represents its activity level. It can fuse infrared and visible images. However, for multi-focus tasks, it is commonly assumed that only objects within the DOF have a sharp appearance in a photograph, while others are likely

✉ Xiaojuan Ban
banxj@ustb.edu.cn

✉ Haiyou Huang
huanghy@mater.ustb.edu.cn

1 Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing, China

2 Beijing Key Laboratory of Knowledge Engineering for Materials Science, University of Science and Technology Beijing, Beijing, China

3 Institute of Artificial Intelligence, School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China

4 Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing, China

5 Faculty of Engineering Sciences, University of Burundi, Bujumbura, Burundi

Published online: 21 September 2020

to be blurred [26]. Therefore, we assume that in the multi-focus task, what really matters is the feature gradient, not feature intensity.

We present a fusion method based on an unsupervised deep convolutional network. It used deep features extracted from an encoder–decoder network, and spatial frequency to measure activity levels. We conducted objective and subjective experiments which demonstrate that this method achieves state-of-the-art fusion performance (especially in gradient-based evaluation) compared to 16 classical fusion methods, and somewhat better than supervised methods.[1] Besides, this unsupervised strategy eliminates the need to manually generate a labeled multi-focus dataset, which is more practical in applications such as the internet of things (IoT).

The remainder of this paper is organized as follows. We briefly review related work in Sect. 2. Section 3 describes the proposed fusion method. Our experiments are described in Sect. 4. We discuss the future direction of this work in Sect. 5, and conclude the paper in Sect. 6.

2 Related work

Various methods of image fusion have been presented, which can be classified as either transform domain methods or spatial domain methods [39]. Classical transform domain fusion methods are based on the theories of the multi-scale transform, such as the Laplacian pyramid (LP) [2], ratio of low-pass pyramid (RP) [41], discrete wavelet transform (DWT) [16], dual tree complex wavelet transform (DTCWT) [14], curvelet transform (CVT) [29], non-subsampled contourlet transform (NSCT) [50], sparse representation (SR) [49], and image-matting-based fusion (IMF) [19]. The key is to use the decomposed coefficients from a selected transform domain to measure the activity level from source input images. Obviously, the selection of the transform domain plays an important role.

Spatial domain fusion methods measure activity levels based on gradient information. Early methods used a manually-fixed-size block strategy to calculate activity levels, such as spatial frequency [17], which usually causes undesirable artifacts. To handle this problem, V. Aslantas acquired an adaptive optimal block size for the decision map using a differential evolution algorithm [1]. Some pixel-wise gradient-based spatial domain algorithms have recently been proposed, such as guided filtering (GF) [18], multi-scale weighted gradient (MWG) [52], and dense SIFT (DSIFT) [25].

The CNN has achieved great success in image processing in the last five years. Some researchers have tried to measure the activity level by a high-capacity deep convolutional model. Yu Liu first applied a CNN in multi-focus image fusion [26]. Ram Prabhakar proposed a deep-learning-based unsupervised approach, DeepFuse [32], for the exposure fusion problem. A pixel-wise CNN was proposed to address the multi-focus image fusion problem [40]. The construction of the fused image was accelerated, and 12 kinds of artificial design masks were used to create the dataset. The algorithm included post-processing to modify the decision map. The training image has both focused and defocused parts, but is still much different from the real multi-focus image data due to the artificial design masks. A deep learning-based algorithm was proposed to handle different kinds of image fusion tasks [51]. An RGB image and depth image generated a large-scale multi-focus image dataset that was more similar to real multi-focus image data. The algorithm was designed in the end-to-end manner and trained on the proposed dataset. However, as it was designed as a general image fusion framework, it showed limited performance at fusing a specific type of image. MMFNet focused on the fusion quality near the focused/defocused boundary. It used a novel α-matte boundary defocus spread model to generate the dataset so as to simulate the real-world images, and a cascaded network was used to refine the focused/defocused boundary of the fusion image, with good results [27]. An unsupervised deep image fusion framework was proposed that utilized the structure tensor as the loss function to let the fused image preserve the overall contrast of the multi-channel input [12]. H. Li presented DenseFuse to fuse infrared and visible images [15], using an encoder–decoder unsupervised strategy to obtain useful features, and fused them by an l1-norm. In an infrared image, objects with thermal radiation will have more pixel intensity than a visible image. The key is the pixel intensity and not the pixel gradient, and the l1-norm is well suited to infrared-visible image fusion. If the object with thermal radiation is large, then this region will be totally white in an infrared image and totally black in a visible image, which will lead them to have the same gradient. In this situation, a gradient-based indicator, such as spatial frequency, cannot be used because it will cause detection error.

In our work, we train the network in an unsupervised encoder–decoder manner, and apply spatial frequency as a fusing rule to obtain the activity level and decision map of source images, which accords with the key assumption that only objects within the depth-of-field have a sharp appearance.

Following the pre-print version and public code, our algorithm has been extended to other datasets [47]. Some authors have improved upon this work and proposed some novel ideas to combine deep learning and gradient-based

[1] Experimental data and code can be found at https://github.com/Keep-Passion/SESF-Fuse.

methods to further advance the performance of multi-focus image fusion. Han Xu proposed a gradient and connected regions-based fusion method (GCF) [46], which used an encoder–decoder structure to output a gradient relation map to generate a decision map, using a deep learning model to implicitly calculate gradient information. In contrast, we use an encoder–decoder to output the original image, and spatial frequency to explicitly calculate the gradient information in deep features to generate a decision map. We speculate that our method is more robust than implicit decoder mapping. Jun Huang proposed gradient-based loss to train a generative adversarial network (GAN) fusion method [8]. Thanks to further experiments and the confirmation of the above work, SESF-Fuse achieves state-of-the-art fusion performance in gradient-based metrics.

3 Method

3.1 Overview of proposed method

We show the overall fusion strategy of our multi-fusion algorithm in Fig. 1. We train an encoder–decoder network to extract high-dimensional features in the training phase. We calculate the activity level using those deep features from the encoder in the inference phase, finally obtaining the decision map to fuse two multi-focus source images. Our algorithm only aims to fuse two source images. For multiple image fusion, it could straightforwardly fuse them, one by one, in series.

3.2 Extraction of deep features

We use encoder and decoder architecture to reconstruct the original input image and discard fusion operations in the training phase. After the network is trained and its trainable parameters are fixed, we use spatial frequency to calculate the activity levels of deep features drawn from the encoder in the inference phase.

As shown in Fig. 1, the encoder consists of a cascade of five convolutional layers, with the output of each connected to the other layers. This strengthens feature propagation and reduces the number of parameters [9]. The kernel number appears below the feature map. To precisely reconstruct the image, there are no pooling layers.

Squeeze and excitation (SE) has shown promising performance at image recognition and segmentation. It can effectively enhance spatial feature encoding by adaptively recalibrating channel-wise or spatial-feature responses [10]. We use an SE module to enhance the robustness and representativeness of deep features.

We use three versions of the SE module, spatial squeeze and channel excitation (sSE), channel squeeze and spatial excitation (cSE), and concurrent spatial and channel squeeze and channel excitation (scSE), which lead the network to learn more meaningful spatial and/or channel-wise feature maps [34]. sSE uses a convolutional layer with one 1×1 kernel to acquire a projection tensor. Each unit of the projection refers to the combined representation for all channels C at a spatial location, and is used to spatially recalibrate the original feature map. cSE uses a global average pooling layer to embed the global spatial information in a vector, which passes through two fully connected layers to acquire a new vector. This encodes the channel-wise dependencies, which can be used to recalibrate the original feature map in the channel direction. scSE is the element-wise addition of cSE and sSE, which concurrently recalibrate the spatial and channel-wise information of the input. The decoder consists of four convolutional layers, which are utilized to recover the input image.

The loss function L combines pixel loss L_{p} and structural similarity (SSIM) loss L_{ssim}, i.e.,

$$L = \lambda L_{\mathrm{ssim}} + L_{\mathrm{p}}, \tag{1}$$

and is minimized to train the encoder and decoder. λ is a constant weight to balance the importance of two losses.

The pixel loss L_{p} is the Euclidean distance between the output (O) and input (I),

$$L_{\mathrm{p}} = ||O - I||_2. \tag{2}$$

The SSIM loss L_{ssim} is the structural difference between O and I,

$$L_{\mathrm{ssim}} = 1 - \mathrm{SSIM}(O, I), \tag{3}$$

where

$$\mathrm{SSIM}(O,I) = \frac{2\mu_O\mu_I + C_1}{\mu_O^2 + \mu_I^2 + C_1} \cdot \frac{2\sigma_O\sigma_I + C_2}{\sigma_O^2 + \sigma_I^2 + C_2} \cdot \frac{\sigma_{O,I} + C_3}{\sigma_O^2\sigma_I^2 + C_3}, \tag{4}$$

where SSIM represents the structural similarity operation. Expressions of brightness, contrast, and structural similarity are on the right of the formula, from left to right; μ is the mean; and σ is the standard deviation. C_1, C_2, and C_3 are constants [43].

3.3 Fusion strategy

In the inference phase, we obtain two deep features of images from the encoder. We utilize spatial frequency to calculate an initial decision map, and we apply some common consistency verification methods to remove small errors. Finally, we obtain the decision map to fuse two multi-focus source images.

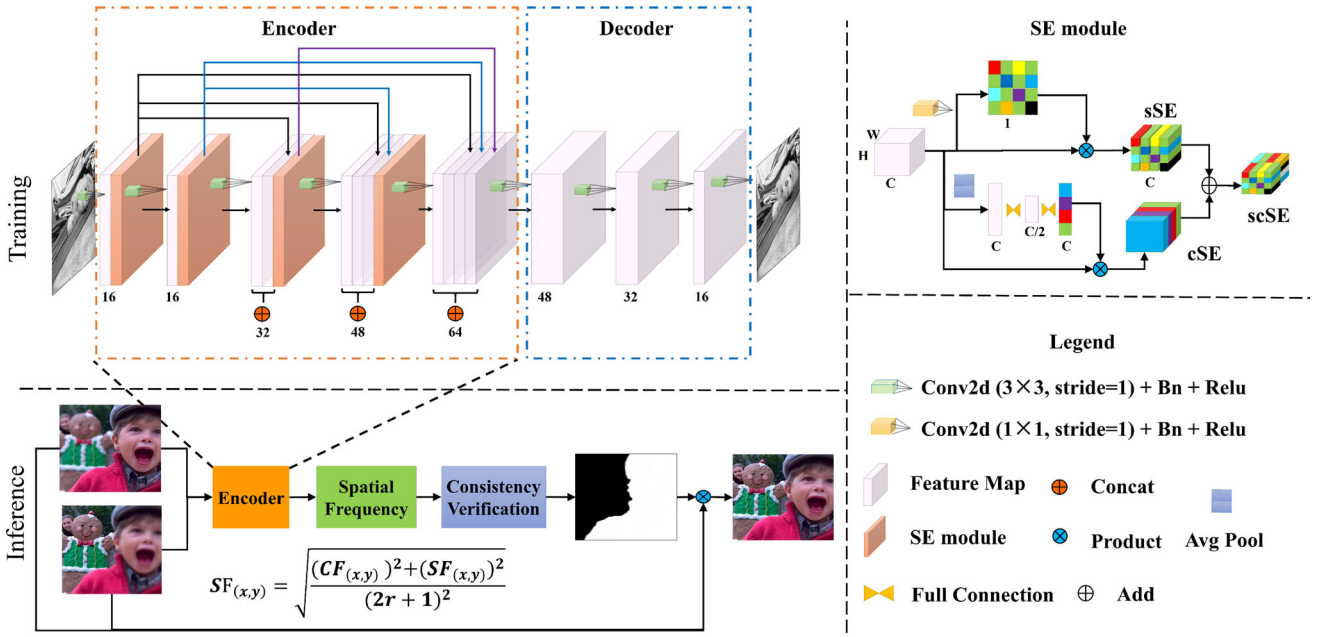

Fig. 1 Schematic diagram of proposed algorithm

3.3.1 Spatial frequency calculation using deep features

Different from the l1-norm in DenseFuse, we use the feature gradient instead of feature intensity to calculate an activity level. We apply spatial frequency to perform this task using deep features.

The encoder provides high-dimensional deep features for each pixel in an image. However, the original spatial frequency is calculated on a gray image with a single channel. We modify the spatial frequency calculation method for deep features. Let F be the deep features driven from the encoder block. $F_{(x,y)}$ is one feature vector, and (x, y) are the coordinates of these vectors in the image. We calculate its spatial frequency (SF) by

$$\mathrm{RF}_{(x,y)} = \sqrt{\sum_{a=-r}^{r}\sum_{b=-r}^{r}\left[F_{(x+a,y+b)} - F_{(x+a,y+b-1)}\right]^2} \tag{5}$$

$$\mathrm{CF}_{(x,y)} = \sqrt{\sum_{a=-r}^{r}\sum_{b=-r}^{r}\left[F_{(x+a,y+b)} - F_{(x+a-1,y+b)}\right]^2} \tag{6}$$

$$\mathrm{SF}_{(x,y)} = \sqrt{\frac{\left(\mathrm{CF}_{(x,y)}\right)^2 + \left(\mathrm{RF}_{(x,y)}\right)^2}{(2r+1)^2}}, \tag{7}$$

where RF and CF are, respectively, the row and column vector frequencies, and r is the kernel radius. The original spatial frequency is block-based, while it is pixel-based in our method. We apply the same padding strategy at the borders of feature maps.

Thus, we can compare the spatial frequencies of SF1 and SF2, where k in SFk is the index of the source image, and we obtain the initial decision map as

$$D_{(x,y)} = \begin{cases} 1, & \text{if } \mathrm{SF1}_{(x,y)} \geq \mathrm{SF2}_{(x,y)} \\ 0, & \text{otherwise} \end{cases}. \tag{8}$$

3.3.2 Consistency verification

Although the decision map calculated from deep features is robust enough to recognize the location of a focused region, there may be some small faults from inappropriate decisions caused by fluctuations, or noises in the image generated by the imaging system. For example, Gaussian or salt-and-pepper noise may randomly occur in pixels during imaging, which will influence the calculation of activity levels. Salt-and-pepper noise will seriously impact the calculation of spatial frequency. We propose some consistency verification methods. Morphology operations, such as opening and closing, and a small-region-removal strategy will eliminate incorrect focused pixels in the decision map.

We applied a morphology operator (opening and closing calculation) with a small disk structuring element [4] as a post-process to rectify the decision map. When the radius of the disk structuring element equals that of the spatial frequency kernel, the faults can be well detected and the adjacent regions can be connected correctly. For each channel of the binary decision map, we applied a small-region-removal strategy to reverse regions smaller than an area threshold ($0.01 \times H \times W$, where H and W are the

height and width, respectively, of the original input image) [26].

We utilized an efficient edge-preserving filter, called guided filter [7], to remove some undesirable artifacts around the boundaries between the focused and defocused regions. We used the initial fused image as the guidance filter for the initial decision map. We experimentally set a window radius $r = 4$ and regularization parameter $\varepsilon = 0.1$ in the guided filter algorithm.

3.3.3 Fusion

We drew out the fused result F using a pixel-wise weighted fusion rule,

$$F_{(x,y)} = D_{(x,y)} Img1_{(x,y)} + (1 - D_{(x,y)}) Img2_{(x,y)}, \quad (9)$$

where the input images are denoted as $Imgk$ and are pre-registered, and k is the index of the source images. The visualization of fused results is shown in Fig. 2.

4 Experiments

4.1 Experimental settings

We used 38 pairs of publicly available multi-focus images [28, 35] as a testing set for evaluation.

Due to the unsupervised strategy, we first trained the encoder–decoder network using MS-COCO [21], with 82783 images as a training set and 40504 images to validate the reconstruction ability in each epoch. All images were transformed to grayscale and resized to 256×256. Note that images were grayscale in the training phase, while images for testing could be grayscale or color images with RGB channels. For color images which needed to be fused, we transformed the images to grayscale and calculated a decision map to fuse them. The initial learning rate was 1×10^{-4}, and this was decreased by a factor of 0.8 at every two epochs. We set $\lambda = 3$ and optimized the objective function by Adam [13]. The batch size and number of epochs were 48 and 30, respectively. After training was done and the trainable parameters were fixed, we used these parameters to perform image fusion on the testing set.

Our implementation was derived from the publicly available Pytorch framework [5]. The network's training and testing were performed on a system using four NVIDIA 1080Ti GPUs with a total of 44 GB memory.

4.2 Objective image fusion quality metrics

The proposed fusion method was compared to 16 representative image fusion methods: Laplacian pyramid (LP) [2], ratio of low-pass pyramid (RP) [41], non-subsampled contourlet transform (NSCT) [50], discrete wavelet transform (DWT) [16], dual-tree complex wavelet transform (DTCWT) [14], sparse representation (SR) [49], curvelet transform (CVT) [29], guided filtering (GF) [18], multi-scale weighted gradient (MWG) [52], dense SIFT (DSIFT) [25], spatial frequency (SF) [17], FocusStack [44], image matting fusion (IMF) [19], DeepFuse [32], DenseFuse (both add and l1-norm fusion strategy) [15], and CNN-Fuse [26]. The implementations of GF and IMF were derived from Dongkang Xu's website [45], and NSCT, CVT, DWT, DTCWT, LP, RP, SR, and CNN-Fuse from Yu Liu's website [22], with parameters set to recommended values.

We adopted Q_g [48], Q_m [31], and Q_{cb} [3] as fusion quality metrics to objectively assess performance. These metrics are described as follows.

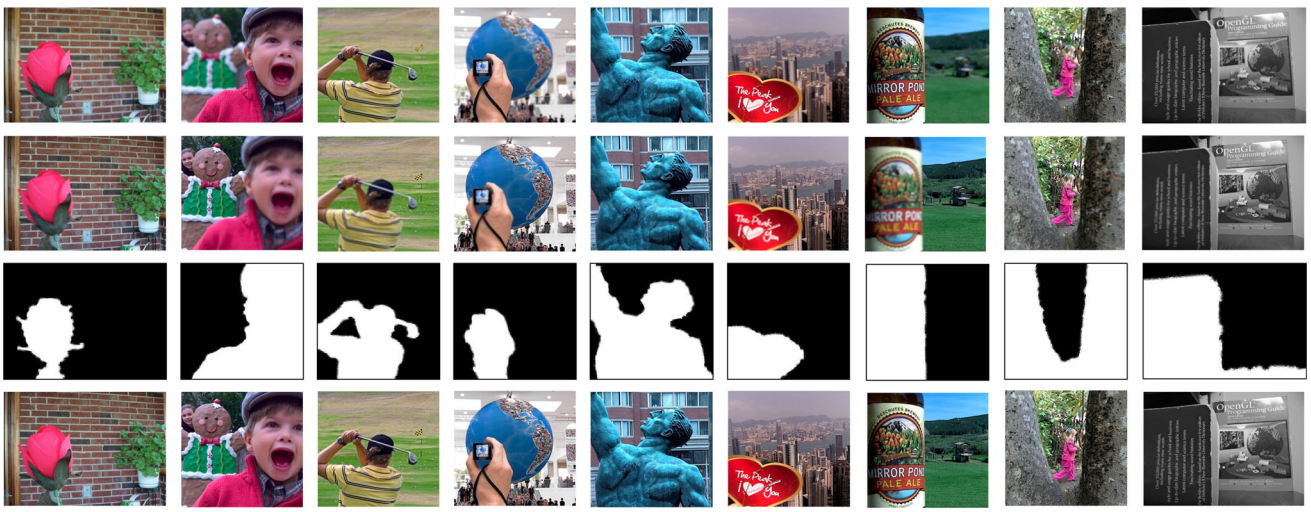

Fig. 2 Visualization of fused results. The first row is the near-focused source image, the second row is the far focused source image, the third row is the decision map, and the final row is the fused result

Q_g evaluates the amount of edge information transferred from input images to the fused image [48]. Consider two input images A and B, and a resulting fused image F. A Sobel edge operator is applied to yield the edge strength $g(x, y)$ and orientation $\alpha(x, y)$ of each pixel. Thus, for an input image A :

$$g_A(x,y) = \sqrt{s_A^x(x,y)^2 + s_A^y(x,y)^2} \tag{10}$$

$$\alpha_A(x,y) = \tan^{-1}\left(\frac{s_A^y(x,y)^2}{s_A^x(x,y)^2}\right), \tag{11}$$

where $s_A^x(x,y)$ and $s_A^y(x,y)$ are the respective convolved results with the horizontal and vertical Sobel templates. The relative strength $G^{AF}(x,y)$ and orientation values $\Delta^{AF}(x,y)$ between input image A and fused image F are defined as:

$$G^{AF}(x,y) = \begin{cases} \dfrac{g_F(x,y)}{g_A(x,y)}, & \text{if} g_A(x,y) > g_F(x,y) \\ \dfrac{g_A(x,y)}{g_F(x,y)}, & \text{otherwise} \end{cases} \tag{12}$$

$$\Delta^{AF}(x,y) = 1 - \frac{|\alpha_A(x,y) - \alpha_F(x,y)|}{\pi/2}. \tag{13}$$

The edge strength and orientation preservation values, respectively, can be derived as:

$$Q_g^{AF}(x,y) = \frac{\Gamma_g}{1 + e^{k_g(G^{AF}(x,y) - \sigma_g)}} \tag{14}$$

$$Q_\alpha^{AF}(x,y) = \frac{\Gamma_\alpha}{1 + e^{k_\alpha(\Delta^{AF}(x,y) - \sigma_\alpha)}}. \tag{15}$$

The constants Γ_g, k_g, σ_g and Γ_α, k_α, σ_α determine the shapes of the respective sigmoid functions used to form the edge strength and orientation preservation value. Normally, $\Gamma_g = \Gamma_a = 1$, $k_g = -10$, $k_\alpha = -20$, $\sigma_g = -0.5$, $\sigma_\alpha = -0.75$. The edge information preservation value is then defined as

$$Q^{AF}(x,y) = Q_g^{AF}(x,y) Q_\alpha^{AF}(x,y). \tag{16}$$

The final assessment is obtained from the weighted average of the edge information preservation values:

$$Q_g = \frac{\sum_{x=1}^{N}\sum_{y=1}^{M} Q^{AF}(x,y) w^A(x,y) + Q^{BF}(x,y) w^B(x,y)}{\sum_{x=1}^{N}\sum_{y=1}^{M} w^A(x,y) + w^B(x,y)}, \tag{17}$$

where $w^A(x,y) = [g_A(x,y)]^L$ and $w^B(x,y) = [g_B(x,y)]^L$. L is a constant, and usually $L = 1$.

Q_m is a metric based on a two-level Haar wavelet. The edge information is retrieved from the high and band-pass components of the decomposition [31]. Consider two input images A and B, and a resulting fused image F. For the i-th scale, the respective horizontal, vertical, and diagonal edge preservation values can be calculated as:

$$\begin{aligned} \varphi_{H_i}^{AF}(x,y) &= \exp(-|LH_i^A(x,y) - LH_i^F(x,y)|) \\ \varphi_{V_i}^{AF}(x,y) &= \exp(-|HL_i^A(x,y) - HL_i^F(x,y)|) \\ \varphi_{D_i}^{AF}(x,y) &= \exp(-|HH_i^A(x,y) - HH_i^F(x,y)|) \end{aligned} . \tag{18}$$

Then, the global edge preservation value at scale i can be derived as

$$EP_i^{AF}(x,y) = \frac{\varphi_{H_i}^{AF}(x,y) + \varphi_{V_i}^{AF}(x,y) + \varphi_{D_i}^{AF}(x,y)}{3}. \tag{19}$$

A normalized weighted performance metric of scale i can be obtained based on EP_i^{AF} and EP_i^{BF}:

$$Q_i^{AB/F} = \frac{\sum_x \sum_y (EP_i^{AF}(x,y) w_i^A(x,y) + EP_i^{BF}(x,y) w_i^B(x,y))}{\sum_x \sum_y (w_i^A(x,y) + w_i^B(x,y))}. \tag{20}$$

The weight coefficient is defined by the high-frequency energy of the input images:

$$w_i^A(x,y) = LH_{A_s}^2(x,y) + HL_{A_s}^2(x,y) + HH_{A_s}^2(x,y). \tag{21}$$

The overall metric is obtained by combining the measurements at different scales:

$$Q_M = \prod_{s=1}^{N} (Q_s^{AB/F})^{\alpha_s}, \tag{22}$$

where α_s is a constant to adjust the relative importance of different scales, and N is the decomposition level. Normally, $N = 2$, $\alpha_1 = 2/3$, $\alpha_2 = 1/3$.

Q_{cb} is a perceptual quality measure for image fusion, which employs the major features in a human visual system model [3]. Consider two input images I_A and I_B, and a resulting fused image I_F. All of the images are filtered by an empirical CSF using a DOG filter and Fourier transform. The local contrast is defined as

$$C_A(x,y) = \frac{\varphi_k(x,y) * I_A(x,y)}{\varphi_{k+1}(x,y) * I_A(x,y)} - 1. \tag{23}$$

A common choice for φ_j would be a Gaussian kernel with a standard deviation of $\sigma_k = 2^k$:

$$\varphi_k(x,y) = \frac{1}{(\sqrt{2\pi}\sigma_k)^2} e^{-\frac{x^2+y^2}{2\sigma_k^2}}. \tag{24}$$

Then, the masked contrast map for input image $I_A(x,y)$ is calculated as

$$C'_A = \frac{t((C_A)^p)}{h((C_A)^q) + Z}, \tag{25}$$

where t, h, p, q, and Z are real scalar parameters that determine the shape of the nonlinearity of the masking function. Normally, t = 1, h = 1, p = 3, q = 2, Z = 0.0001. After the masked contrast map is calculated, the salience map for $I_A(x, y)$ is defined as

$$\lambda_A(x, y) = \frac{C'_A(x, y)^2}{C'_A(x, y)^2 + C'_B(x, y)^2}. \tag{26}$$

The information preservation value is

$$Q_{AF}(x, y) = \begin{cases} \dfrac{C'_F(x, y)}{C'_A(x, y)}, & \text{if } C'_A(x, y) > C'_F(x, y) \\ \dfrac{C'_A(x, y)}{C'_F(x, y)}, & \text{otherwise} \end{cases}. \tag{27}$$

We can obtain the global quality map as

$$Q_{GQM}(x, y) = \lambda_A(x, y)Q_{AF}(x, y) + \lambda_B(x, y)Q_{BF}(x, y). \tag{28}$$

Finally, the metric value is obtained by averaging the global quality map:

$$Q_{CB} = \overline{Q_{\mathrm{GQM}}(x, y)}. \tag{29}$$

A larger value of any of the above three metrics indicates better fusion performance. A good comprehensive survey of quality metrics can be found in Liu et al. [24]. For fair comparison, we use appropriate default parameters for these metrics, and all codes are derived from their public codes [23].

4.3 Ablation experiments

We evaluated our method with different settings to verify the contribution of the gradient module. We picked eight fusion modes to explore the use of deep features. We use '(s)(c)SE_Y(_dm)' form to represent different method. 'sSE', 'cSE' and 'scSE' are three versions of the squeeze and excitation blocks in the former of this paper. 'Y' means fusion mode, such as max, absmax, l1-norm and sf, and sf denotes spatial frequency. 'dm' means the algorithm use decision map which export from encoder to fuse images. For the algorithm without 'dm' term, it denotes that the network directly output fusing result from decoder, which receives the input of fused feature. DenseFuse [15] compared add and l1-norm fusion strategies and drew the conclusion that the l1-norm of deep features could be used to fuse infrared-visible images. It utilized feature intensity to calculate the activity level. We find that the feature gradient (calculated by spatial frequency) is suited to the multi-focus fusion task. Table 1 shows the mean average scores for different methods. The bold value shows the best performance among all fusion modes. The number in parentheses denotes the number of first place results in 38 pairs of multi-focus images (mentioned in Sect. 4.1) ordered by compared methods. With the cSE module, sf outperforms abs-max, max, average, and l1-norm fusion modes in metric evaluation. In addition, while deep learning has promising representative ability, it cannot perfectly recover the image. Thus, if we use sf to fuse the deep features and input to the decoder, the fused result cannot recover every detail of the in-focus region. Therefore, we propose to use deep features to calculate the decision map and fuse the original images. As shown in the experimental results, cSE_sf_dm outperforms cSE_sf. We conducted an experiment to verify the influence of the SE architecture [10], and found that the first-place numbers of Q_g and Q_m of cSE_sf_dm are higher than those of sSE_sf_dm and scSE_sf_dm. We think that channel-wise recalibration is more important and robust than spatial recalibration for gradient calculation. Therefore, we represented SESF-Fuse by cSE_sf_dm in the next experiment.

Compared to the traditional spatial frequency method, SESF-Fuse analyzes the sharp appearance from deep features instead of original input images. We demonstrate the visualization of two methods in Fig. 3a, b for near- and far-focused images, respectively. Figure 3c shows spatial frequency, and analyzes the sharp appearance of the original image. Figure 3d shows SESF-Fuse, and it analyzes the sharp appearance in deep features. A detailed region is shown in the red rectangle, indicating that the combination of deep learning and spatial frequency can accurately preserve the focused region.

4.4 Comparison to classical fusion methods

We compare classical fusion methods in subjective and objective assessments. We first compare fusion methods according to visual assessment. For this purpose, three

Table 1 Ablation experiments with different settings

Method	Q_g	Q_m	Q_{cb}
cSE_absmax	0.5204(0)	2.4880(0)	0.6019(0)
cSE_average	0.5033(0)	2.4835(0)	0.5963(0)
cSE_l1_norm	0.5124(0)	2.4961(0)	0.6020(0)
cSE_max	0.5059(0)	2.4851(0)	0.5980(0)
cSE_sf	0.6885(0)	2.7216(2)	0.7526(0)
cSE_sf_dm	**0.7105(17)**	**2.8886(17)**	0.7848(14)
sSE_sf_dm	0.7103(6)	2.8889(7)	0.7849(7)
scSE_sf_dm	0.7104(15)	2.8903(12)	**0.7852(15)**

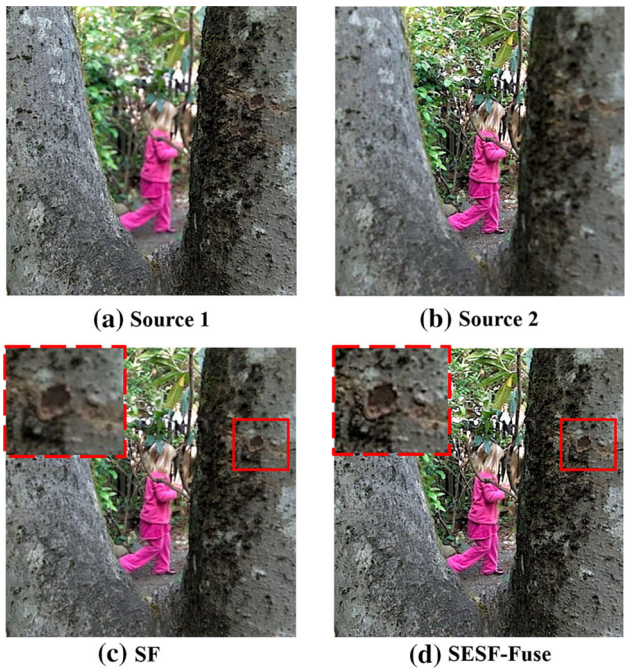

Fig. 3 Visualization of SF and SESF-Fuse fused results

fused results in two visualization modes are presented to show the differences among different methods.

In Fig. 4, we visualize "leaf" image pairs and their fused results. A region is magnified and shown in the upper-left corner in each image. We can see the border of the leaf with different methods. DWT shows a serrated shape, and CVT, DSIFT, SR, DenseFuse, and CNN show undesirable artifacts. For DWT and DenseFuse, the luminance of the leaf at the top-right corner shows an abnormal increase. The same region in MWG is out of focus, which means that the method cannot well detect the focused regions.

For a better comparison, Figs. 5 and 6 show the difference images obtained by subtracting the first source image from each fused image, which is normalized to the range of 0 to 1 for visualization. If the near-focused region is completely detected, the difference image will not show any of its information. Figure 5 shows a beer bottle. CVT, DSIFT, DWT, and DenseFuse-1e3-l1-Norm cannot perfectly detect the focused region. SR, MWG, and CNN perform well except in the region at the border of the bottle, because we still can see the contour of the near-focused region. SESF-Fuse performs well in both the center and border near-focused regions. In Fig. 6, the near-focused region is a man. As above, CVT, DSIFT, DWT, NSCT, and DenseFuse cannot perfectly detect the focused region. MWG and CNN perform well except in the region at the border. For MWG, the region surrounded by arms (red rectangle in Fig. 6) is actually a far-focused region, which MWG cannot correctly detect.

Table 2 shows the objective performance of different fusion methods. We find that SESF-Fuse and CNN-Fuse clearly outperform the other 15 methods on the average values of the Q_g and Q_{cb} fusing metrics. CNN-Fuse and SESF-Fuse achieve comparable performance on Q_g. However, CNN-Fuse is a supervised method which must generate synthetic images with different blur levels to train a two-class image classification network. In contrast, our network only must train an unsupervised model, which does not need to generate synthetic image data. For the Q_m metric, SESF-Fuse has a smaller average score than LP. However, the first place number of the proposed method achieves the highest value, which means it is more robust than other methods. Thanks to the further experiment and confirmation of other works [46, 47], which follow our preprint version and public code, SESF-Fuse performs best on other gradient-based metrics, such as spatial frequency

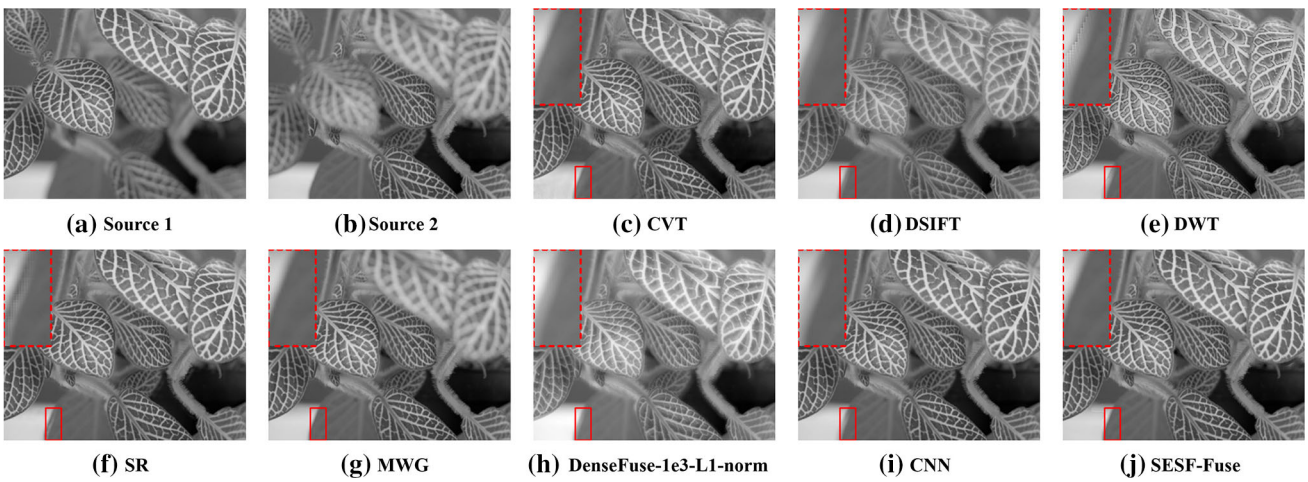

Fig. 4 Visualization of "leaf" fused results

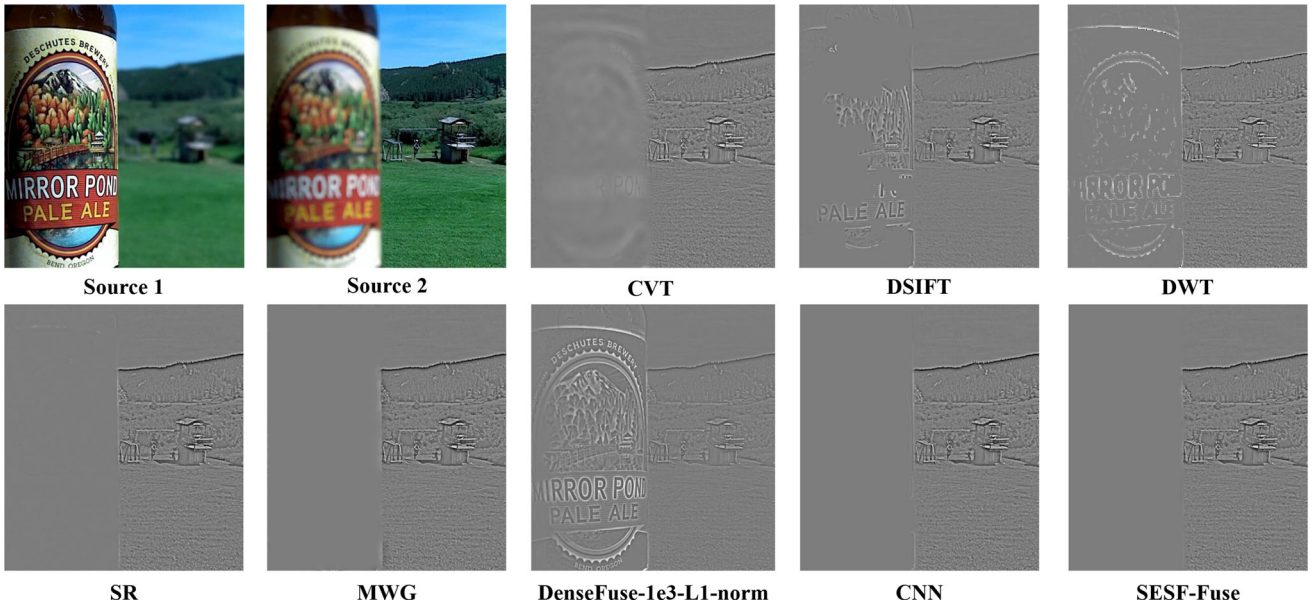

Fig. 5 Difference images for "beer" fused results

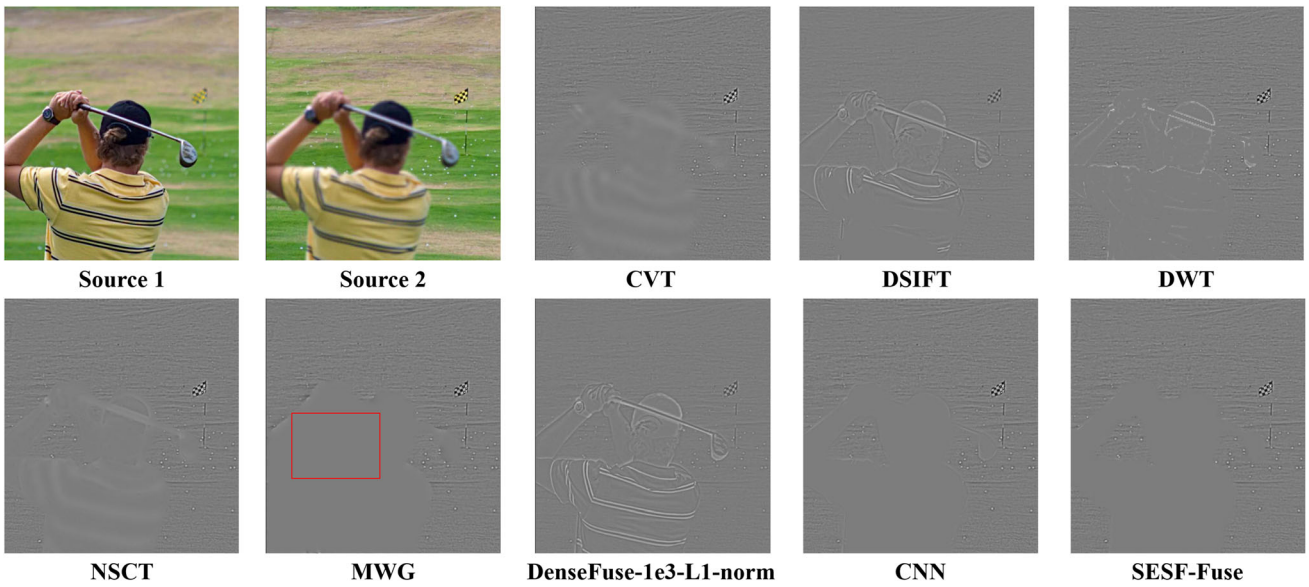

Fig. 6 Difference images of "golf" fused results

(SF), average gradient (AG), gray level difference (GLD), and visual information fidelity (VIF).

Fluctuations, such as Gaussian or salt-and-pepper noise, may randomly occur on each pixel during imaging, which influences the calculation of the activity level. We propose some consistency-verification methods for this. Morphology operations, such as opening and closing, and a small-region-removal strategy, will eliminate incorrect focused pixels in the decision map. Noises may have some positive roles in image fusion, such as stochastic resonance [11, 33, 36], which will induce an enhancement of the dynamical stability of the deep learning model in the training and inference phases, resulting in a less noisy response and robust feature encoder. Some demonstrations of the positive role of noises have been published [37, 38, 42].

Table 3 lists the run times of different fusion methods. Such methods as SESF, CNN-Fuse, DenseFuse, and DeepFuse were tested on a GTX 1080Ti GPU, and others on an E5-2620 CPU. Due to the parallel implementation of convolutional networks and its spatial frequency, SESF-Fuse achieved an average running time of 0.3432 second, which is faster than most of the methods. Although DeepFuse performed best on this test, SESF-Fuse showed a

Table 2 Comparison of multi-focus fusion methods

Metrics	DeepFuse	FocusStack	SF	DenseFuse_1e3_add	DSIFT	DenseFuse_1e3_l1
Q_g	0.4269(0)	0.4709(0)	0.5115(0)	0.5190(0)	0.5267(0)	0.5283(0)
Q_m	2.4618(0)	2.8510(0)	2.8512(0)	2.8530(0)	2.8725(0)	2.8561(0)
Q_{cb}	0.5651(0)	0.6330(0)	0.6024(0)	0.6008(0)	0.6067(0)	0.5972(0)

Metrics	GF	CVT	DWT	IMF	RP	DTCWT
Q_g	0.5631(0)	0.6187(0)	0.6222(0)	0.6324(2)	0.6478(0)	0.6529(0)
Q_m	2.8506(0)	2.9563(0)	2.9465(1)	2.8844(0)	2.9460(0)	2.9583(0)
Q_{cb}	0.7008(3)	0.6908(0)	0.6712(0)	0.7362(4)	0.7101(0)	0.7126(0)

Metrics	NSCT	SR	LP	MWG	CNN-Fuse	SESF-fuse
Q_g	0.6587(0)	0.6686(0)	0.6731(0)	0.6998(0)	0.7102(16)	**0.7105(20)**
Q_m	2.9592(0)	2.9630(2)	2.9642(8)	2.9615(6)	2.9654(7)	**2.8886(14)**
Q_{cb}	0.7169(0)	0.7335(0)	0.7352(0)	0.7764(2)	0.7839(9)	**0.7848(20)**

Table 3 Running times of different methods (unit: second)

Running time	SR	CNN-Fuse	DSIFT	IMF	GF	MWG
Mean	773.3400	190.9777	58.3616	38.9755	28.6036	27.2727
STD	101.5714	34.4013	59.7994	0.8017	0.0194	0.1568

Running time	NSCT	CVT	DTCWT	RP	DWT	LP
Mean	21.8151	15.9019	13.0514	12.4081	12.3742	12.2998
STD	0.1426	0.0342	0.0120	0.0310	0.0112	0.0058

Running time	DenseFuse_1e3_l1	SF	SESF-Fuse	DenseFuse_1e3_add	FocusStack	DeepFuse
Mean	10.8645	2.2039	0.3432	0.2925	0.2337	**0.0948**
STD	0.1887	0.0018	0.0002	0.0064	0.0044	**0.0001**

28% improvement on Q_g compared to DeepFuse. Thanks to the further experiment and confirmation of [46], SESF-Fuse was also faster than other fusion methods, such as DCTvar [6], GBM [30], and GCF [46]. With a fast running time and unsupervised network architecture, SESF-Fuse is easily applied to specific applications, such as IoT, because there is no need to manually or automatically generate a labeled multi-focus dataset to train the network.

Based on the above comparisons of subjective visualization and objective evaluation, our proposed SESF-Fuse fusing algorithm outperforms classical methods and achieves promising performance at the task of multi-focus image fusion.

5 Discussion

We discuss the potential of the combination of deep learning and traditional vision algorithms in multi-focus fusion and explore its advantages over previous methods. According to ours and other experiments [46, 47], our method explicitly calculates gradient information in the inference phase and uses it to generate a decision map, hence it is inclined to detect the clarity region with sharp appearance, which has more gradient information. One can design a more sophisticated gradient computation method to replace our spatial frequency for better performance. We think there is still ample room to improve fusion methods.

One limitation of our method, and methods such as DSIFT [25] and CNN-Fuse [26], is that they cannot be simply applied to multiple image fusion. Most of them inefficiently fuse images, one by one, in series, mainly because the output image from the decoder cannot exactly recover the real pixel value, which causes a low Q_g value in evaluation. So, these methods try to acquire an intermediate result (decision map) to recover the true pixel value of the source image. Some post-processing methods can only be applied to decision maps in two-image fusion, which prevents efficient multi-image fusion. Therefore, the ability of the decoder must improve to solve the problem. If a decoder can precisely recover the source pixel value without post-processing, then the network will perform

feature extraction by multiple encoders and fuse them by one decoder to save inference time. GAN and a well-designed loss function may be a good solution to this problem.

SESF-Fuse cannot address the defocus spread effect. A novel deep learning model, MMFNet, which employs a new data generation method and loss function [27], can produce a better fusion result in objective assessment with a defocus spread effect [47]. However, MMFNet performed less well than SESF-Fuse in gradient-based metrics in an experiment [47]. We speculate that MMFNet acts like a guided filter [7] to smooth the boundary of the focus and defocus regions to produce a better visualization fusing result. However, smooth operation will decrease the gradient-based subjective evaluation scores. We believe a better way will be found to combine the advantages of the two methods.

Multi-image fusion and the defocus spread effect may appear in many industrial applications, such as imaging of electronic components. Many improvements can be researched in future work.

6 Conclusion

We presented an unsupervised deep learning model to address the multi-focus image fusion problem. The key point of our model is that the DOF can be analyzed by the sharp appearance from deep features. First, we trained an encoder–decoder network in an unsupervised manner to acquire deep features of input images. We utilized spatial frequency to calculate the activity level from these features and perform image fusion according to a decision map. Experimental results demonstrate that the presented method achieves promising fusion performance compared to current fusion methods. This paper demonstrates the viability of the combination of unsupervised learning and traditional image processing algorithms.

Acknowledgements We acknowledge the support of the National Key Research and Development Program of China (No. 2016YFB0700500), National Science Foundation of China (No. 6170203, No. 61873299), Key Research Plan of Hainan Province (No. ZDYF2019009), Guangdong Province Key Area R and D Program (No. 2019B010940001), Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB (No. BK19BE030), and Fundamental Research Funds for the University of Science and Technology Beijing (No. FRF-BD-19-012A, No. FRF-TP-19-043A2). The computing work was supported by USTB MatCom of Beijing Advanced Innovation Center for Materials Genome Engineering.

Compliance with ethical standards

Conflicts of interest The authors declare that there is no conflict of interest.

References

1. Aslantas V, Kurban R (2010) Fusion of multi-focus images using differential evolution algorithm. Expert Syst Appl 37(12):8861–8870. https://doi.org/10.1016/j.eswa.2010.06.011
2. Burt P, Adelson E (1983) The laplacian pyramid as a compact image code. IEEE Trans Commun 31(4):532–540. https://doi.org/10.1109/TCOM.1983.1095851
3. Chen Y, Blum RS (2009) A new automated quality assessment algorithm for image fusion. Image Vis Comput 27(10):1421–1432. https://doi.org/10.1016/j.imavis.2007.12.002 **(Special Section: Computer Vision Methods for Ambient Intelligence)**
4. De I, Chanda B, Chattopadhyay B (2006) Enhancing effective depth-of-field by image fusion using mathematical morphology. Image Vision Comput 24(12):1278–1287. https://doi.org/10.1016/j.imavis.2006.04.005
5. Facebook: Pytorch. https://pytorch.org (2019)
6. Haghighat M, Aghagolzadeh A, Seyedarabi H (2011) Multi-focus image fusion for visual sensor networks in DCT domain. Comput Electr Eng 37(5):789–797
7. He K, Sun J, Tang X (2013) Guided image filtering. IEEE Trans Pattern Anal Mach Intell 35(6):1397–1409. https://doi.org/10.1109/TPAMI.2012.213
8. Huang J, Le Z, Ma Y, Mei X, Fan F (2020) A generative adversarial network with adaptive constraints for multi-focus image fusion. Neural Comput Appl. https://doi.org/10.1007/s00521-020-04863-1.pdf
9. Huang G, Liu Z, Van Der Maaten L, Weinberger K.Q (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
10. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: The IEEE conference on computer vision and pattern recognition (CVPR)
11. Itzcovich E, Riani M, Sannita WG (2017) Stochastic resonance improves vision in the severely impaired. Sci Rep 7(1):1–8
12. Jung H, Kim Y, Jang H, Ha N, Sohn K (2020) Unsupervised deep image fusion with structure tensor representations. IEEE Trans Image Process 29:3845–3858
13. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: International conference on learning representations
14. Lewis JJ, O'Callaghan RJ, Nikolov SG, Bull DR, Canagarajah N, (2007) Pixel- and region-based image fusion with complex wavelets. Inf Fusion 8(2):119–130. https://doi.org/10.1016/j.inffus.2005.09.006 Special Issue on Image Fusion: Advances in the State of the Art
15. Li H, Wu X (2019) Densefuse: A fusion approach to infrared and visible images. IEEE Trans Image Process 28(5):2614–2623. https://doi.org/10.1109/TIP.2018.2887342
16. Li H, Manjunath B, Mitra S (1995) Multisensor image fusion using the wavelet transform. Graph Models Image Process 57(3):235–245. https://doi.org/10.1006/gmip.1995.1022
17. Li S, Kwok JT, Wang Y (2001) Combination of images with diverse focuses using the spatial frequency. Inf Fusion 2(3):169–176. https://doi.org/10.1016/S1566-2535(01)00038-0
18. Li S, Kang X, Hu J (2013) Image fusion with guided filtering. IEEE Trans Image Process 22(7):2864–2875. https://doi.org/10.1109/TIP.2013.2244222
19. Li S, Kang X, Hu J, Yang B (2013) Image matting for fusion of multi-focus images in dynamic scenes. Inf Fusion 14(2):147–162. https://doi.org/10.1016/j.inffus.2011.07.001
20. Li S, Kang X, Fang L, Hu J, Yin H (2017) Pixel-level image fusion: a survey of the state of the art. Inf Fusion 33:100–112. https://doi.org/10.1016/j.inffus.2016.05.004

21. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer vision—ECCV 2014. Springer, Cham, pp 740–755
22. Liu Y (2019) Image fusion. http://www.escience.cn/people/liuyu1/Codes.html
23. Liu Z (2012) Image fusion metrics. https://github.com/zhengliu6699/imageFusionMetrics
24. Liu Z, Blasch E, Xue Z, Zhao J, Laganiere R, Wu W (2012) Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study. IEEE Trans Pattern Anal Mach Intell 34(1):94–109. https://doi.org/10.1109/TPAMI.2011.109
25. Liu Y, Liu S, Wang Z (2015) Multi-focus image fusion with dense sift. Inf Fusion 23:139–155. https://doi.org/10.1016/j.inffus.2014.05.004
26. Liu Y, Chen X, Peng H, Wang Z (2017) Multi-focus image fusion with a deep convolutional neural network. Inf Fusion 36:191–207. https://doi.org/10.1016/j.inffus.2016.12.001
27. Ma H, Liao Q, Zhang J, Liu S, Xue JH (2019) An α matte boundary defocus model based cascaded network for multi-focus image fusion
28. Nejati M, Samavi S, Shirani S (2015) Multi-focus image fusion using dictionary-based sparse representation. Inf Fusion 25:72–84. https://doi.org/10.1016/j.inffus.2014.10.004
29. Nencini F, Garzelli A, Baronti S, Alparone L (2007) Remote sensing image fusion using the curvelet transform. Inf Fusion 8(2):143–156. https://doi.org/10.1016/j.inffus.2006.02.001 **(Special Issue on Image Fusion: Advances in the State of the Art)**
30. Paul S, Sevcenco IS, Agathoklis P (2016) Multi-exposure and multi-focus image fusion in gradient domain. J Circuits Syst Comput 25:1650123
31. Peng-wei Wang, Bo Liu (2008) A novel image fusion metric based on multi-scale analysis. In: 2008 9th international conference on signal processing, pp 965–968 . https://doi.org/10.1109/ICOSP.2008.4697288
32. Prabhakar R (2017) Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: The IEEE international conference on computer vision (ICCV)
33. Riani M, Simonotto E (1994) Stochastic resonance in the perceptual interpretation of ambiguous figures: a neural network model. Phys Rev Lett 72(19):3120
34. Roy AG, Navab N, Wachinger C (2018) Concurrent spatial and channel squeeze and excitation in fully convolutional networks. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 421–429
35. Savić S, Babić Z (2012) Multifocus image fusion based on empirical mode decomposition. In: 19th IEEE international conference on systems, signals and image processing (IWSSIP)
36. Simonotto E, Riani M, Seife C, Roberts M, Twitty J, Moss F (1997) Visual perception of stochastic resonance. Phys Rev Lett 78(6):1186
37. Spagnolo B, Valenti D, Guarcello C, Carollo A, Adorno DP, Spezia S, Pizzolato N, Di Paola B (2015) Noise-induced effects in nonlinear relaxation of condensed matter systems. Chaos Solitons Fractals 81:412–424
38. Spagnolo B, Guarcello C, Magazzù L, Carollo A, Persano Adorno D, Valenti D (2017) Nonlinear relaxation phenomena in metastable condensed matter systems. Entropy 19(1):20
39. Stathaki T (2011) Image fusion: algorithms and applications. Elsevier, Amsterdam
40. Tang H, Xiao B, Li W, Wang G (2017) Pixel convolutional neural network for multi-focus image fusion. Inf Sci. https://doi.org/10.1016/j.ins.2017.12.043
41. Toet A (1989) Image fusion by a ratio of low-pass pyramid. Pattern Recogn Lett 9(4):245–253. https://doi.org/10.1016/0167-8655(89)90003-2
42. Valenti D, Magazzù L, Caldara P, Spagnolo B (2015) Stabilization of quantum metastable states by dissipation. Phys Rev B 91(23):235412
43. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
44. Wikipedia: Focus stacking. https://github.com/cmcguinness/focusstack (2019)
45. Xu K (2019) Image fusion. http://xudongkang.weebly.com/index.html
46. Xu H, Fan F, Zhang H, Le Z, Huang J (2020) A deep model for multi-focus image fusion based on gradients and connected regions. IEEE Access 8:26316–26327
47. Xu S, Wei X, Zhang C, Liu J, Zhang J (2020) Mffw: A new dataset for multi-focus image fusion. arXiv preprint arXiv:2002.04780
48. Xydeas CS, Petrovic V (2000) Objective image fusion performance measure. Electron Lett 36(4):308–309. https://doi.org/10.1049/el:20000267
49. Yang B, Li S (2010) Multifocus image fusion and restoration with sparse representation. IEEE Trans Instrum Meas 59(4):884–892. https://doi.org/10.1109/TIM.2009.2026612
50. Zhang Q, Long Guo B (2009) Multifocus image fusion using the nonsubsampled contourlet transform. Signal Process 89(7):1334–1346. https://doi.org/10.1016/j.sigpro.2009.01.012
51. Zhang Y, Liu Y, Sun P, Yan H, Zhao X, Zhang L (2020) IFCNN: a general image fusion framework based on convolutional neural network. Inf Fusion 54:99–118
52. Zhou Z, Li S, Wang B (2014) Multi-scale weighted gradient-based fusion for multi-focus images. Inf Fusion 20:60–72. https://doi.org/10.1016/j.inffus.2013.11.005