End-to-end learning for simultaneously generating decision map and multi-focus image fusion result

Boyuan Ma [a,b,c,d,1], Xiang Yin [e,1], Di Wu [f], Haokai Shen [g], Xiaojuan Ban [a,b,c,d,⇑], Yu Wang [h,⇑]

[a] *Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology, Beijing, China*
[b] *Institute of Artificial Intelligence, University of Science and Technology, Beijing, China*
[c] *Shunde Graduate School of University of Science and Technology Beijing, Foshan, China*
[d] *Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, China*
[e] *Collaborative Innovation Center of Steel Technology, University of Science and Technology, Beijing, China*
[f] *Department of ICT and Natural Sciences, Norwegian University of Science and Technology, Aalesund, Norway*
[g] *North Automatic Control Technology Institute, Taiyuan, China*
[h] *School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing, China*

article info

Article history:
Received 13 July 2021
Revised 24 October 2021
Accepted 29 October 2021
Available online 06 November 2021 Communicated by Zidong Wang

Keywords:
Multi-focus image fusion
Multiple images fusion
Convolution neural network
Loss function

abstract

The general aim of multi-focus image fusion is to gather focused regions of different images to generate a unique all-in-focus fused image. Deep learning based methods become the mainstream of image fusion by virtue of its powerful feature representation ability. However, most of the existing deep learning struc- tures failed to balance fusion quality and end-to-end implementation convenience. End-to-end decoder design often leads to unrealistic result because of its non-linear mapping mechanism. On the other hand, generating an intermediate decision map achieves better quality for the fused image, but relies on the rectification with empirical post-processing parameter choices. In this work, to handle the requirements of both output image quality and comprehensive simplicity of structure implementation, we propose a cascade network to simultaneously generate decision map and fused result with an end-to-end training procedure. It avoids the dependence on empirical post-processing methods in the inference stage. To improve the fusion quality, we introduce a gradient aware loss function to preserve gradient information in output fused image. In addition, we design a decision calibration strategy to decrease the time con- sumption in the application of multiple images fusion. Extensive experiments are conducted to compare with 19 different state-of-the-art multi-focus image fusion structures with 6 assessment metrics. The results prove that our designed structure can generally ameliorate the output fused image quality, while implementation efficiency increases over 30% for multiple images fusion.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

The multi-focus image fusion is an important topic in image processing. The limitation of optical lenses naturally presents that only objects within the depth-of-field (DOF) have a focused and clear 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 [1]. Many algorithms have been designed to create an all-in-focus image by fusing multiple source images that capture the same scene with different focus points. The fused image can be used for visualization and further processing, such as object recognition and segmentation.

Deep learning based solutions [2] are accepted to be the prevail- ing choice for image fusion by virtue of its powerful feature repre- sentation ability. Yu Liu introduced a convolution neural network (CNN) to image fusion and proposed a CNN-Fuse fusion method to recognize which part of the image is in-focus with a supervised deep learning structure [3]. CNN-Fuse reached a better perfor- mance compared to traditional fusion algorithms based on the handcrafted features. Boyuan Ma moved further in applying an unsupervised training strategy to fuse images, termed as SESF- Fuse [4]. It avoided heavy labeling work for images to train the network.

Although deep learning has reached relatively good perfor- mance in multi-focus image fusion, the new problems yielded with complex structure design remain unsolved. There are three ques-

⇑ Corresponding authors at: Institute of Artificial Intelligence, University of Science and Technology, Beijing, China.
E-mail addresses: banxj@ustb.edu.cn (X. Ban), wangyubit@163.com (Y. Wang).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.neucom.2021.10.115
0925-2312/© 2021 Elsevier B.V. All rights reserved.

tions that deserve higher priorities. 1) The balance between fusion quality and end-to-end implementation convenience [5]. Some structures tried to use a decoder to directly output the final fused result [6,7]. However, they did not preserve true pixel values in the source image and hardly achieve good performance in fusing eval- uation due to the nonlinear mapping mechanism in the decoder. Some other structures generated intermediate decision map (DM) to reconstruct fused result with high quality [3,8]. But they relied highly on post-processing method (or consistency verifica- tion) choices. These methods require empirical parameters to rec- tify the DM, resulting in limits the generalization of them to different scenes of image fusion. 2) The gradient feature contains rich beneficial information for multi-focus image fusion. However, it was overlooked in many designs. Some deep learning structures used the l2 and SSIM objective functions to optimize the network [9,10]. These made the gradient feature completely lost during training procedure. 3) The efficiency in multiple images fusion. Currently, most of the multi-focus fusion structures focus on two images based fusion application. With multiple images fusion, the strategy is to go one by one in serial sequence [4]. However, the time consumption is scarcely acceptable for big volume image fusion.

In order to counterpoise the requirements of fused image qual- ity and training simplicity, we design a gradient aware cascade structure, termed GACN.[2] It simultaneously generates decision map and fused result with an end-to-end training procedure. The original pixel values in the source are retained to optimize output fused image bypassing empirical post-processing methods. Further- more, we modify a commonly used gradient based evaluation metric as the training loss function in order to preserve gradient informa- tion. For multiple images fusion, we simplify redundant calculations by proposing a calibration module to acquire the activity levels of all images. It helps to significantly decrease the time consumption. We highlight our contributions as follows:

- We propose a network to simultaneously generate decision map and fused result with an end-to-end training procedure.
- We introduce a gradient aware loss function to preserve gradi- ent information and improve output fusion quality.
- We design a decision calibration strategy for multiple images fusion in order to increase implementation efficiency.
- To prove the feasibility and efficiency of the proposed GACN, we conduct extensive experiments to compare with 19 different state-of-the-art (SOTA) multi-focus image fusion structures with 6 assessment metrics. We implement ablation studies additionally to test the impact of different loss function in our structure. The results prove that our designed structure can generally ameliorate the output fused image quality, and increase implementation efficiency over 30% for multiple images fusion.

2. Related work

The existing solutions for multi-focus image fusion can be gen- eralized into two orientations: handcrafted feature based and deep learning based algorithms.

2.1. Handcrafted feature based fusion algorithms

Handcrafted feature based fusion algorithms concentrate on the profound image analysis of transform or spatial domains. Trans- form domain based algorithms adopt decomposed coefficients from a selected transform domain to measure different activity levels in the input source images, such as laplacian pyramid (LP) [11] and non-subsampled contourlet transform (NSCT) [12]. Spatial domain based algorithms measure activity levels with gradient features, such as spatial frequency [13], multi-scale weighted gra- dient (MWG) [14], and dense SIFT (DSIFT) [15].

2.2. Deep learning based fusion algorithms

Deep learning based algorithms provide prevalent solutions to image fusion problems. CNN-Fuse [3] first used a convolutional network to automatically learn features in each patch of image and decided which patch is the clarity region, which achieved bet- ter performance compared to handcrafted feature based algo- rithms. Afterward, some researchers tried to modify the network to improve the fusion quality or efficiency. Han Tang proposed a pixel-wise fusion CNN to further improve the fusion quality [16]. Dense-Fuse [9], U2Fusion [17], and SESF-Fuse [4] fused images in the unsupervised training procedure. IFCNN [18] presented a gen- eral image fusion framework to handle different kinds of image fusion tasks. However, there are still other parts of deep learning based algorithm that need to refine.

The output mode is an important module in network designing [5]. Some algorithms tried to use a decoder to directly output the fused result. Hao Zhang [7] used only one convolutional layer in decoder to fuse multi-scale features and generate fused result. To improve the reconstructive ability, Hyungjoo Jung [19] used resid- ual block to improve the efficiency of gradient propagation, and some works [6,20] used generative adversarial network to auto- matically ameliorate fusion quality. However, due to nonlinear mapping in the decoder, these structures cannot precisely recon- struct fused result. This leads to relatively unrealistic performance in fusion evaluation. Therefore, some structures resorted to gener- ate an intermediate DM, to decide which pixel should appear in fused result. Some works [3,21] used CNN to directly output DM. SESF-Fuse [4] used spatial frequency to calculate gradient in deep features and draw out DM. Han Xu [8] used a binary gradient rela- tion map to further ask decoder to preserve gradient information in DM. Despite the highly fusing quality of these structures, they need some post-processing methods (or consistency verification) with empirical parameters to rectify the DM, such as morphology oper- ations (opening and closing calculation) and small region removal strategy, which limits the generalization of the structure to differ- ent scenes of image fusion.

The objective function is a key point in structure optimization. In the field of multi-focus image fusion, the gradient in source images is an important factor to decide which part of the image is clear. However, many deep learning structures only used the l2 norm and SSIM objective function to optimize the network [9,10], which did not ask the network to preserve the gradient information in fused image. Hyungjoo Jung [19] proposed structure tensor to preserve the overall contrast of images. Jinxing Li [21] used an edge-preserving loss function to preserve gradient infor- mation, but it only considered gradient intensity and not took ori- entation information into account. In this work, we try to modify the commonly used classical gradient based evaluation metric as the loss function to directly optimize the network to export clearly fused result.

Most applications of multi-focus fusion are based on multiple images. However, almost multi-focus fusion structures concen- trated on two images scene and only used one by one serial fusion strategy for multiple images [3,17], which has in-acceptable time consumption. To the best of our knowledge, we are the first work to concentrate on the implementation efficiency in multiple images fusion scene.

[2] The code and data are available at https://github.com/Keep-Passion/GACN.

3. Method

In this section, we illustrate the details of the main contributions of this work, such as the network structure, the loss function, and the decision calibration strategy.

3.1. Network structure

The overall fusion network structure is shown in Fig. 1. It includes two paths of convolutional operations, feature extraction and decision. First, we use the feature extraction path to collect multi-scale deep features of each source image. Second, we take the spatial frequency (SF) module to calculate activity level of each scale. Third, in the decision path, we concate multi-scale activity levels and feed them into some convolutional operations to draw out the initial DM, which records the probability of each pixel should be in-focused in each source image. Then we apply guided filter [22] to smooth the boundary of DM and acquire final DM. Finally, we cascade the fusion module in our structure and generate the fusion result.

3.1.1. Feature extraction path

As shown in Fig. 1, the feature extraction path is a siamese architecture [23], which uses the same architectures with the same weights. It consists of a cascade of four convolutional layers to extract multi-scale deep features from each source image, and uses densely connection architecture to connect the output of each layer to the other layers, which strengthens feature propagation and reduces the number of parameters [24,25]. To precisely localize the details of the image, there are no pooling layers in our network.

In addition, we use the squeeze and excitation (SE) module after each convolutional layer, which showed good performance at image recognition and segmentation [26]. It can effectively enhance spatial feature encoding by adaptive recalibrating channel-wise or spatial-feature responses. Same with [4], we use channel SE module (CSE) [27] in feature extraction path. 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.

After feature extraction, we calculate multi-scale activity levels using the SF module [4]. Consider two input images A and B, and a fused image F. Let DF be the deep features drawn from the convolutional layer of each scale. $DF^A_{(m,n)}$ is one feature vector of pixel i in source image A with (m, n) coordinates. We calculate its SF by:

$$RF^A_{(m,n)} = \sqrt{\sum_{-r\leqslant a,b\leqslant r}\left[DF^A_{(m+a,n+b)} - DF^A_{(m+a,n+b-1)}\right]^2} \tag{1}$$

$$CF^A_{(m,n)} = \sqrt{\sum_{-r\leqslant a,b\leqslant r}\left[DF^A_{(m+a,n+b)} - DF^A_{(m+a-1,n+b)}\right]^2} \tag{2}$$

$$SF^A_{(m,n)} = \sqrt{\frac{\left(CF^A_{(m,n)}\right)^2 + \left(RF^A_{(m,n)}\right)^2}{(2r+1)^2}} \tag{3}$$

where RF and CF are respectively the row and column vector frequencies. r is the kernel radius and $r = 5$ in our work. 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.

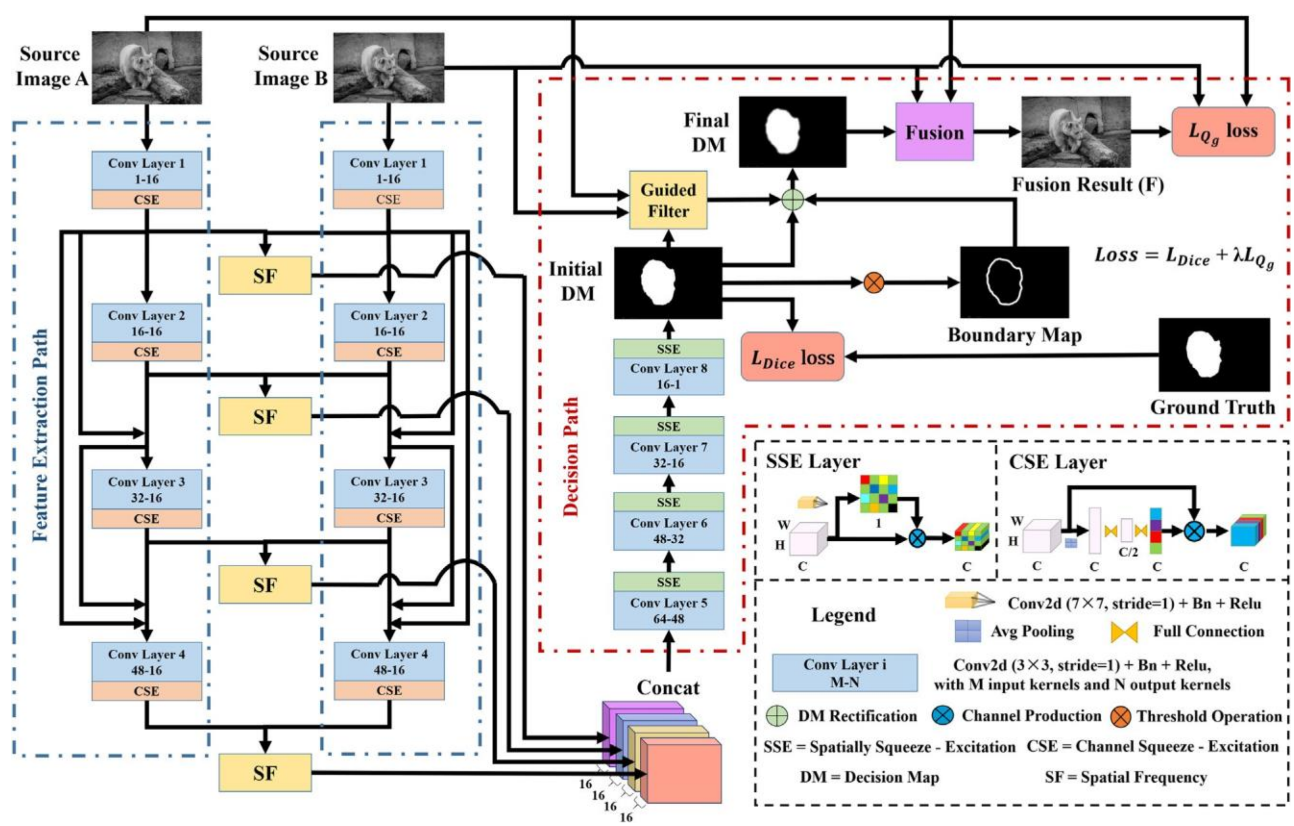

Fig. 1. The network structure of the proposed algorithm.

We subtract SF^B from SF^A to obtain activity level maps for each scale. Then we concate multi-scale activity level maps and feed them into the decision path.

3.1.2. Decision path

In the decision path, we first use four convolutional layers to generate the initial DM, which records the probability (p_i) that each pixel (i) of the source image A is more clear than that of the source image B. Because the output of lastest SSE module is non- binarized, so we add a sigmoid projection as Eq. 4 to project the non-binarized pixel value into range (0,1) after the lastest SSE module to generate the nearly binarized initial decision map.

$$y = \frac{1}{1 + e^{-kx}} \tag{4}$$

where k controls the steepness of the curve and closeness to the original Heaviside function, larger k means closer approximation (k ¼ 1000 in our work). The initial DM is optimized by loss function with ground truth DM, as shown in the next section.

In addition, we also use the SE module in the decision path. Specifically, we use spatial squeeze and channel excitation (SSE) [27], to enhance the robustness and representatives of deep fea- tures. SSE uses a convolutional layer with one k_s x k_s kernel to acquire a projection tensor (k_s ¼ 7 in our work). 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 orig- inal feature map.

To smooth the boundary of the fused result, we first use gaus- sian filter to filter the initial DM, then utilize threshold operation to obtain the boundary region. We found that thinner boundary regions make the boundary area of fusion result not smooth enough while thicker boundary regions make the boundary area of fusion result loss detail information. In this work, after multiple tests, we choose the pixels which value between ð0:1; 0:8] as the boundary region in order to make the boundary width within an acceptable range subjectively, and all images adopt a fixed thresh- old range. And then we use guided filter [22] to obtain the smooth DM. Finally, we use the boundary region as threshold region to combine the smooth DM and the initial DM to form the final DM. That is the boundary of the final DM is the smooth DM and the cen- ter of the final DM is the initial DM. Note that we only use a thresh- old operation to generate boundary region and do not hinder the backpropagation of network, which means that our structure can be trained by an end-to-end training procedure. In addition, we do not use non-differentiable post-processing methods with empirical parameters, such as morphology operation and small region removal strategy. Then, we cascade a fusion module using the final DM and source images to generate the fused result. As shown in Eq. 5, each pixel of fused image (F_i) can be obtained by:

$$F_i = p_i \times Img_i^A + (1 - p_i) \times Img_i^B \tag{5}$$

where the probability (p_i) in DM also means the fusion ratio of each pixel in the source images.

Finally, we use gradient aware loss function to optimize the net- work to preserve gradient information in fusion result.

In general, the network can simultaneously generate DM and fusion result with end-to-end training procedure.

3.2. Loss function

We define a gradient aware loss function to optimize the net- work to simultaneously output DM and clear fusion result. The final loss function is defined in Eq. 6.

$$L = L_{Dice} + \lambda L_{Q_g} \tag{6}$$

where k is a weight to balance the importance between two losses, and k ¼ 1 in this work.

L_{Dice} is a classical loss function in semantic segmentation [28], which is defined in Eq. 7.

$$L_{Dice} = 1 - \frac{2\sum_i^{N_P} p_i g_i + 1}{\sum_i^{N_P} p_i^2 + \sum_i^{N_P} g_i^2 + 1} \tag{7}$$

where the sums run over the N_P pixels, of the predicted binary seg- mentation map p_i 2 DM and the ground truth map g_i 2 G. Adding 1 is to mitigate the gradient vanishing issue.

In addition, we propose to use L_{Q_g} to optimize the network to export the final clear fused result. In the field of multi-focus image fusion, it is commonly speculated that only objects within the DOF have a sharp appearance in a photograph, while others are likely to be blurred [3]. However, lots of previous works did not consider preserving gradient information in network training. In this work, we focus on a classical gradient based fusion evaluation metric, Q_g or Q_F^{AB} [29], and make it differentiable as loss function in an end-to- end training procedure. By using this optimization, we lead the network to preserve gradient information in the final fused result. Q_g is an evaluation metric that measures the amount of edge information transferred from input images to the fused image [29]. Consider two input images A and B, and a fused image F. A sobel edge operator is applied to yield the edge strength g_i and ori- entation α_i of each pixel i. Thus, for an input image A:

$$g_i^A = \sqrt{(s_i^{Ax})^2 + (s_i^{Ay})^2} \tag{8}$$

$$\alpha_i^A = tan^{-1}\left(\frac{(s_i^{Ay})^2}{(s_i^{Ax})^2}\right) \tag{9}$$

where s^{Ax} and s^{Ay} are the respective convoluted results with the hor- izontal and vertical sobel templates.

The relative strength G_i^{AF} and orientation value Δ_i^{AF} between input image A and fused image F are defined as:

$$G_i^{AF} = \begin{cases} \frac{g_i^F}{g_i^A}, & \text{if } g_i^A > g_i^F \\ \frac{g_i^A}{g_i^F}, & \text{if } g_i^A \leqslant g_i^F \end{cases} \tag{10}$$

$$\Delta_i^{AF} = 1 - \frac{|\alpha_i^A - \alpha_i^F|}{\pi/2} \tag{11}$$

Unfortunately, the Heaviside function in Eq. 10 and absolute func- tion in Eq. 11 are not differentiable and thus cannot be included in training stage. Therefore, we propose to use the sigmoid function as a smooth approximation to the Heaviside function which is defined as:

$$f(x,y) = \frac{1}{1 + e^{-k(x-y)}} \tag{12}$$

where k controls the steepness of the curve and closeness to the original Heaviside function, larger k means closer approximation (k ¼ 1000 in our work). Then, Eq. 10 can be rewritten as Eq. 13.

$$G_i^{AF} \approx f(g_i^F, g_i^A) \times \frac{g_i^A}{g_i^F} + (1 - f(g_i^F, g_i^A)) \times \frac{g_i^F}{g_i^A} \tag{13}$$

And Eq. 11 can be rewritten as Eq. 14.

$$\Delta_i^{AF} \approx 1 - \frac{(\alpha_i^A - \alpha_i^F) \times (2f(\alpha_i^A, \alpha_i^F) - 1)}{\pi/2} \tag{14}$$

Note that in pytorch implementation [30], the gradient of abso- lute function is 0 when input of that equals 0, which is differen- tiable. Thus it can use Eq. 11 rather than Eq. 14 in pytorch. The detailed analysis can be found in the experiment section.

The edge strength and orientation preservation values, respec- tively, can be derived as:

$$Q^{AF}_{g_i} = \frac{\Gamma_g}{1 + e^{k_g(G^{AF}_i - \sigma_g)}} \tag{15}$$

$$Q^{AF}_{\alpha_i} = \frac{\Gamma_\alpha}{1 + e^{k_\alpha(A^{AF}_i - \sigma_\alpha)}} \tag{16}$$

where the constants C_g; k_g; r_g and C_a; k_a; r_a determine the shapes of the respective sigmoid functions (same with Eq. 12) used to form the edge strength and orientation preservation value. Normally, C_g ¼ C_a ¼ 1; k_g ¼ -10; k_a ¼ -20; r_g ¼ 0:5; r_a ¼ 0:75. The edge information preservation value is then defined as:

$$Q^{AF}_i \text{ ¼ } Q^{AF}_{g_i} \times Q^{AF}_{a_i} \tag{17}$$

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

$$L_{Q_g} = 1 - Q_g = 1 - \frac{\sum_i^{N_P}(Q^{AF}_i w^A_i + Q^{BF}_i w^B_i)}{\sum_i^{N_P}(w^A_i + w^B_i)} \tag{18}$$

where $w^A_i = [g^A_i]^\gamma$ and $w^B_i = [g^B_i]^\gamma$. γ is a constant, and usually sets γ = 1.

In total, we modify a gradient based classical fusion evaluation metric (Q_g) as a loss function to optimize the network to export clearly fused result.

We further show an experiment to visualize the comparison of L_{Dice} þ L_{Q_g} and L_{Q_g}, as shown in Fig. 2. It is shown that the fusion model trained with L_{Q_g} have less noise in the decision map com- pared to the model without it, which means L_{Q_g} can act as a post-processing method to improve the fusion quality because it can preserve gradient information in the image.

3.3. Decision calibration for multiple images fusion

Most applications of multi-focus fusion are based on multiple images. However, currently almost multi-focus fusion structures concentrated on two images scene and only used one by one serial fusion strategy for multiple images fusion. As shown at the top of Fig. 3, one-by-one serial strategy needs to run 2 x ðN_I - 1Þ times feature extraction paths and N_I - 1 times decision paths, where N_I is the number of the source images. In this work, we propose a decision calibration strategy, which shown at the bottom of Fig. 3. It only needs to run N_I times feature extraction paths and N_I - 1 times decision paths by using the calibration module, which can generally decrease time consumption.

In the decision calibration strategy, the first image is used as baseline, and feeds it to the structure with other images. Thus, we can save the parameters of the first image in the feature extraction path and avoid repeating computation. Then it uses final DMs drawn from each decision path to calculate the decision volume (DV), which records the activity levels of all the source images. The calculation process is acting as normalization to draw out rel- ative clarity of each source images, which is shown below:

$$DV^j_i = \begin{cases} p_i^2, & \text{if } j = 1 \\ 1 - p_i^2, & \text{if } j = 2 \\ \frac{p_i^2 \times (1 - p_i^j)}{p_i^j}, & others \end{cases} \tag{19}$$

where j belong to {1,..., N_I}, is the index of the source image and p^j_i is the value of pixel i in final DM when fuses the source image 1 and the source image j.

Then, we choose the index of maximum in DV^j_i for each pixel i as the index of the most clarity pixel i in the source images. According to the above indices, we can obtain the entire resulting fusion image.

It is important to notice that the decision calibration strategy can only applied to the DM based network structure without the empirical post-processing methods. Because those empirical post-processing methods, such as morphology operation and small region removal strategy, which used in CNN-Fuse [3] and SESF- Fuse [4], firstly require to convert the initial DM to the binary DM, which loss the relative clarity information and can not be used in the process of decision volume calculation. Our method, GACN, can draw out the decision map without the empirical post-processing methods, which is more suited to the decision calibra- tion strategy in the application of multiple images fusion.

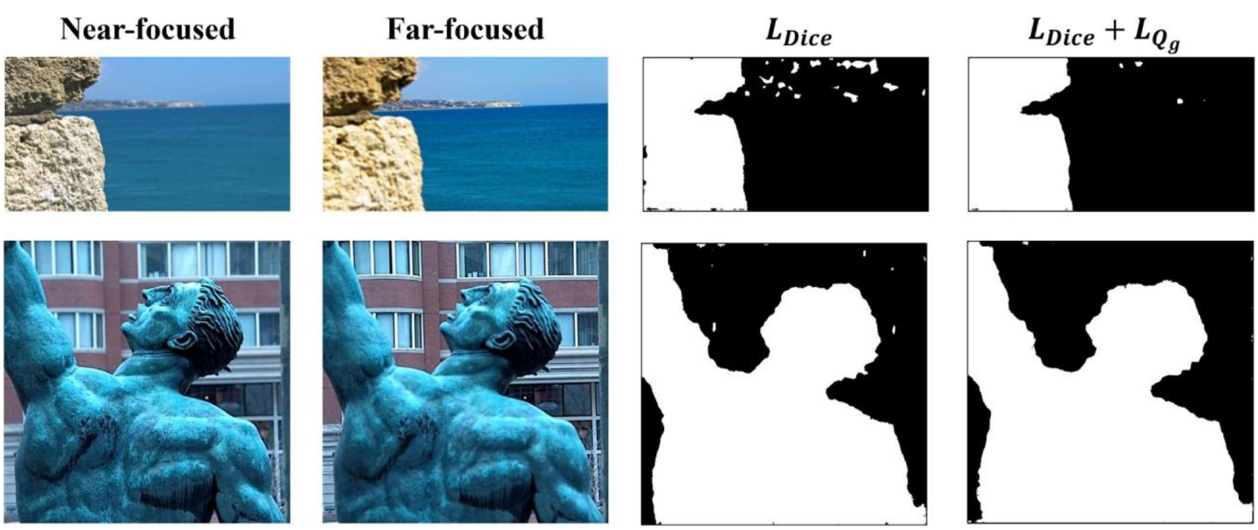

Fig. 2. Visualization of decision maps of the model trained with or without Q_g.

4. Experiment

4.1. Dataset

4.1.1. Training set

In this paper, we generate multi-focus image data based on MS COCO dataset [31]. The MS COCO dataset contains annotations for instance segmentation, and our method uses the original image and its segmentation annotation to generate multi-focus image data. That is, we use annotation as threshold region to decide which part of the image should be filtered by gaussian blurring. As shown in Fig. 4, the original image 'truck' and its annotation are obtained by MS COCO. We use gaussian filter to blur the back- ground to form near-focused image and blur the foreground form far-focused image. And we use the defocused spread effect model proposed in [32] to further improve the realness of the generated data. Thus we have two inputs of multi-focus images, one ground truth fused result (original image) and one decision map (label) for network training. Because some data in MS COCO dataset con- tains multiple instances that are not at the same depth-of-field (DOF), so we only select images that contain one instance. Besides, we regard the multi-focus image fusion problem as an image seg- mentation problem. The imbalance of the foreground and back- ground category often affects the segmentation results, so we further select the image with the foreground size between 20,000 and 170,000 pixels as the training data. Finally, we obtain 5786 images and divide these into training set and validation set according to the ratio of 7:3.

4.1.2. Testing set

We use 26 image pairs of publicly available multi-focus images from [33,34] as the testing set for evaluation.

4.2. Training procedure

During training, all images were transformed to gray-scale and resized to 256 x 256, then random cropped to 156 x 156. Note that images were gray-scale in the training phase, while images for testing can be gray-scale or color images with RGB channels. For color images that needed to be fused, we transformed the images to gray-scale and calculated a decision map to fuse them. In addi- tion, we used random crop, random blur, random offset, and gaus-

sian noise as data augmentation methods [35,36]. The initial

learning rate was 1 x 10^{-4}, and this was decreased by a factor of

0.8 at every two epochs [37,38]. We optimized the objective func- tion by Adam [39]. The batch size and number of epochs were 16 and 50, respectively [40]. Our implementation was derived from the publicly available Pytorch framework [30]. The network's training and testing were performed on a station using an NVIDIA Tesla V100 GPU with 32 GB memory.

4.3. Evaluation metrics

We use six classical fusion metrics: Q_g [29], Q_y [41], Q_{ncie} [42], Q_{cb} [43], *FMI EDGE* and *FMI DCT* [44] to evaluate the quality of fused result. Q_g evaluates the amount of edge information trans- ferred from input images to the fused image. Q_y calculates the sim-

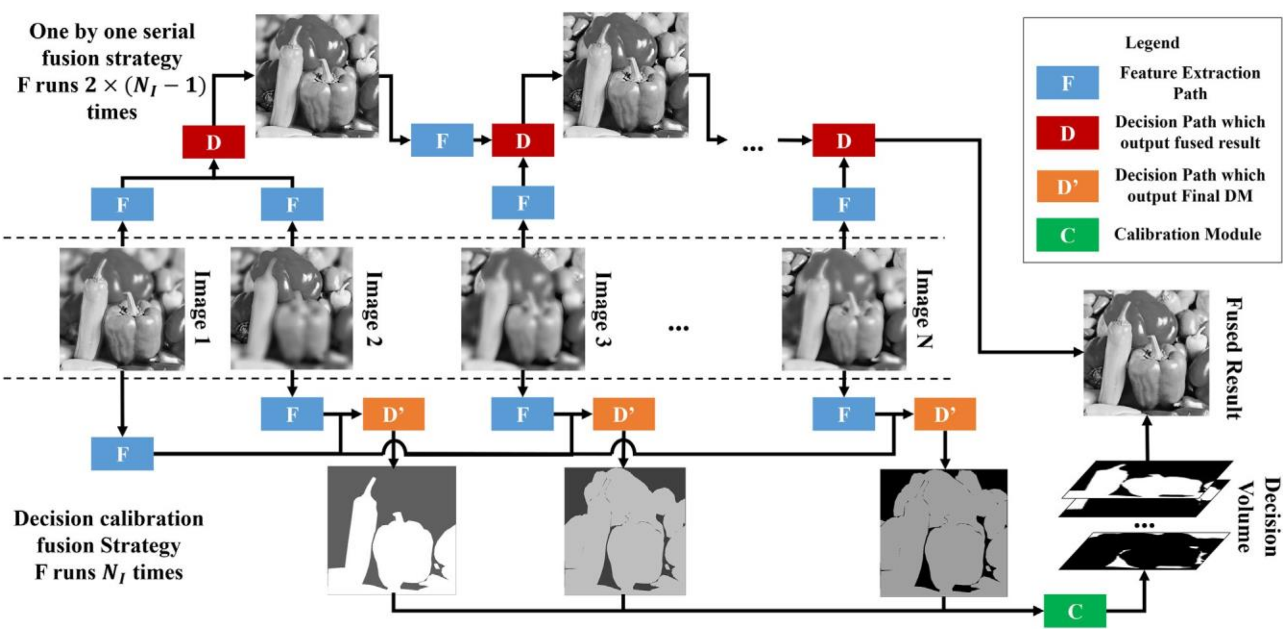

Fig. 3. The flowchart of the traditional one-by-one serial fusion strategy (Top) and the proposed decision calibration strategy (Bottom).

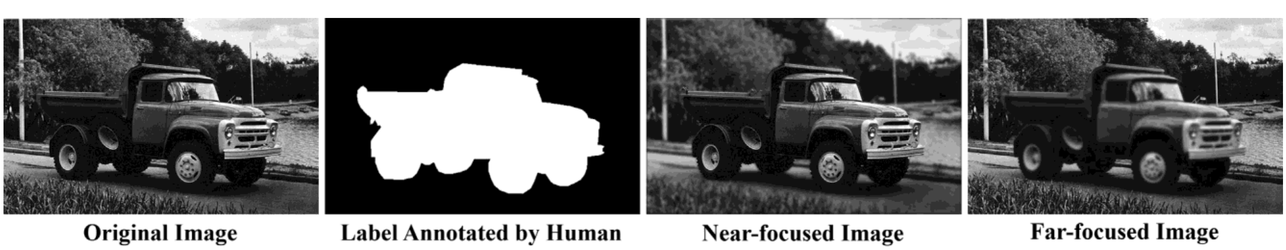

Fig. 4. Visualization of a training example generated by MS COCO dataset.

Table 1
Comparison with traditional methods in testing set. The bold value denotes best performance in each metric.

Methods	Q_g	Q_y	Q_{ncie}	Q_{cb}	*FMI_EDGE*	*FMI_DCT*	Time (s)
GACN	0.7169	0.97769	0.8411	0.7948	0.897806	0.4058	0.16
MFF-GAN (2021)	0.5623	0.88652	0.8210	0.6437	0.884512	0.3699	0.33
MFF-SSIM (2020)	0.7020	0.96712	0.8331	0.7678	0.895163	0.4009	36.19
DRPL (2020)	0.6919	0.96535	0.8333	0.7771	0.895190	0.3640	0.16
FusionDN (2020)	0.5216	0.82352	0.8209	0.6106	0.878785	0.3050	0.49
U2Fusion (2020)	0.5590	0.86993	0.8210	0.6388	0.882694	0.3118	0.75
IFCNN (2020)	0.6486	0.93751	0.8265	0.7158	0.891569	0.3757	0.06
PMGI (2020)	0.4803	0.80668	0.8209	0.5805	0.880374	0.3527	0.21
SESF-Fuse (2019)	0.7150	0.97761	0.8397	0.7965	0.897133	0.3953	0.30
Dense-Fuse (2019)	0.5329	0.83965	0.8239	0.6109	0.886998	0.4046	0.38
CNN-Fuse (2017)	0.7153	0.97706	0.8396	0.7676	0.897800	0.4079	188.16
DSIFT (2015)	0.5419	0.84643	0.8255	0.6306	0.889215	0.3900	49.28
MWG (2014)	0.7041	0.97720	0.8376	0.7878	0.898504	0.3965	24.99
Focus-Stack (2013)	0.5098	0.78907	0.8276	0.6628	0.868776	0.2332	0.19
SR (2010)	0.6792	0.95132	0.8326	0.7523	0.896763	0.3924	698.44
NSCT (2009)	0.6721	0.94886	0.8272	0.7326	0.896647	0.4037	19.99
CVT (2007)	0.6373	0.93765	0.8252	0.7111	0.895890	0.4055	14.76
DTCWT (2007)	0.6688	0.95190	0.8267	0.7304	0.896893	0.4031	12.21
SF (2001)	0.5202	0.82904	0.8239	0.6173	0.889395	0.4145	2.25
DWT (1995)	0.6444	0.91346	0.8326	0.6997	0.890219	0.3293	11.51
RP (1989)	0.6652	0.94001	0.8280	0.7330	0.892010	0.3574	11.34
LP (1983)	0.6834	0.95369	0.8286	0.7509	0.897242	0.3911	11.58

ilarity between fused image and the sources[41]. Q_{ncie} measures the nonlinear correlation information entropy between the input images and the fused image [42]. Q_{cb} is a perceptual quality mea- sure for image fusion, which employs the major features in a human visual system model [43]. *FMI EDGE* and *FMI DCT* calcu- lates the mutual information of the edge features and discrete cosine transform feature between the input images and the fused image [44]. A larger value of any of the above six metrics indicates better fusion performance. For fair comparison, we use appropriate default parameters for these metrics, and all codes are derived from their public resources [45,46].

4.4. Comparison

To demonstrate the performance of our method, we compare it with recent SOTA fusion methods in objective and subjective assessments.

4.4.1. Objective assessment

The comparison of our method with existing multi-focus fusion methods are listed in Table 1, such as MFF-GAN [20], FusionDN [25], U2Fusion [17], IFCNN [18], PMGI [7], DRPL [21], MFF-SSIM [47], SESF-Fuse [4], Dense-Fuse [9], CNN-Fuse [3], dense SIFT (DSIFT) [15], multi-scale weighted gradient (MWG) [14], Focus- Stack [48], sparse representation (SR) [49], non-subsampled con- tourlet transform (NSCT) [12], curvelet transform (CVT) [50], dual-tree complex wavelet transform (DTCWT) [51], spatial fre- quency (SF) [13], discrete wavelet transform (DWT) [52], ratio of low-pass pyramid (RP) [53], and Laplacian pyramid (LP) [11]. Specifically, we further show detailed comparison of each image pair with nine SOTA deep learning based methods in Fig. 5. With two of them are DM based methods (CNN-Fuse and SESF-Fuse), and six of them are decoder based methods (MFF-GAN, FusionDN, U2Fusion, DenseFuse, IFCNN and PMGI). In addition, for CNN-Fuse and SESF-Fuse, we also compare different versions of whether to use post-processing (pp) methods (or consistency verification) with empirical parameters. According to experiment, DM based algorithms generate an intermediate decision map to decide which pixel should appear in the fused result, which can precisely pre- serve true pixel values of the source image. And decoder based algorithms directly use a decoder to draw out the fused result and cannot preserve true pixel values because of the nonlinear mapping mechanism in the decoder. Therefore, DM based algo- rithms achieve high performance in objective assessments, while decoder based algorithms show unrealistic performance. In addi- tion, DM based algorithms rely on post-processing methods to rec- tify DM, so the performance will degrade if we remove it. Our algorithm, GACN can simultaneously generate decision map and fused result with end-to-end training procedure, and gradient information can be preserved by the gradient loss function. Our method, achieves robust promising performance compared to above traditional methods.

In addition, the run times of different fusion methods per image pair on the test set are listed in Table 1. Such methods as GACN, MFF-GAN, MFF-SSIM, DRPL, FusionDN, U2Fusion, IFCNN, PMGI, SESF-Fuse, CNN-Fuse, and DenseFuse are tested on a GTX 1080Ti GPU, and others on an E5-2620 CPU. GACN achieves an average running time of 0.16 s, which is faster than most of the methods and can be applied to actual application. Although the IFCNN is fas- ter than GACN, it achieves lower fusion quality compared to GACN.

4.4.2. Subjective assessment

We show some visualization results of GACN and other SOTA methods, DM based and decoder based methods, respectively. Firstly, we present the decision maps of GACN with some classical DM based methods (CNN-Fuse and SESF-Fuse) in Fig. 6. The influ- ence of post processing method is shown in detail. According to the experiment, the SESF-Fuse and CNN-Fuse require post-processing methods with empirical parameters, such as morphology opera- tion and small size removal strategy, to eliminate noise. If we remove these post processing methods, there will be some artifacts that appear on the results, such as blob noisy in the decision map. Besides, the threshold of kernel size in morphology operation and region removal strategy are empirical parameters which hard to adjust. While our method GACN can draw out good decision map without post-processing methods.

Secondly, we demonstrate the fusion results and difference images of GACN with some classical decoder based methods (Dense-Fuse, PMGI, FusionDN, U2Fusion and MFF-GAN) in Fig. 7. The red rectangles and their magnified regions (shown in upper right of the figure) denote the detailed fusion results of different methods. It is shown that there is artifact area at the border of near-focused and far-focused regions for the classical decoder based methods. While GACN shows clear result. The difference

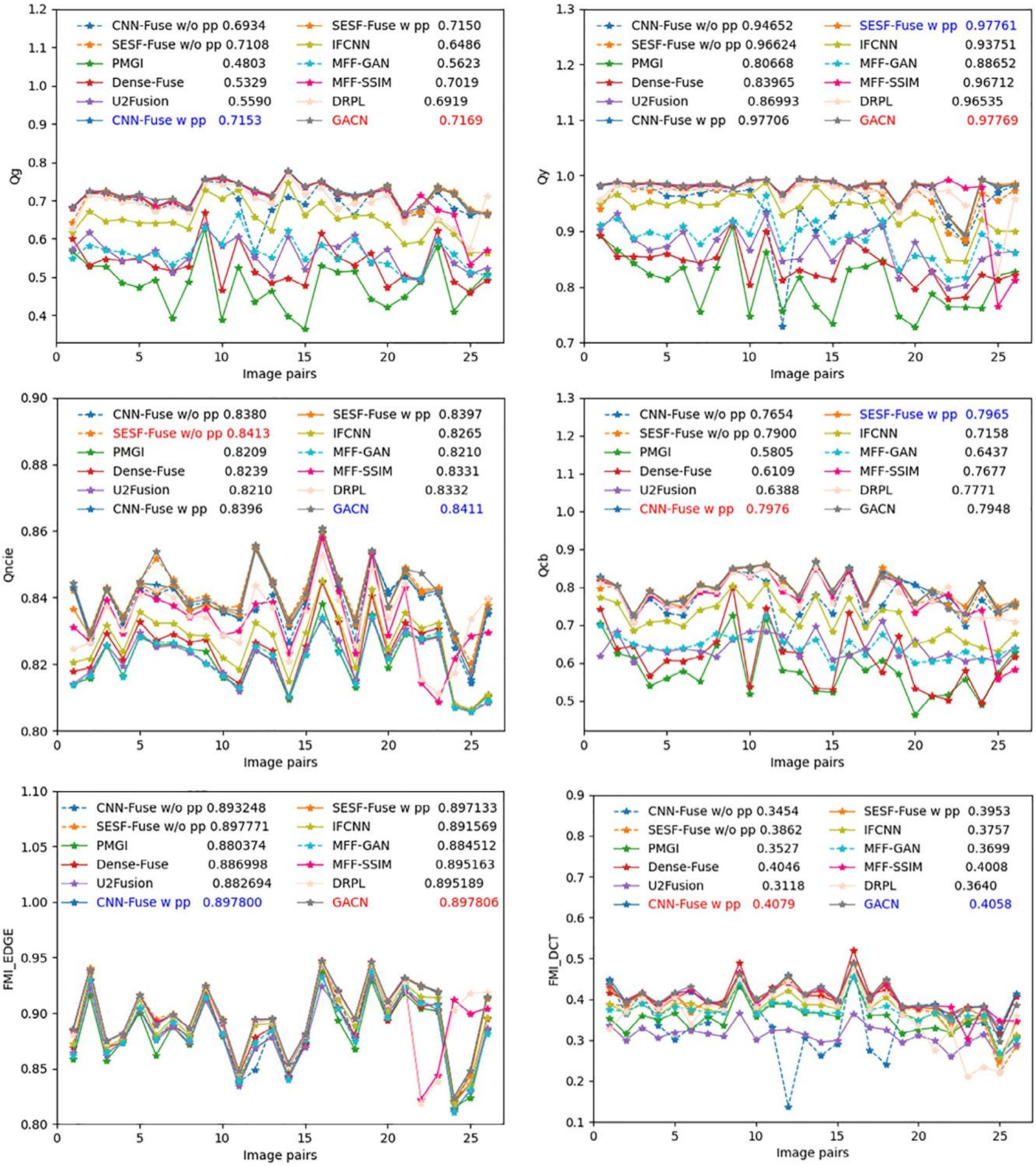

Fig. 5. Objective Assessments of our GACN with other SOTA algorithms. Means of metrics for different algorithms are shown in the legends, and evaluation for each image pair is shown in the plot. Optimal values are shown in red and sub-optimal values in blue. 'pp' means post processing methods.

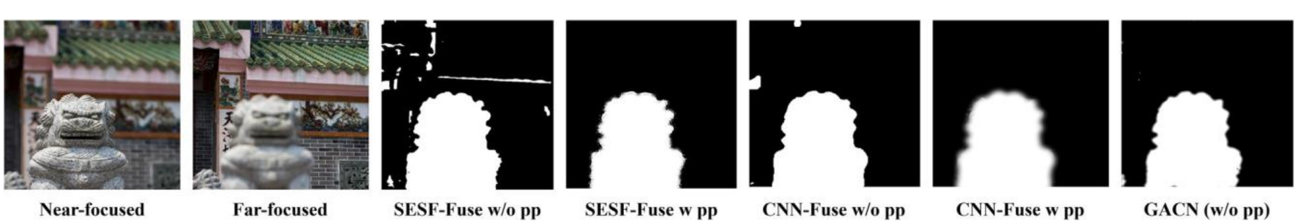

Fig. 6. Visualization of decision map of DM based methods (SESF-Fuse and CNN-Fuse) and GACN. pp means post processing methods.

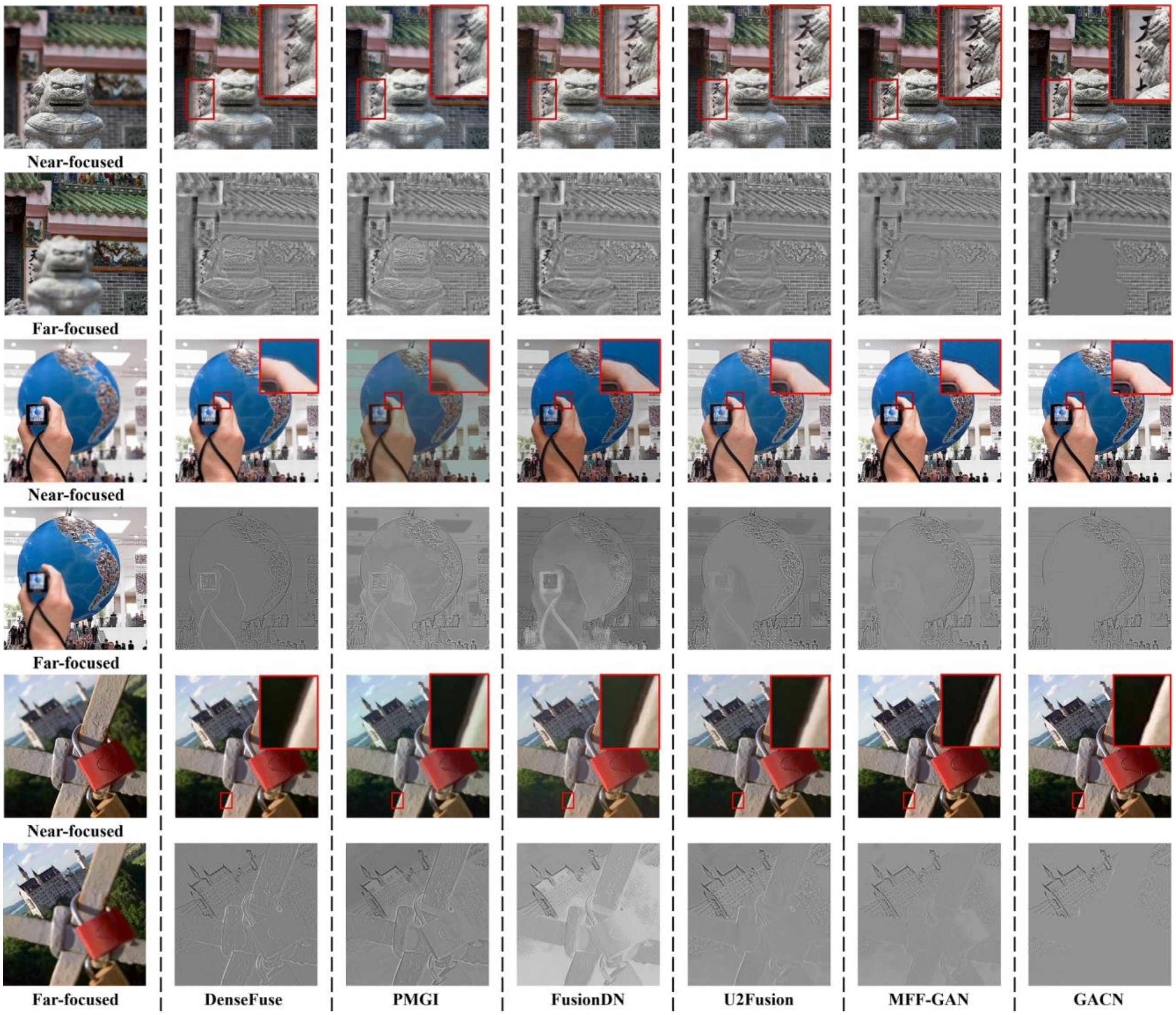

Fig. 7. Visualization of fusion result and difference images of decoder based methods (Dense-Fuse, PMGI, FusionDN, U2Fusion, and MFF-GAN) and GACN. For each example, the top image is fusion result and the bottom image is difference image, which is obtained by subtracting the near-focused image from the fusion result.

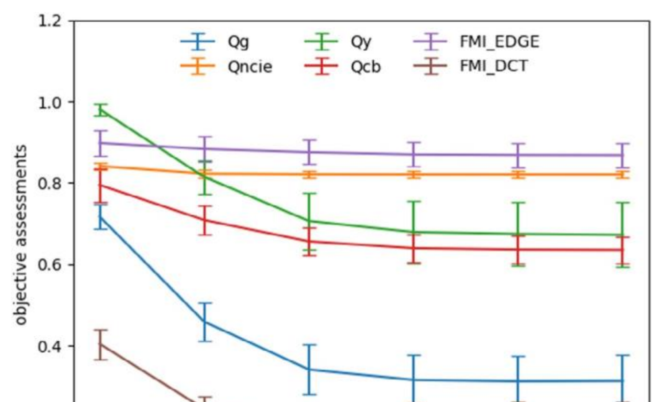

Fig. 8. Variation of different metrics.

image is obtained by subtracting the near-focused image from the fusion result, which is normalized to the range of 0 to 1 for visual- ization. If the near-focused region is completely detected, the dif- ference image will not show any of its information. Decoder based methods cannot precisely recover the true pixel values in fusion result due to the nonlinear mapping mechanism in the decoder. Therefore most of them have clear contour information in the near-focused region on the difference images. Besides, there is some color distortion in the fusion result of PMGI. And the fusion result of DenseFuse is more blurred than other methods. Fortu- nately, our method, GACN, achieves robust promising fusing per- formance on all examples.

4.5. *Ablation study*

We evaluate our method with different settings to verify the contribution of each module.

4.5.1. *Loss function study*

We first conducted an experiment to figure out which metric is more suitable for evaluation of quality of multi-focus image fusion. We introduce Gaussian blurring with different standard deviations

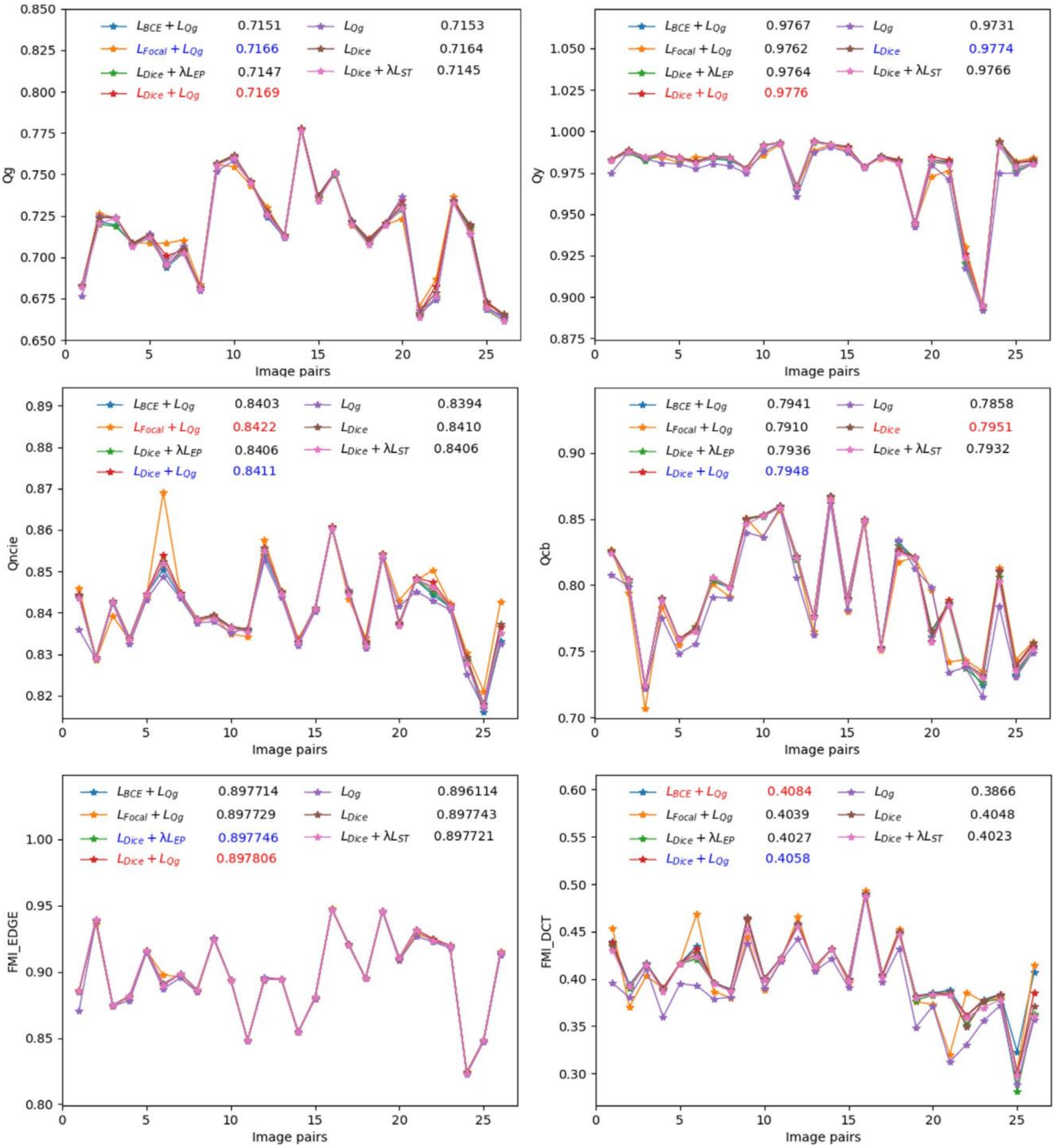

Fig. 9. Loss function analysis.

Table 2
Differentiation Comparison. 'Abs' means absolute function, and 'Smooth' denotes smooth approximation.

Settings	Q_g	Q_y	Q_{ncie}
Abs	0.7169	0.9776	0.8411
Smooth	0.7162	0.9773	0.8410
Settings	Q_{cb}	FMI_EDGE	FMI_DCT
Abs	0.7948	0.8978	0.4058
Smooth	0.7952	0.8977	0.4048

Table 3
Time consumption per image for multiple 'chip' images fusion with multi-focus points. The bold value denotes the best performance in each method. CNN-Fuse is running on CPU mode according to its public code.

Runtime(s)	One by one serial	Decision calibration
CNN-Fuse	886.6872	687.5352
SESF-Fuse	1.1880	0.5293
GACN	0.7138	0.4905

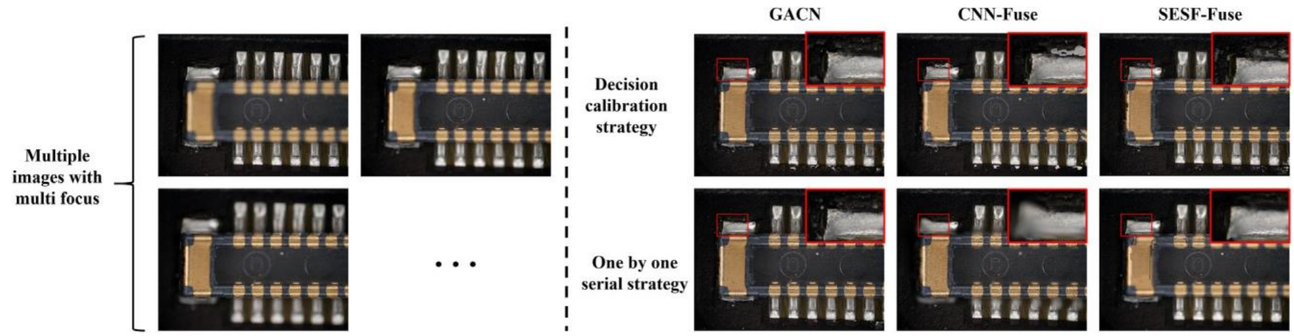

Fig. 10. Visualization of multiple images fusion.

to the fusion result of the testing set. As shown in Fig. 8, with the increase of standard deviation of Gaussian kernel, the metric Q_g degenerates most obviously compared to other metrics. It is shown that the metric Q_g can better reflect the clarity of the fusion result, which means that metric Q_g is beneficial to be the loss function for model training.

In addition, we compared the performance of the different com- binations of mask-based and gradient-based loss functions to ver- ify the contribution of proposed loss functions, shown in Fig. 9. The mask based loss functions include L_{Dice} [28], L_{Focal} [54], and L_{BCE} [55]. While gradient-based loss functions include L_{Q_g}; L_{EG} [21], and L_{ST} [19]. L_{BCE} denotes balanced cross entropy which is a classical loss function in image segmentation [55], which can eliminate the impact of imbalance pixels in the foreground and background. L_{Focal} denotes focal loss [54], which leads the network to focus on and correctly detect hard examples. Where *EG* refers to edge- preserving loss and ST means structure tensor loss. For the last two losses, we conducted an experiment and pick the best perfor- mance with k ¼ 0:0001 to balance the importance with L_{Dice}. According to the experiment, we noted that the performance of the combination of L_{Dice} and L_{Q_g} outperforms other loss settings in most the metrics, which means that the above two losses will both lead the network to export promising fusing result. Besides, we find that L_{Dice} is better than L_{BCE}, and L_{Focal}, which means that L_{Dice} can precisely recognize the decision map. And L_{Q_g} is better than L_{EP}, and L_{ST} , which means that L_{Q_g} can better lead the structure to preserve gradient information in the fused result.

4.5.2. Differentiation study

We compared the performance of the absolute function and the smooth approximation for angle calculation (Eq. 11) in L_{Q_g} in Table 2. We found that directly using the absolute function is a lit- tle better than the smooth approximation by using pytorch frame- work, which might be the reason for the gradient vanish in the sigmoid calculation.

4.6. Multiple images fusion with multi-focus

The example of multiple images fusion is shown in Table 3 and Fig. 10. The microscopic image 'chip' (with the size of 2700 x 1800) was obtained by a microscope that took pictures with lots of differ- ent focus points. Decision calibration for 'chip' images fusion can actually increase execution efficiency by about 30.65% compared to one-by-one serial strategy (0.7138's to 0.4905's for each image by using GACN), which is more feasible for industrial application. And the same increase of efficiency can also be found in CNN- Fuse and SESF-Fuse, which means that the decision calibration can be applied to other networks. Note that for decision calibra- tion, we deleted the post-processing operations of DM in CNN- Fuse and SESF-Fuse for fair comparison.

The visualization of fusion result of GACN is more clear than that of CNN-Fuse and SESF-Fuse whether in decision calibration or serial strategy. The decision calibration strategy can reduce nearly half of time cost during the image feature extraction process of the decision-map-based image fusion methods. But it requires the feature extraction module to have strong and effective feature expression ability, otherwise it will bring error propagation in the multi images fusion result. Although the SOTA decision-map-based image fusion methods are well trained, the error propagation prob- lem can also influence the objective assessment of the fusion result to a certain extent. In the future work, we try to overcome the error propagation problem as well as eliminate the impact of defocused spread effect for multi-focus image fusion.

5. Conclusion

In this work, we propose a network to simultaneously generate decision map and fused result with an end-to-end training proce- dure. It avoids utilizing empirical post-processing methods in the inference stage. Besides we introduce a gradient aware loss func- tion to lead the network to preserve gradient information. Also we design a decision calibration strategy to fuse multiple images, which can increase implementation efficiency. Extensive experi- ments are conducted to compare with existing SOTA multi-focus image fusion structures, which shows that our designed structure can generally ameliorate the output fused image quality for multi-focus images, and increase implementation efficiency over 30% for multiple images fusion. We will further improve the fusion performance of multiple images fusion in future work.

CRediT authorship contribution statement

Boyuan Ma: Conceptualization and Writing - original draft. Xiang Yin: Data curation and Formal analysis. Di Wu: Writing - review and editing. Haokai Shen: Validation. Xiaojuan Ban: Supervision. Yu Wang: Funding acquistion.

Declaration of Competing Interest

The authors declare that they have no known competing finan- cial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank professor Jiayi Ma of Wuhan University for the advice about the visualization of subjective assessment. In addition, we thank Zhuhai Boming Vision Technology Co., Ltd for providing the dataset of multiple images fusion and LetPub for its linguistic assistance during the preparation of this manuscript. This work was supported by the National Key Research and Development Program of China under Grant 2019YFC0605300, National Natural Science Foundation of China under Grant 6210020684 and Grant 61873299, Scientific and Technological Innovation Foundation of Shunde Graduate School of USTB under Grant BK19AE034 and Grant BK20AF001 and Grant BK21BF002, and Fundamental Research Funds for the Central Universities of China under Grant 00007467 and Grant FRF-TP- 19-043A2. The computing work is supported by USTB MatCom of Beijing Advanced Innovation Center for Materials Genome Engineering.

References

[1] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of the state of the art, Information Fusion 33 (2017) 100–112, https://doi.org/ 10.1016/j.inffus.2016.05.004.

[2] L. Yann, B. Yoshua, H. Geoffrey, Deep learning, Nature 521 (7553) (2015) 436–444.

[3] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Information Fusion 36 (2017) 191–207, https:// doi.org/10.1016/j.inffus.2016.12.001.

[4] B. Ma, Y. Zhu, X. Yin, X. Ban, H. Huang, Sesf-fuse: An unsupervised deep model for multi-focus image fusion, arXiv (2019)..

[5] X. Zhang, Deep learning-based multi-focus image fusion: A survey and a comparative study, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1, https://doi.org/10.1109/TPAMI.2021.3078906.

[6] J.H. Huang, Z. Le, Y.T. Ma, X. Mei, F. Fan, A generative adversarial network with adaptive constraints for multi-focus image fusion, Neural Computing and Applications (2020) 1–11.

[7] H. Zhang, H. Xu, Y. Xiao, X. Guo, J. Ma, Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 12797–12804..

[8] H. Xu, F. Fan, H. Zhang, Z. Le, J. Huang, A deep model for multi-focus image fusion based on gradients and connected regions, IEEE Access 8 (2020) 26316– 26327.

[9] H. Li, X. Wu, Densefuse: A fusion approach to infrared and visible images, IEEE Transactions on Image Processing 28 (2019) 2614–2623.

[10] K. Ram Prabhakar, V. Sai Srikar, R. Venkatesh Babu, Deepfuse, A deep unsupervised approach for exposure fusion with extreme exposure image pairs, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4714–4722.

[11] P. Burt, E. Adelson, The laplacian pyramid as a compact image code, IEEE Transactions on Communications 31 (4) (1983) 532–540, https://doi.org/ 10.1109/TCOM.1983.1095851.

[12] Q. Zhang, B. long Guo, Multifocus image fusion using the nonsubsampled contourlet transform, Signal Processing 89 (7) (2009) 1334–1346, https://doi. org/10.1016/j.sigpro.2009.01.012.

[13] S. Li, J.T. Kwok, Y. Wang, Combination of images with diverse focuses using the spatial frequency, Information Fusion 2 (3) (2001) 169–176, https://doi.org/ 10.1016/S1566-2535(01)00038-0.

[14] Z. Zhou, S. Li, B. Wang, Multi-scale weighted gradient-based fusion for multi- focus images, Information Fusion 20 (2014) 60–72, https://doi.org/10.1016/j. inffus.2013.11.005.

[15] Y. Liu, S. Liu, Z. Wang, Multi-focus image fusion with dense sift, Information Fusion 23 (2015) 139–155, https://doi.org/10.1016/j.inffus.2014.05.004.

[16] H. Tang, B. Xiao, W. Li, G. Wang, Pixel convolutional neural network for multi- focus image fusion, Information Sciences (2018) 125–141, https://doi.org/ 10.1016/j.ins.2017.12.043.

[17] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2fusion: A unified unsupervised image fusion network, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[18] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, L. Zhang, Ifcnn: A general image fusion framework based on convolutional neural network, Information Fusion 54 (2020) 99–118.

[19] J.H.K.Y.J.H.H.N. Sohn, Unsupervised deep image fusion with structure tensor representations, IEEE Transactions on Image Processing 29 (2020) 3845–3858.

[20] H. Zhang, Z. Le, Z. Shao, H. Xu, J. Ma, Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi- focus image fusion, Information Fusion 66 (2021) 40–53, https://doi.org/ 10.1016/j.inffus.2020.08.022.

[21] J. Li, X. Guo, G. Lu, B. Zhang, Y. Xu, F. Wu, D. Zhang, Drpl: Deep regression pair learning for multi-focus image fusion, IEEE Transactions on Image Processing 29 (2020) 4816–4831.

[22] K. He, J. Sun, X. Tang, Guided image filtering, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (6) (2013) 1397–1409, https://doi.org/ 10.1109/TPAMI.2012.213.

[23] Z. Sergey, K. Nikos, Learning to compare image patches via convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4353–4361.

[24] H. Gao, L. Zhuang, V.D.M. Laurens, W.K. Q, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708..

[25] H. Xu, J. Ma, Z. Le, J. Jiang, X. Guo, Fusiondn: A unified densely connected network for image fusion, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 12484–12491..

[26] H. Jie, S. Li, S. Gang, Squeeze-and-excitation networks, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[27] R.A. Guha, N. Nassir, W. Christian, Concurrent spatial and channel squeeze and excitation in fully convolutional networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 421–429.

[28] M. Fausto, N. Nassir, A. Seyedahmad, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: International Conference on 3D Vision, 2016, pp. 565–571.

[29] C.S. Xydeas, V. Petrovic, Objective image fusion performance measure, Electronics Letters 36 (4) (2000) 308–309, https://doi.org/10.1049/ el:20000267.

[30] P. Adam, G. Sam, M. Francisco, L. Adam, B. James, C. Gregory, K. Trevor, L. Zeming, G. Natalia, A. Luca, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8026–8037..

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755..

[32] H. Ma, Q. Liao, J. Zhang, S. Liu, J.H. Xue, An α-matte boundary defocus model-based cascaded network for multi-focus image fusion, IEEE Transactions on Image Processing 29 (2020) 8668–8679.

[33] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary- based sparse representation, Information Fusion 25 (2015) 72–84, https://doi. org/10.1016/j.inffus.2014.10.004.

[34] S. Savic´, Z. Babic´, Multifocus image fusion based on empirical mode decomposition, in: 19th IEEE International Conference on Systems, Signals and Image Processing, 2012, pp. 91–94.

[35] B. Ma, X. Ban, H. Huang, Y. Chen, W. Liu, Y. Zhi, Deep learning-based image segmentation for al-la alloy microscopic images, Symmetry 10 (4) (2018) 107.

[36] B. Ma, X. Wei, C. Liu, X. Ban, H. Huang, H. Wang, W. Xue, S. Wu, M. Gao, Q. Shen, et al., Data augmentation in microscopic images for material data mining, NPJ Computational Materials 6 (1) (2020) 1–9.

[37] C. Chen, K. Zhou, M. Zha, X. Qu, X. Guo, H. Chen, Z. Wang, R. Xiao, An effective deep neural network for lung lesions segmentation from covid-19 ct images, IEEE Transactions on Industrial Informatics (2021).

[38] C. Chen, R. Xiao, T. Zhang, Y. Lu, X. Guo, J. Wang, H. Chen, Z. Wang, Pathological lung segmentation in chest ct images based on improved random walker, Computer Methods and Programs in Biomedicine 200 (2021) 105864.

[39] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015, pp. 1–15..

[40] Z. Li, J. He, X. Zhang, H. Fu, J. Qin, Toward high accuracy and visualization: An interpretable feature extraction method based on genetic programming and non-overlap degree, in: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2020, pp. 299–304.

[41] C. Yang, J.Q. Zhang, X.R. Wang, X. Liu, A novel similarity based quality metric for image fusion, Information Fusion 9 (2) (2008) 156–160.

[42] W. Qiang, S. Yi, J. Jing, Performance evaluation of image fusion techniques, Image Fusion: Algorithms and Applications 19 (2008) 469–492.

[43] Y. Chen, R.S. Blum, A new automated quality assessment algorithm for image fusion, Image and Vision Computing 27 (10) (2009) 1421–1432, https://doi. org/10.1016/j.imavis.2007.12.002.

[44] M.B.A. Haghighat, A. Aghagolzadeh, H. Seyedarabi, A non-reference image fusion metric based on mutual information of image features, Computers and Electrical Engineering 37 (5) (2011) 744–756, https://doi.org/10.1016/ j.compeleceng.2011.07.012.

[45] Z. Liu, Image fusion metrics, https://github.com/zhengliu6699/ imageFusionMetrics (2012)..

[46] H. Mohammad, R.M. Amirkabiri, Fast-fmi: non-reference image fusion metric, in: 2014 IEEE 8th International Conference on Application of Information and Communication Technologies, IEEE, 2014, pp. 1–3.

[47] S. Xu, L. Ji, Z. Wang, P. Li, K. Sun, C. Zhang, J. Zhang, Towards reducing severe defocus spread effects for multi-focus image fusion via an optimization based strategy, IEEE Transactions on Computational Imaging 6 (2020) 1561–1570.

[48] C. Andrew, Focus stacking made easy with photoshop, https://github.com/ cmcguinness/focusstack (2013)..

[49] B. Yang, Li, Multifocus image fusion and restoration with sparse representation, IEEE Transactions on Instrumentation and Measurement 59 (4) (2010) 884–892, https://doi.org/10.1109/TIM.2009.2026612.

[50] F. Nencini, A. Garzelli, S. Baronti, L. Alparone, Remote sensing image fusion using the curvelet transform, Information Fusion 8 (2) (2007) 143–156, https://doi.org/10.1016/j.inffus.2006.02.001.

[51] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Information Fusion 8 (2) (2007) 119–130, https://doi.org/10.1016/j.inffus.2005.09.006.
[52] H. Li, B. Manjunath, S. Mitra, Multisensor image fusion using the wavelet transform, Graphical Models and Image Processing 57 (3) (1995) 235–245, https://doi.org/10.1006/gmip.1995.1022.
[53] A. Toet, Image fusion by a ratio of low-pass pyramid, Pattern Recognition Letters 9 (4) (1989) 245–253, https://doi.org/10.1016/0167-8655(89)90003-2.
[54] L. Tsung-Yi, G. Priya, G. Ross, H. Kaiming, D. Piotr, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[55] X. Saining, T. Zhuowen, Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1395–1403.

Boyuan Ma received the B.E. degree in information engineering from Beijing Technology and Business University in 2015, the M.E degree in computer tech- nology from University of Science and Technology Bei- jing (USTB) in 2017, and the Ph.D. degree in computer science and engineering from USTB, Beijing, China, in 2021. He is currently a associate professor in the University of Science and Technology Beijing (USTB). His research interests include image fusion, image seg- mentation, and deep learning.

Xiang Yin received the B.E. degree from the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China, where he is currently pursuing the MA.Sc degree in the Arti- ficial Intelligence and 3D Visualization Lab, University of Science and Technology Beijing, Beijing. His research interests include image fusion, machine learning, deep learning and federated learning.

Di Wu received her BS degree from the Department of Automation, Beijing Institute of Technology, China in 2004. And she received PhD. on Pattern Recognition and Intelligent System at the School of Automation, Beijing Institute of Technology in 2010. She worked as Post- doc/Lecturer with the department of Computer and Communication Engineering in the University of Science and Technology Beijing, China. Currently she is a PhD candidate on Data Science in Norwegian University of Science and Technology. She has already published over
35 papers for international journals and conferences.
Her research interest covers smart data analysis as deep learning in multiple industrial

Haokai Shen received the B.E. degree in North Univer- sity of China University in 2015, the M.E degree in computer technology from University of China Univer- sity of Petroleum, Beijing (CUP) in 2020. He is working for North Automatic Control Technology Institute, Taiyuan, China and his research interests include image fusion and deep learning.

Xiaojuan Ban received the Ph.D. degree from the University of Science and Technology Beijing, Beijing, in 2003. She is currently a Ph.D. Supervisor with the University of Science and Technology Beijing (USTB). She has authored more than 300 articles. She is also the Managing Director of the Chinese Association for Arti- ficial Intelligence (CAAI). She is also a member of the standing committee of the human-computer interaction specialty and the theoretical computer science specialty in the China Computer Society (CCF). She has received the New Century Excellent Talent of the Ministry of Education.

Yu Wang received the Ph.D. degree from the School of Computer and Communication Engineering from University of Science and Technology Beijing, Beijing, in 2020. The B.E. degree in network engineering from National University of Defense Technology, Chang Sha, China. His research interests include pattern recognition and deep learning.