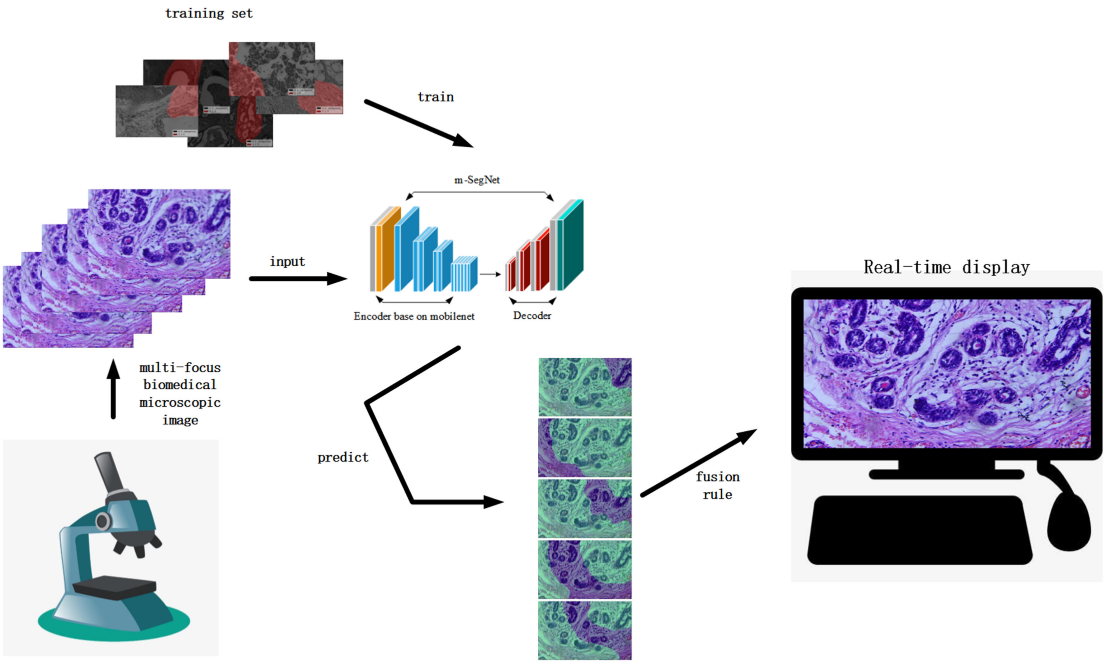
IEEE Photonics Journal
An IEEE Photonics Society Publication

Open Access

Real-Time Multi-Focus Biomedical Microscopic Image Fusion Based on m-SegNet

Volume 13, Number 3, June 2021

Ronghao Pei
Weiwei Fu
Kang Yao
Tianli Zheng
Shangshang Ding
Hetong Zhang
Yang Zhang



DOI: 10.1109/JPHOT.2021.3073022

IEEE photonics SOCIETY
IEEE

Real-Time Multi-Focus Biomedical Microscopic Image Fusion Based on m-SegNet

Ronghao Pei,[1,2] **Weiwei Fu**,[1,2] **Kang Yao**,[1] **Tianli Zheng**,[1] **Shangshang Ding**,[1] **Hetong Zhang**,[1] **and Yang Zhang**[2]

[1]School of Biomedical Engineering(Suzhou), Division of Life Science and Medicine, University of Science and Technology of China, Hefei 230026, China
[2]Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Science, Suzhou 215000, China

DOI:10.1109/JPHOT.2021.3073022

Manuscript received March 17, 2021; revised April 5, 2021; accepted April 8, 2021. Date of publication April 13, 2021; date of current version May 4, 2021. This work was supported in part by the Natural Science Foundation of Jiangsu province (BK20190197) and in part by Excellent member of Youth Innovation Promotion Association of Chinese Academy of Sciences (YIPA)(E1290301). Corresponding author: Yang Zhang (e-mail: zhangyang@sibet.ac.cn).
This article has supplementary downloadable material available at https://doi.org/10.1109/JPHOT.2021.3073022, provided by the authors.

Abstract: Activity level measurement and fusion rules are the two key factors of image fusion. In the fusion method based on neural networks, the activity level measurements are realized by dividing the image into small blocks and predicting the sharpness of each block; then, the global decision graph guiding fusion is generated according to the predicted results. However, these two tasks are serial in nature, which makes it difficult to complete them simultaneously while achieving satisfactory fusion performance. Therefore, a new multi-focus microscopic image fusion method is proposed in this paper to quickly fuse multiple histological microscopic images from different focusing planes to generate full-focus images. The improved SegNet network was used to detect the unfocused regions. Considering that two or more images are needed for fusion, a parallel fusion strategy is proposed herein to generate clear fusion images based on multiple images instead of pairwise decision graphs. Compared with the convolutional neural network, the proposed network has better representation ability and can extract and fuse the most ideal features to provide a more accurate fusion decision. Compared with the traditional Segnet network, it is lightweight, which greatly improves computing speed and achieves real-time fusion.

Index Terms: Histological microscopic image, Segnet network, multi-focus microscopic image fusion, parallel fusion strategy.

1. Introduction

The depth of field (DOF) parameter represents the imaging range within which a microscope can provide clear images after focusing. There is an inter-limiting relationship between the DOF and magnification of a microscope. When a sample is observed with a high-magnification microscope, the DOF is small, and only the parts located within a certain range of the focal plane will be clearly imaged, while other parts will be recorded with different degrees of blur. For example, in gland cell morphological observations, the gland cells form the main basis of cancer diagnosis, so obtaining clearly focused histological microscopic images can enable clinical diagnosis; however, only those cells that are near the focal plane or have clear structures can be imaged, while simple structures with depth information in relatively flat glandular cells cannot be completely clearly imaged. It is often necessary to constantly adjust the focal length of the microscope during observations to obtain clear images of samples, which causes inconvenience with respect to observation and diagnosis of the morphology of the cancerous gland cells. A common method to solve this problem is to use a multi-focus image fusion technology, which aims to collect more effective information by combining several images with different focal lengths into a full-focus image. At the same time, to ensure minimal loss in the time dimension, real-time fusion and real-time display are ideal, so the speed of the fusion algorithm is also an important requirement. In the past few years, several focused image fusion methods have been proposed. Based on the fusion strategies used, these methods can be divided into two categories: transformation-domain-based and spatial-domain-based methods.

Multi-focus image fusion in the transformation domain adopts multi-scale transformation (decomposition and reconstruction) to realize the fusion of multiple source images. In the early research on multi-scale transformation tools, Burt *et al.* [1] proposed the Laplacian pyramid (LP) transform in 1983, which has the advantage of being able to extract more high-frequency details from the source images. However, owing to the lack of directivity and redundancy of the decomposed data, fuzzy distortions are easily formed in the fusion image. Based on this, Toet *et al.* [2] proposed a series of improved tower transformation methods, such as contrast pyramid and morphology pyramid. Hui *et al.* [3] applied the discrete wavelet transform (DWT) to multi-focus image fusion in 1994. The adaptive wavelet denoising approach proposed by Xu *et al.* [4] provides higher accuracy for image fusion based on wavelet transform. To obtain better a multi-scale transformation tool, Do *et al.* [5] proposed a two-dimensional image representation method in the real sense: contourlet transform, which has the characteristics of multi-resolution and multi-direction but not translation invariance; moreover, it is easy to cause spectrum aliasing effects in the process of multi-scale transformation. Thereafter, non-subsampled contourlet transform (NSCT), dual-tree complex wavelet transform (DT-CWT), and shiftable complex directional pyramid transform (SCDPT) have been proposed. From the perspective of fused image quality, compared with the wavelet and contourlet transforms, a transform with translation invariance property (such as NSCT, SCDPT, and DT-CWT) has better fusion performance. However, owing to the complexity of multi-scale decomposition and reconstruction and the complexity of the image content, this method may easily cause loss of useful information from the source image (especially in the focus area). In recent years, a new transformation-domain fusion method [9]–[14] has been proposed. This method is not based on multi-scale transformation but converts images to single-scale feature domains with advanced signal representation theories, such as independent component analysis (ICA) and sparse representation (SR).

The spatial-domain-based method uses a block-based, region-based, or pixel-based fusion strategy to achieve the fusion. Aslantas *et al.* adopted the method of block segmentation to achieve multi-focus image fusion [18]; this method realizes segmentation and fusion of the source images according to the source image partitioning and fundamentally improves the correlations between the pixels. However, the main problem with this method is the selection of the image block size because unreasonable block sizes can easily cause the “block effect” in the fused image. In this regard, Zhang *et al.* [19] further adopted the particle swarm optimization algorithm to determine the optimal block size, which significantly improved fusion performance. Hariharan *et al.* [20] segmented the source image according to the connected regions, and these partitions were seamlessly stitched to form a fused image; this method uses focus connections and does not depend on the physical attributes of the image, so it can achieve better fusion. In addition, De *et al.* [21] improved the accuracy of segmentation of different regions in an image and reduced the “block effect” of the fused image based on continuous subdivision of the transition regions (between the focused and defocused regions) in the source image. The focused region segmentation method is another important approach in spatial multi-focus image fusion; this method divides and fuses the source image according to the focus region [22], so it generally does not produce the “block effect” in the fused image. However, the accuracy and integrity of region segmentation largely determine the performance of the algorithm, typical examples of which include guided filtering (GF) [15], dense SIFT (DSIFT) [16], and multi-scale weighted gradient (MWG) [17]. These methods have the advantages of simple calculation and fast fusion, but the obtained fused images have poor contrast, and some details may be lost.

With the wide application of deep-learning theory, researchers have proposed multi-focus image fusion methods based on convolutional neural networks (CNNs). Liu *et al.* [24] first proposed a CNN-based multi-focus fusion strategy, using the CNN trained by overlapping image blocks and their blurred version, and for each corresponding position in each image to be fused Predict the image blocks separately to form an initial decision diagram, and then obtain a more accurate decision diagram through a series of traditional image processing methods to guide the fusion Tang *et al.* [25] proposed a pixel-level CNN (P-CNN) to distinguish between the focus and defocus pixels, conversion of the P-CNN into the conventional image-CNN method, and directly using the image-CNN on the entire image, thereby reducing computation time. Guo *et al.* [26] used the full CNN to perform multi-focus image fusion; however, the initial decision mapping generated by the network still needs to be refined by a fully connected conditional random field (CRF). The above CNN-based method fundamentally determines the definition of the image block, which requires multiple uses of the CNN to predict artificially segmented blocks in a group of images to be fused, involves time complexity, and cannot meet real-time requirements. In addition, using this method to obtain the initial decision mapping generally involves fuzzy data and errors; at the boundaries of the focus and focal region, inaccuracies are caused and cannot be directly used for image fusion, such that the final figure used to guide fusion depends largely on the quality of the various post-processing technologies (such as morphology and guide filter etc.) that are not visible in an end-to-end manner. Li *et al.* [27] used the U-NET to perform semantic segmentation on two non-fully focused images to be fused to form a decision graph to guide fusion. Compared with the CNN-based methods, this method can better distinguish the focus and defocus areas in the source images and reduce the time complexity. However, the traditional U-NET network structure is still large. Although the speed is greater than the CNN-based methods, the fusion of multiple non-fully focused pathological microscopic images still fails to meet real-time requirements. In addition, the deep-learning methods ultimately require generating binary decision diagrams to guide image fusion; the decision-making step considers that the source image A is clear for pixel values of 0, whereas pixel values of 1 represent that source image B is clear. The corresponding pixels are able to do so because these methods would be geared to the needs of the scenery, the character such as simple photography scene two fusion, namely binary classification problems, the clear the area of A or B. However, in the field of biomedicine, the images are more detailed, especially when obtained with high-power microscopes, where there may be multiple clear images from the same field of view to be fused. Taking the DOF fusion of pathological sections of cancerous cells as an example, such high-magnification and high-resolution images motivate higher requirements for both the DOF and resolution. We believe that extending the DOF to five times the original (which is the optimal multiple obtained from a large number of experiments) can encompass the entire DOF of a normal thick section. That is to say, five microscopic images of non-fully focused pathological sections from different focal planes can be perfectly fused into a single full-focus image; however, the traditional U-Net method cannot generate a decision map that can guide the fusion of five images at once.

To solve the above problems, this paper proposes an improved SEGNET network architecture for multi-focus image fusion. The main contributions of this paper can be summarized as follows: (1) a SegNet model decoded by the lightweight mobile model is proposed, which enables our network to identify the non-focused regions in non-fully focused images in less time while reducing the misjudgments between the focused and non-focused edge regions; (2) an end-to-end multi-focus image fusion method with adaptive multi-image input is proposed. No longer need to generate decision figure, but with the segmentation results to guide the fusion SegNet model directly, in the original Image are more irregular repeat the sentence to a clear area, use the No - the Reference Image Quality Assessment using the Blur judge which original Image of the area more clearly, further enhance the efficiency of integration and discriminant accuracy. (3) The output of each level

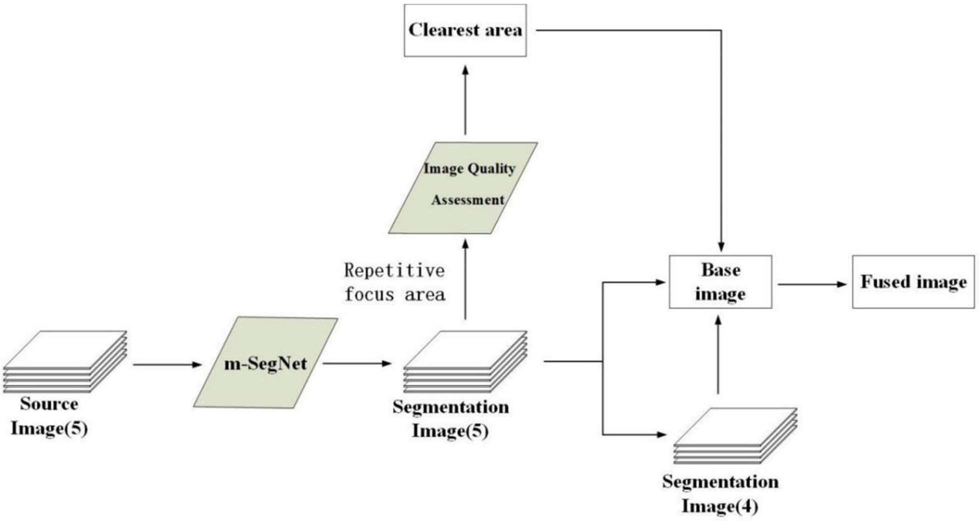


Fig. 1. Multi-focus image fusion scheme based on m-SegNet.

in the decoder is deeply supervised by the focal loss function, and it shows superior performance in the face of small samples with unbalanced categories

The remainder of this article is organized as follows. Section 2 introduce the multi-focus fusion method based on the m-SegNet network, including its structure, training, and evaluation, parallel fusion strategy, and post-processing steps for multiple source images that are repeatedly judged as having clear irregular regions in the fusion process. Section 3 presents the results of the fusion effect of the proposed method for a variety of cancer cell samples. The advantages of this method with respect to time complexity and clarity are discussed comparatively. Section 4 summarizes the observations of this study.

2. Multi-Focus Image Fusion Based on m-SegNet

The multi-focus biomedical microscopic image fusion scheme based on m-SegNet is shown in Fig. 1. After conducting a large number of experiments, we believe that in the imaging process of the pathological section of cancer cells, five images from different focal planes are sufficient to cover all the clear areas to allow perfect fusion of the full-focus pathological microscopic images. It can be seen from the figure that on the basis of the trained m-SegNet, our multi-focus microscopic image fusion approach mainly includes four steps. First, we use the built fast zoom microscope platform to collect five pathological images of cancer cells with different focus areas at the best focus plane of cancer cells, above and below the plane. Each source image is scored by m-SegNet according to pixels, and finally the labeled non focus area segmentation image is formed(not labeled is the focus area).Second, the clearest areas in the five segmented images are identified by a secondary blur algorithm to determine the image that provides the clearest focus area. Third, the source image corresponding to the one with the largest proportion of focal area is selected as the base image from among the five segmented images. Finally, the clearest area identified in step 2 and the focal areas from the remaining four source images are fused to the base image to create a full-focus pathological microscopic image.

2.1 Proposed m-SegNet

As described above, the multi-focus image fusion task is primarily a focus area detection task or a classification task that categorizes image blocks or regions into focused or defocused parts.

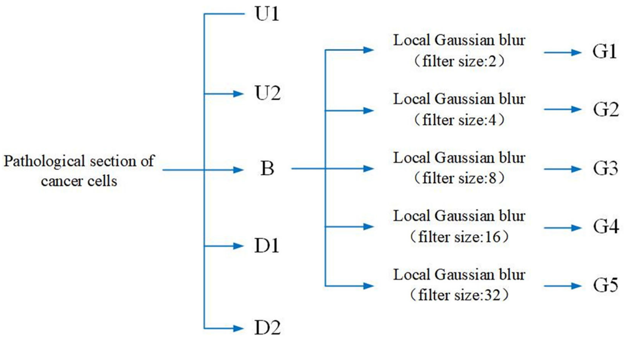

Pathological section of cancer cells
U1
U2
B
D1
D2
Local Gaussian blur (filter size:2)
Local Gaussian blur (filter size:4)
Local Gaussian blur (filter size:8)
Local Gaussian blur (filter size:16)
Local Gaussian blur (filter size:32)
G1
G2
G3
G4
G5

Fig. 2. Generate training images for the m-SegNet.

As the CNN approach has shown success in terms of image classification, many people have studied image fusion methods based on the CNN; this method is used to classify the artificially divided image blocks, guide the fusion decision diagram, perform figure-scale decision-making that is affected by the block size, especially the in-focus and focal parts of an image block, which may otherwise inevitably lead to inaccurate classifications. Therefore, most of the post-processing steps deal with misjudged blocks and imperfect boundaries. While the SegNet network is a kind of semantic network segmentation at the pixel level, the low-resolution encoder features of its decoder network can be mapped to the input resolution feature mapping to classify pixels directly, improve the accuracy of detection of the focusing area, combine the lightweight mobile model for SegNet network coding part, and significantly improve the computational efficiency with more focus on image fusion. In this section, we first introduce the set of generated training images, and then introduce the structure of the m-SegNet. Next, the efficiency of the proposed m-SegNet for identifying the focused and defocused pixels is demonstrated.

2.1.1 Training Image Set: A key requirement for training the m-SegNet for multi-focus biomedical microscopic images is a large number of labeled training images. However, there are no publicly available image databases with tags for focus and defocus images. Therefore, a training image set was created using various pathological microscopic images of cancer cells, which included focused images of correctly labeled partial regions, to meet the training requirements. In order for the m-SegNet network to learn more of the available features, we need to ensure some criteria: 1) a large enough sample size; 2) sufficient sample richness; 3) balanced distribution among the various samples.

In this work, we selected the pathological micrographs of 30 types of cancer cells, including lymph node adenocarcinoma, lymph node metastatic carcinoma, sarcoma carcinoma, colon carcinoid carcinoma, rectum adenocarcinoma, maternal carcinoma, and thyroid papillary carcinoma, among others, as the image database. Each type of cancer cell was captured from 10 different locations of the microscopic images, and each position contained 10 different focus levels of the image, including the optimum focus plane (B), the two best focus planes above it (U1, U2), optimum two focus planes below (D1, D2), and standard deviation for 2, 4, 8, 16, and 32, gaussian filter for optimum focus plane that a local area of fuzzy five (G1, G2 and G3, G4 and G5), as shown in Fig. 2. Each type of cancer cell thus had 100 partially focused pathological microscopic images, resulting in 3000 samples, 90% of which were used for training the m-SegNet network and the remaining 10% were used for testing. The deep-learning framework Keras 2.2.4 was adopted along with an Nvidia Quadro P620 GPU with 16 GB of RAM for training and testing.

2.1.2 Structure of the m-SegNet: Google has proposed MobileNet, which is a lightweight deep neural network for embedded devices such as mobile phones. The core idea of this method is to use depth-wise convolution (DEP) as its basic unit. Depth-level separable convolution is actually a factorized convolution, which can be decomposed into two smaller operations: depthwise convolution and pointwise convolution. It uses the method of depthwise and 1 × 1 pointwise operation to decompose the convolution instead of channel fusion when using 3 × 3 (or larger)

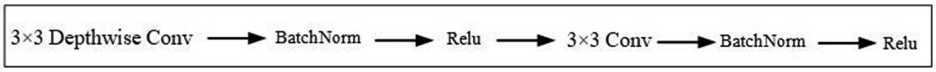


Fig. 3. MobileNet with depthwise convolution.

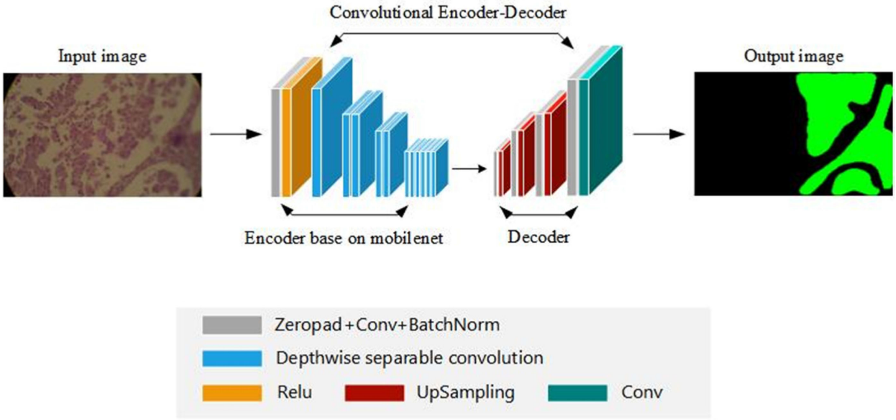

Convolutional Encoder-Decoder
Input image
Output image
Encoder base on mobilenet
Decoder
Zeropad+Conv+BatchNorm
Depthwise separable convolution
Relu
UpSampling
Conv

Fig. 4. Overview of m-SegNet network structure.

convolution. Depthwise convolution is used to conduct convolution on different input channels, and pointwise convolution is used to combine the above outputs. The overall effect is similar to that of a standard convolution, but it greatly reduces the amount of computation and number of model parameters.

In addition to the depthwise convolution, which is a basic component of Mobile Net, in this study, Batch Norm is added and the ReLU activation function is used. Therefore, the basic structure of (S) is as shown in Fig. 3. The SegNet network is based on a part of the classification network as an encoder, removing its fully connected layer and then adding a decoder at the back. The SegNet puts Softmax classification at the end, outputs a category probability for each pixel, and records the element in the pooled block that produces the largest result from pooling. When sampling on the decoder, it recovers according to the previously recorded pixel index value, so that more accurate results can be obtained theoretically. Among them, the first 13 layers of the VGG network are used in the encoder part. Although high classification accuracy can be achieved, the large size of the network affects classification efficiency and fails to meet the requirements for real-time performance. Thus, the network architecture was improved as shown in Fig. 4.on the premise of without changing the classification accuracy, as small as possible to set up each layer, the number of nerve cell using the first four layers as MobileNet decoders, each layer can be separated by one or more depth convolution units, the corresponding convolution, standardization, and many times in the decoder on sampling steps, end up with a number of the filter is the same as the input image size is 2 layers. That is, each pixel in the input image can be considered as one of two categories: clear or fuzzy. Although the network layer number is small, the experimental results show that ideal accuracy can be achieved, and the speed is significantly improved compared with the traditional SegNet network.

2.1.3 Loss Function of m-SegNet: In the process of network training, cross entropy is generally used as the loss function:

$$L = -\left[y\log\hat{y} + (1-y)\log(1-\hat{y})\right] \tag{1}$$

However, our training set positive and negative samples, that is, the focus area and the defocus area are not balanced. In most cases, the proportion of the positive sample is greater than the negative sample, as shown in Fig. 5. The red area is the positive sample, while the gray area is

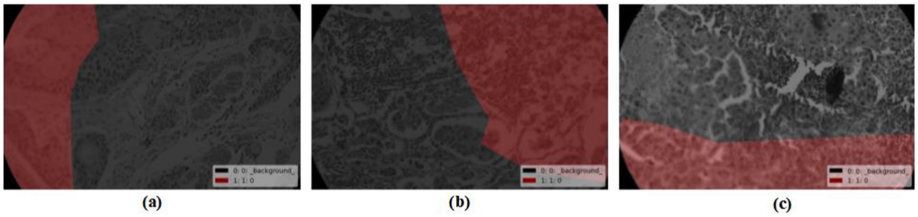


Fig. 5. Examples of unbalanced training samples.

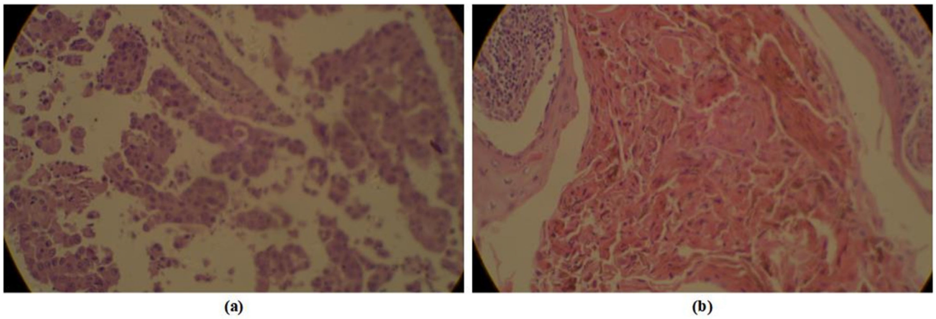


Fig. 6. Examples of sample classification difficulty differences.

the negative sample, and the proportion of the former is significantly smaller.In addition, due to the different degree of focus on each focal plane, the classification difficulty is also different. As shown in Fig. 6, the focusing region and defocusing region of (b) are not clearly distinguished, so the classification difficulty is obviously greater than that of (a).a large class imbalance during training will lead to cross-entropy loss. The easily divided negative sample occupies the dominant position in gradient and loss. Therefore, we adopted focal loss as the loss function during network training. This loss function is modified on the basis of the standard cross-entropy loss,used to solve the imbalance of classification and the difference in classification difficulty in classification problems. By adding suppression parameters and reducing the weight of easy-to-classify samples, the loss function is inclined to difficult-to-classify samples, so that the model is trained to focus on difficult-to-classify samples, thereby improving the accuracy of difficult-to-classify samples. The equation to compute focal loss is given as

$$L_{fl} = \begin{cases} -\alpha(1-\hat{\text{y}})^{\gamma}\log\hat{\text{y}}, \text{ while } y=1 \\ -(1-\alpha)\,\hat{\text{y}}^{\gamma}\log(1-\hat{\text{y}}), \text{ while } y=0 \end{cases} \tag{2}$$

Where ŷ is the probability that each pixel in the source image predicted by the network belongs to the fuzzy category, y is the true probability that each pixel in the source image marked in advance belongs to the fuzzy category and γ is the regulating factor.

For dichotomies, we almost always activate them with the sigmoid function such that

$$\hat{y} = \sigma(x), \textit{ and } 1 - \sigma(x) = \sigma(-x) \tag{3}$$

Therefore, the final form of the focal loss in the dichotomy is

$$L_{fl} = \begin{cases} -\alpha\sigma(-x)\log\sigma(x), \text{ while } y=1 \\ -(1-\alpha)\sigma^{\gamma}(x)\log\sigma(-x), \text{ while } y=0 \end{cases} \tag{4}$$

2.1.4 m-SegNet Training: We trained our model using the stochastic gradient descent method and used the focal loss function described above to monitor the depth of the output at each level of the encoder. We set the batch size to 16, and learning rate to 0.001. The learning rate is set to decline if val_loss does not decrease for three times, and this rate will be reduced to the original

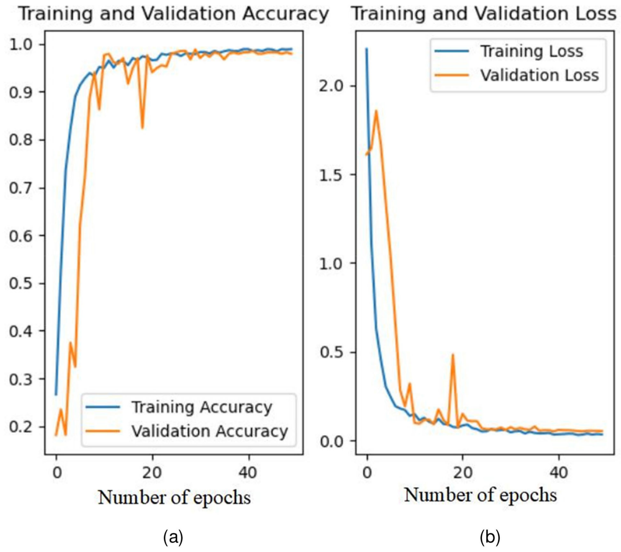


Fig. 7. Accuracy and loss of network training.

10%,aiming to improve the stability of training and ensure that training can converge to the optimal solution; we then set early stop aims to solve the problem that the number of epochs needs to be set manually: when it is found that the training of loss has not decreased compared with the last epoch, the training will stop after 10 epochs. In this way, the accuracy of training can reach the optimal solution and avoid over-fitting of the network. Finally, the training results are obtained as shown in Fig. 7, where (a) and (b) respectively show the accuracy and loss of m-SegNet during training,The optimal solution is reached after 50 epochs of training, and the early stop is triggered, and the training stops automatically. From the figure, the accuracy of m-SegNet is 98.25%. The proposed m-SegNet can thus effectively detect the focus and defocus regions.

2.2 Generating Split Maps for Each Source Image

We used the proposed m-SegNet network to perform semantic segmentations on five images from different focal planes and to judge and mark the clear regions in each image. Compared with the traditional CNN method for image block sharpness evaluation, the advantage of the proposed approach is that the pixel-based sharpness evaluation does not have obvious block effects, there is no unclear edge discrimination, and the clear part of each image can be accurately identified. As a result, all the clear parts are combined in the final fused image. At the same time, because the first four layers of MobileNet are used as the decoder for m-SegNet, the network can generate a segmentation map quickly, and the segmentation of each image only requires about 0.2 s, which guarantees real-time outputs for the full-focus images.

2.3 Searching for the Clearest Area in the Clearly Marked Sections

Relying on the accurate judgment of m-SegNet, the focus and non-focus regions in each focal plane image can be accurately identified. However, there is a clear overlap between images in different focal planes, and the image in which the focal plane shows the clearest region can cause confusion during fusion. Therefore, the no-reference image quality assessment using blur method is used to find the clearest area that is most convenient for fusion of these images from different focal planes.

For fuzzy estimation, there are two steps, as shown in Fig. 8 . The first step is edge detection, and the second is fuzzy determination. Here, the fuzzy estimation is obtained by calculating the

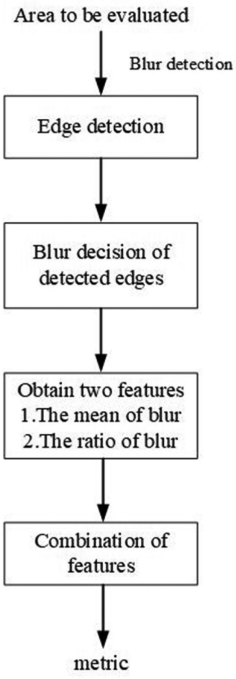


Fig. 8. No-reference image quality assessment using blur.

differences between the mean values of the pixel points in front of the point and those in the neighbourhood. The absolute difference of level is thus defined as follows:

$$D_h = |f(x, y+1) - f(x, y-1)| \tag{5}$$

where $f(x, y)$ is a clear overlap area judged from five images with different focus positions.

The mean of the absolute horizontal difference of the entire region is

$$D_{h-\text{mean}} = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} D_h(x, y) \tag{6}$$

If the current pixel point $D_h(x, y)$ is greater than $D_{h-\text{mean}}$, then this pixel is a candidate edge point $C_h(x, y)$; if$C_h(x, y)$ is larger than its two adjacent horizontal points $\{C_h(x, y-1), C_h(x, y+1)\}$, then this pixel is confirmed as an edge point. The judgment of the edge point $E_h(x, y)$ is summarized as follows:

$$C_h(x, y) = \begin{cases} D_h(x, y) \quad \text{while } D_h(x, y) > D_{h-mean} \\ 0 \quad \text{otherwiser} \end{cases} \tag{7}$$

$$E_h(x, y) = \begin{cases} 1 \quad \text{while } C_h(x, y) > C_h(x, y+1) \text{ and} \\ \qquad\qquad C_h(x, y) > C_h(x, y-1) \\ 0 \quad \text{otherwiser} \end{cases} \tag{8}$$

Next, we check to see if the edge points are blurry. We define

$$A_h(x, y) = \frac{1}{2} D_h(x, y) \tag{9}$$

$$BR_h(x, y) = \frac{|f(x, y) - A_h(x, y)|}{A_h(x, y)} \tag{10}$$

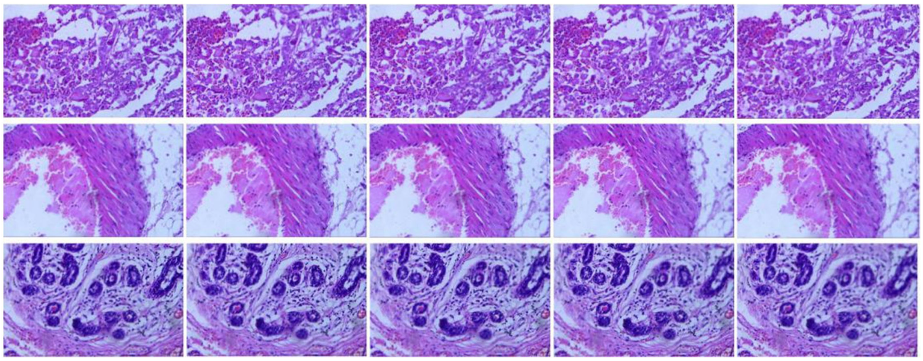


Fig. 9. Example of multi-focal microscopic images for multiple groups of cancer cells.

Similarly, the value BR_h in the vertical direction can be calculated by following the steps above. The larger value between BR_h and BR_v is called the inverse blurriness. For the final fuzzy decision basis, we have

$$B(x, y) = \begin{cases} 1 & if \ \max(BR_h, BR_h) < Th_B \\ 0 & otherwiser \end{cases} \tag{11}$$

Inverse blurriness below a threshold Th_B is considered fuzzy. The experiments show that the threshold here is 0.1. Finally, the mean and ratio of edge blur are given as

$$Blur_{mean} = \frac{Sum_{blur}}{Blur_{cnt}}, \quad Blur_{ratio} = \frac{Blur_{cnt}}{Edge_{cnt}} \tag{12}$$

where Sum_{blur},$Blur_{cnt}$ are the number of inverse blurriness and fuzzy points respectively; $Edge_{cnt}$ is the total number of edge points. The overlapping area definition index is given as

$$Mrtric = 1 - (w_1 Blur_{mean} + w_2 Blur_{ratio}) \tag{13}$$

where $w_1 = 1$, $w_2 = 0.95$.

2.4 Fusion to Generate Full-Focus Image

The five images from different focus planes were semantically segmented through the m-SegNet network, and the focus and non-focus regions were marked. One of the five images was randomly selected as the base map, and the focus regions in the remaining four images were integrated with the base map to replace the original regions. At the same time, the no-reference image quality assessment using blur algorithm was used to evaluate the sharpness of the overlapping parts in the focus areas of the five images, and the areas with the highest indexes were integrated with the base image. At this point, the base image aggregates all the focus areas that are provided by the five different focal planes to generate a full-focus biomedical microscopic image.

3. Experiments and Analysis

3.1 Experimental Set

To verify the effectiveness of the proposed method, 10 groups of multi-focus microscopic images of different types of cancer cells were collected as test data. The size of each image is 1936 × 1216 pixels, and some example images are shown in Fig. 9.

In the following experiments on the multi-focus microscopic image dataset of cancer cells, the proposed method is compared with other well-known multi-focus image fusion methods, namely the LP [1], non-subsampled contourlet transform (NSCT) [6], DWT [3], DTCWT [7], curvelet transform

(CVT) [5], CNN-shaped transform [24], and U-shaped transform [27]. In addition, the parameters of these methods are set to the recommended values from their original papers.

3.2 Fusion Quality Assessment

3.2.1 Evaluation Index: It is difficult to quantitatively evaluate the quality of multi-focus image fusion owing to the lack of corresponding full-focus images as the ground truths. To evaluate the performance of a fusion algorithm comprehensively, several evaluation indexes are thus needed. In recent years, a variety of fusion metrics have been proposed. The commonly used fusion metrics can be divided into four categories: based on information theory, based on image features, based on image structure similarity, and based on human perception. Therefore, five measures covering the above four methods were chosen to assess the results more comprehensively. For these measures, the higher their value, the better is the fusion quality. The details of these metrics are as follows:

1) Gradient-based measure $Q^{AB/F}$: This is an objective non-reference quality assessment index of the fused image; the algorithm to obtain $Q^{AB/F}$ uses the local measure to estimate the performance degree of the significant information from the input in the fused image. The higher the value of $Q^{AB/F}$, the better is the quality of the fused image. The specific definitions for this measure are as follows:

$$Q^{AF}(i,\ j) = Q_{\delta}^{AF}(i,\ j) * Q_{\sigma}^{AF}(i,\ j) \tag{14}$$

$$Q_k^{AB/F} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(Q^{AF}(i,\ j)W^A(i,\ j))}{\sum_{i=1}^{n_1}\sum_{i=1}^{n_1}(W^A(i,\ j)+W^B(i,\ j))} \tag{15}$$

2) The average gradient Q^{AVG}: This is also known as sharpness and reflects the contrast of minute details and texture changes in the image as well as sharpness of the image. The higher the value of Q^{AVG}, the better is the quality of the fused image. The specific definitions for this measure are as follows:

$$Q^{AVG} = \frac{1}{(M-1)(N-1)}\sum_{i=1}^{M-1}\sum_{j=1}^{N-1}\sqrt{\frac{(H(i+1,\ j)-H(i,\ j))^2+((H(i,\ j+1))-(H(i,\ j)))^2}{2}} \tag{16}$$

where H represents the fused image, and M and N represent the height and width of the image, respectively.

3) Structural similarity index SSIM: We denote the structural similarity between samples X and Y based on three comparative measures: luminance, contrast, and structure.
The value range of this index is [−1, 1], and the closer the value is to 1, the higher is the similarity and the better is the fusion quality. The specific definitions of this index are as follows:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2+\mu_y^2+c_1},\ c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2+\sigma_y^2+c_2},\ s(x,y) = \frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3} \tag{17}$$

$$SSIM(x,y) = \left[l(x,y)^{\alpha}\cdot c(x,y)^{\beta}\cdot s(x,y)^{\gamma}\right]$$

4) Peak signal-to-noise ratio PSNR: This value is based on the error between the corresponding pixel points, that is, based on error-sensitive image quality evaluations. This ratio is usually defined by the mean-square error (MSE):

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,\ j)-K(i,\ j)]^2 \tag{18}$$

$$PSNR = 10\cdot\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{19}$$

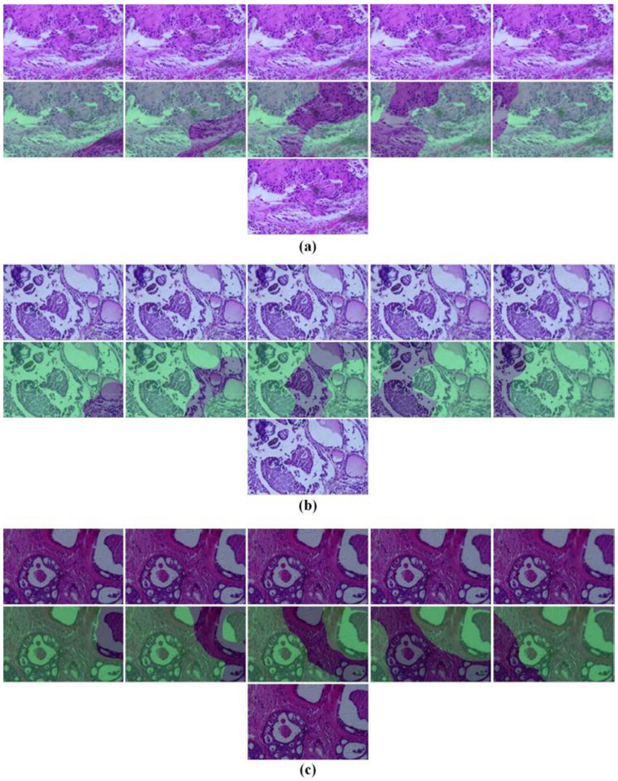


Fig. 10. Visualization results: (a) pathological microscopic images of lung cancer; (b) pathological microscopic images of thyroid papillary carcinoma; (c) pathological microscopic images of adenoid cystic carcinoma. The first row of each group is the microscopic image of a different focus condition under the same field of view; the second row shows the decision mapping generated by the proposed network; the third row shows the corresponding fusion result.

where *I* is the fused image, *K* is the source image, and *MAX* is the maximum possible pixel value of the image.

5) Mutual information MI: Mutual information measures the degree of similarity between two images, that is, the amount of information obtained about the original images from the fused image. The larger the mutual information, the more is the source image information retained by the fused image and the better is the quality. Mutual information is defined according to the information entropy H(A) and joint information entropy H(A,B) of the image:

$$MI(A, B) = H(A) + H(B) + H(A, B) \tag{20}$$

3.2.2 Experiments: The decision mapping generated by the network and the final fusion results are shown in Fig. 10, where the green area indicates the area with clear decision results while the unmarked area indicates the area with fuzzy decision results. The decision map generated by the proposed network is used to guide the fusion process. When the recognition accuracy of the decision map is high, there is no obvious misclassification region, and the boundary between the focus and defocus regions is clear. This means that the designed network can fully and clearly detect the focus region from the source image.

A total of 50 groups of multi-focus microscopic images of different types of cancer cells were collected, and each group included five images from different focus positions for performance comparison of the proposed method with other methods. The fusion results of the different methods are shown in Figs. 11–13. In addition, for better comparison, the difference images of each of the fusion results are also displayed, which are obtained by subtracting the source images from the fused images, as shown in Figs. 14–16. It is important to note here that if the focus regions are

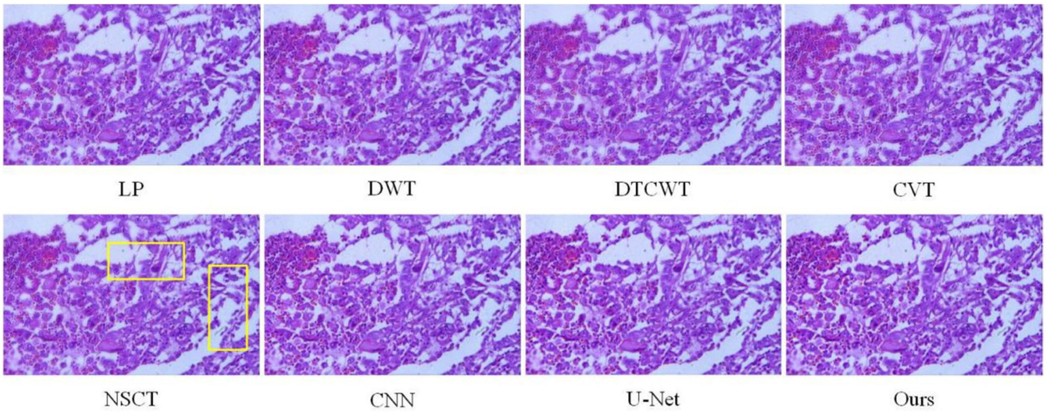


Fig. 11. Fusion results of pathological microscopic images of lung cancer using various methods.

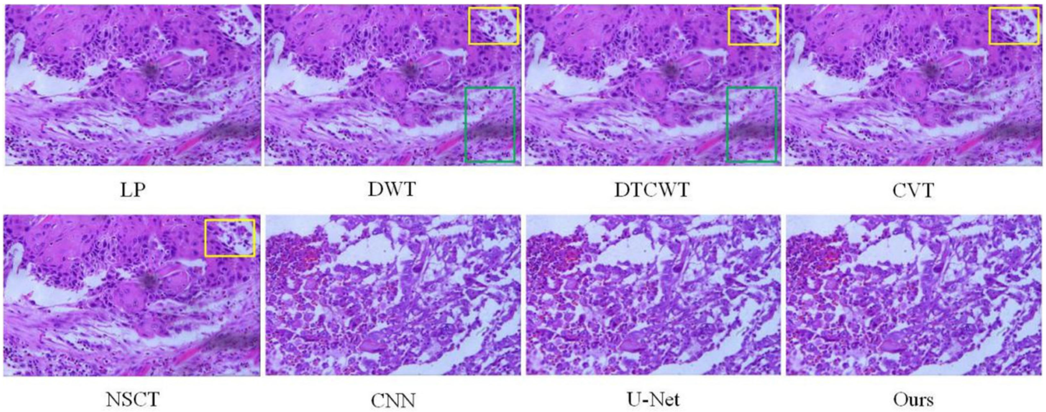


Fig. 12. Fusion of pathological microscopic images of keratinized squamous cell carcinoma using various methods.

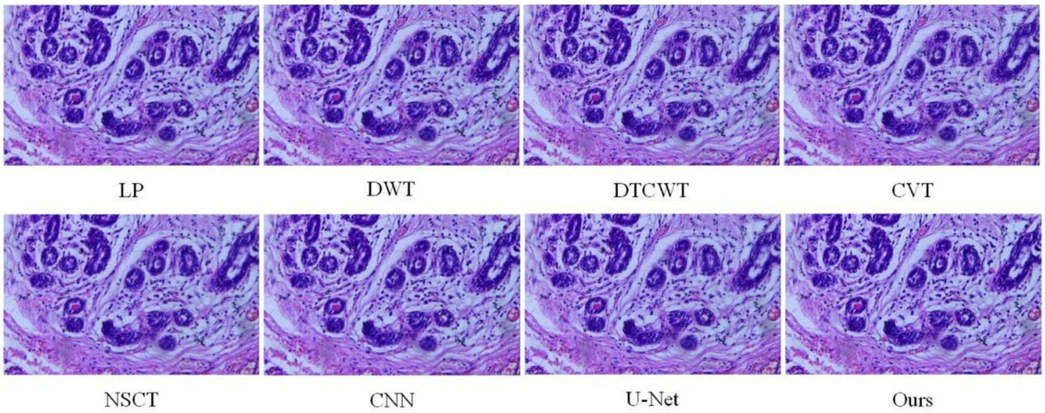


Fig. 13. Fusion of pathomicrographic images of infiltrating ductal carcinoma of the breast using various methods.

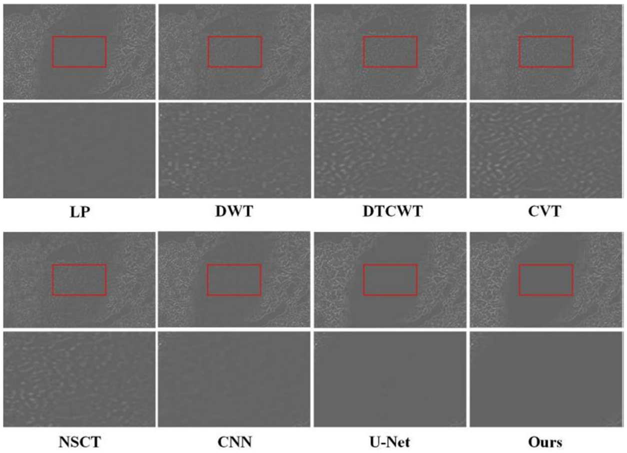


Fig. 14. Difference between local and full-focus pathological microscopic images of lung cancer.

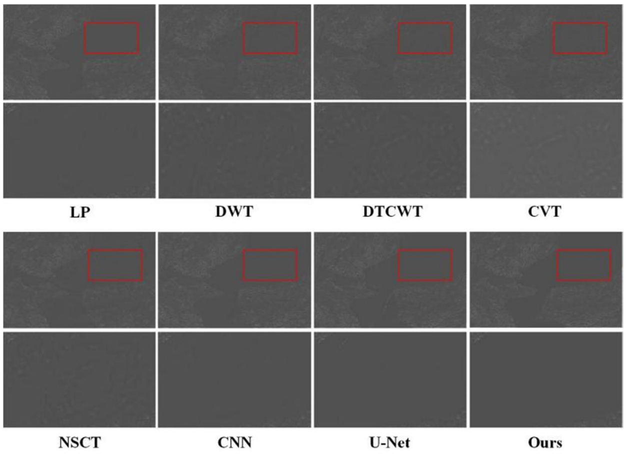


Fig. 15. Differences in pathomicrographic images of locally and fully focused keratinized squamous cell carcinoma.

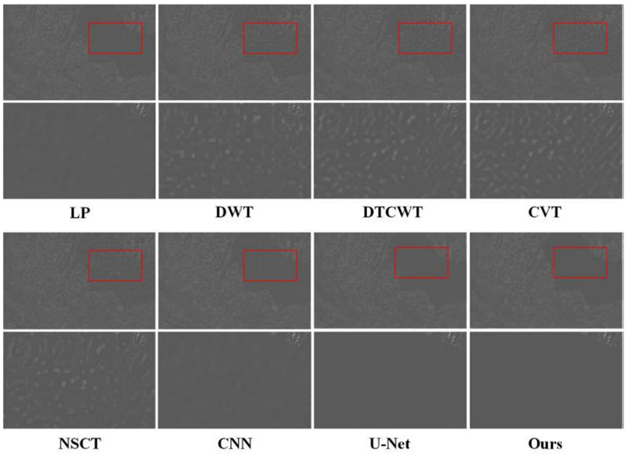


Fig. 16. Differences in pathomicrographic images of locally and fully focused infiltrating ductal carcinoma.

fully detected, there should be no residuals in the corresponding regions on the difference images. That is, in these differential images, lesser remaining traces of the focus areas mean more image information on the focus part being extracted from the fusion map, indicating better fusion effect. To facilitate observation, we show magnified focus areas for each of the differential images.

Fig. 11 shows the lung cancer pathological microscopic image fusion result; the DWT, DTCWT, and CVT image qualities are poor and overall fuzzy, while the LP, NSCT, U-Net, and proposed method have better the image qualities. However, the NSCT and LP still fail to achieve optimal fusion results in some areas, and these areas are marked with yellow boxes in the figure, showing that the fusion is relatively vague. Fig. 14 shows the differential image of pathological microscopic images of lung cancer. It is obvious that the less traces of the focused area left in different maps, the more image information of the focused part extracted from the fusion map, the better the fusion effect.the DWT, DTCWT, and CVT have obvious residual traces in the difference images, and the LP and CNN performances are relatively good, even with a little residual information. The U-Net and proposed method have no residual traces, indicating that the images are completely integrated.

Fig. 12 shows the keratinized squamous cell carcinoma pathological microscopic image fusion results; in terms of image quality, the performance is the most optimal for LP and CNN, while the DWT, DTCWT, CVT, U-Net, and proposed method have no degree of image quality variation. The difference here is that the in the DWT, DTCWT, and CVT, the fusion of NSCT is obviously not in the ideal area, which have been marked with yellow and green boxes in the figure. They are more obscure, and The black areas in the green boxes are contaminated pathological sections. These contaminants impact the clear and non-clear decisions for some of the image fusion rules. Fig. 15 shows the differential pathological microscopic images of the keratinized squamous cell carcinoma. In these images, the residual images of the various fusion methods all have small residuals, but it can still be seen that the DWT, DTCWT, and CVT have a lot of residuals, while LP has very subtle residuals, and the CNN, U-Net, and proposed method have no residuals.

Fig. 13 shows the fusion results of the pathological microscopic images of infiltrating ductal carcinoma of the breast. The NSCT image quality is significantly deteriorated, and the areas marked in red are clearly blurred. All the other methods show good performance. Fig. 16 shows the difference images of the pathological microscopic images of infiltrating ductal carcinoma of the breast, in which the residual traces of the DWT, DTCWT, CVT, and NSCT are still obvious, while the performances of LP and CNN are relatively good; however, there are still subtle residuals. The U-Net and proposed method have no residual traces, indicating that they are completely integrated.

3.2.3 Quantitative Results: To quantitatively evaluate the performance of the proposed method, we used the five measures described in Section 3.2 to comprehensively evaluate the fusion results of different methods. The average scores of the various methods on the five normalized measures above are listed in Table 1. The highest values are in red, and the second highest values are shown in green font.

It can be seen that the proposed m-SegNet method has the best score for all indicators except the QSSIM, for which it is inferior to the NSCT. This is because NSCT has multi-resolution, multi-directionality and shift invariance at the same time. As a pixel-level fusion method, it can accurately capture the clear pixels in the source image, which will show certain advantages in SSIM (an indicator used to measure the similarity between the fused image and grand truth).

However, In the deep learning method we used, the accuracy of the neural network largely depends on the labeling accuracy of the training set.While, due to the manual labelling, some pixel-level labelling cannot be completely accurate, so there are some labelling errors in some details. As shown in Fig. 17, the red outline surrounds the focus area marked by manual labelling, while the other areas are the defocus area.In the upper right corner of the picture, we zoom in on the area marked imperfect, it is obvious that some areas that should have been marked as focus are mistakenly marked as blur. These errors are only a small area, even a few pixels, which results in the m-SegNet method being 0.0013 lower than the NSCT method in the SSIM index.

The above situation is worthy of our attention.The neural network can improve the accuracy by adjusting various hyperparameters, but the essence of deep learning is the learning of training

TABLE 1
Objective Evaluation of Different Methods for Source Image Fusion

Metric	$Q^{AB/F}$	Q^{AVG}	SSIM	PSNR	MI
LP	0.7153	**6.7582**	0.9027	72.0151	6.8833
DWT	0.6872	6.7400	0.9020	72.0076	6.7999
DTCWT	0.7016	6.6477	0.9017	**72.1412**	6.7907
CVT	0.7992	6.6458	0.8976	72.0783	6.7711
NSCT	0.7043	6.6005	**0.9071**	72.0961	6.8429
CNN	0.7078	6.7390	0.9022	72.0239	6.8537
U-Net	**0.7197**	6.7505	0.9046	72.0686	**6.9017**
Proposed method	**0.7213**	**6.7724**	**0.9058**	**72.1435**	**6.9436**

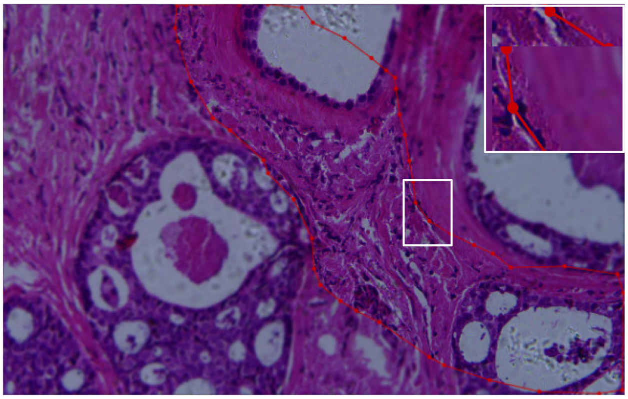


Fig. 17. Example of training set annotation error.

samples, and the accuracy upper limit of the neural network depends on the training set. The accurate training set labeling is beneficial to fundamentally improve the performance of the network.

3.3 Evaluation of Fusion Speed

To assess the computational efficiency, Table 2 shows the average run times of the various algorithms on a dataset of 50 multi-focus microscopic images of cancer cells. All eight methods are implemented in Python3.6 on a computer with an Intel(R) Core(TM) i7-8700 3.20 GHz CPU and 16 GB RAM, the size of the source image to be fused is 1936 × 1216 pixels. Obviously, the time consumed by the proposed method to generate the full-focus image is only 1.13 s, which

TABLE 2
Average Elapsed Time (s) for Different Fusion Methods

Method	LP	DWT	DTCWT	CVT
Time	2.37	9.88	21.44	66.78
Method	NSCT	CNN	U-Net	Proposed
Time	101.16	571.88	2.38	1.13

is the least among the compared algorithms.The reason why our algorithm can achieve such a speed is that the semantic segmentation network is based on the image as a whole to classify and segment, and the output results can be directly used as the decision map guiding fusion, which is end-to-end.In addition, our proposed algorithm is highly parallel, which can input five or more non-fully focused source images in parallel for fusion, rather than needing pairwise fusion as previous algorithms.Therefore, our proposed fusion algorithm based on m-SegNet has great advantages in accomplishing the real-time fusion task of multi-focus images.

4. Conclusion

In this paper, a multi-focus biomedical microscopic image fusion method based on m-SegNet is proposed. Herein, we first created a training image set to train the m-SegNet with pathological microscopic images of various cancer cells to learn the characteristic information for focused and defocused pixels. Second, for the microscopic scenes to be fused, images from five different focal planes were obtained at the same positions and input to the trained m-SegNet to evaluate the focus levels of the pixels in the source images and to generate decision graphs to guide fusion. Third, on the basis of determining the focus area of each focal plane image, all the focal plane images are judged as clear overlap areas, and the no-reference image quality assessment using blur algorithm is used to find the clearest area. Finally, the fused image is realized through a final decision mapping. This proposed method was compared with five other advanced fusion methods in terms of subjective visual perception and objective evaluation indicators. The experimental results show that the proposed method is comparable to or better than some of the other advanced methods considered herein.

The main contributions of this work are as follows. First, this work proposes a Segnet model that uses the lightweight MobileNet model for decoding, which improves the speed of network operation under the premise of ensuring high accuracy. Second, we propose an end-to-end multi-focus image fusion method where it is no longer necessary to generate the decision graph and the Segnet model can be directly used for segmentation to guide fusion; this shows the great potential of the semantic segmentation method in multi-focus image fusion. Third, the focal loss function was used for in-depth supervision of the output at each level in the decoder, showing its superior performance for small sample sizes with class imbalances.

In general, the proposed multi-focus image fusion algorithm based on m-SegNet can reduce the error to the pixel level when classifying the focus and defocus regions of the source image, thus effectively avoiding the block effect and reducing the imperfect boundary.In addition, our algorithm classifies and segments images as a whole, and the output results can be directly used as decision map guiding fusion. It is end-to-end and highly parallel, so it is also better than traditional fusion methods in speed.However, there are a few shortcomings in the proposed method, which needs to be studied further. Including reducing operation or memory consumption to meet the requirements of real-time applications, transplanting the proposed method to many fields that require multi-focus image fusion, such as metallography and textiles.

Acknowledgment

The authors would like to thank the anonymous reviewers and editor for their helpful and valuable comments, which substantially improved the quality of the paper.The authors sincerely thank Prof. Xiaobin Xu from Hohai University for his meaningful guidance.

References

[1] P. J. Burt , and E. H. Adelson, “The laplacian pyramid as a compact image code,” *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, 2003.

[2] A. Toet, “Image fusion by a ratio of low-pass pyramid,” *Pattern Recognit. Lett.*, vol. 9, no. 4, pp. 245–253, 1989.

[3] L. Hui , B. S. Manjunath, and S. K. Mitra, “Multi-sensor image fusion using the wavelet transform,” *Graphical Models Image Process.*, vol. 57, no. 3, pp. 235–245, 2002.

[4] X. Xu *et al.*, “Echo signal extraction method of laser radar based on improved singular value decomposition and wavelet threshold denoising,” *Infrared Phys. Technol.*, vol. 92, pp. 327–335, 2018.

[5] M. N. Do and M. Vetterli, “The contourlet transform: An efficient directional multiresolution image representation,” *IEEE Trans. Image Process.*, vol. 14, no. 12, p.pp. 2091–2106, 2005.

[6] Q. Z. A and B.-.. G. b, “Multifocus image fusion using the nonsubsampled contourlet transform,” *Signal Process.*, vol. 89, no. 7, pp. 1334–1346, 2009.

[7] J. J. Lewis *et al.*, “Pixel- and region-based image fusion with complex wavelets,” *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, 2007.

[8] H. Qian *et al.*, “Multi-sensor image fusion with SCDPT transform,” in *Proc. 15th IEEE Int. Conf. Commun. Technol. (ICCT)*, 2013.

[9] N. Mitianoudis , and T. Stathaki, “Pixel-based and region-based image fusion schemes using ICA bases,” *Inf. Fusion*, vol. 8, no. 2, pp. 131–142, 2007.

[10] B. Yang , and S. Li, “Multifocus image fusion and restoration with sparse representation,” *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 884–892, 2010.

[11] J. Liang *et al.*, “Image fusion using higher order singular value decomposition,” *IEEE Trans Image Process*, vol. 21, no. 5, pp. 2898–2909, 2012.

[12] Y. J. A. B, and M. W. A, “Image fusion with morphological component analysis,” *Inf. Fusion*, vol. 18, no. 1, pp. 107–118, 2014.

[13] Z. Liu *et al.*, “A novel multi-focus image fusion approach based on image decomposition,” *Inf. Fusion*, vol. 35, pp. 102–116, 2017.

[14] Z. Jing *et al.*, “Evaluation of focus measures in multi-focus image fusion,” *Pattern Recognit. Lett.*, vol. 28, no. 4, 2007, pp. 493–500. https://doi.org/10.1016/j.patrec.2006.09.005.

[15] S. Li , X. Kang, and J. Hu, “Image fusion with guided filtering,” *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, 2013.

[16] L. Yu, , S. Liu,and, and Z. Wang.“Multi-focus image fusion with dense SIFT,” *Inf. Fusion*, vol. 23 no. 2015:139–155.

[17] Z. Zhou, S. Li, B. Wang Multi-scale weighted gradient-based fusion for multi-focus images[J]. *Inf. Fusion*, 2014, vol. 20, pp. 60–72.

[18] V. Aslantas , and R. Kurban, “Fusion of multi-focus images using differential evolution algorithm,” *Expert Syst. with Appl.*, vol. 37, no. 12, pp. 8861–8870, 2010.

[19] Y. Guo , Y. Zhang, and D. Zhou, “A block-based fusion method with SSIM criterion for multi-focus images,” in *Proc. Int. Symp. Photon. Optoelectron. Int. Soc. for Opt. Photon.*, 2014.

[20] H. Hariharan , A. Koschan, and M. A. Abidi, “Multifocus image fusion by establishing focal connectivity,” in *Proc. IEEE Int. Conf. Image Process.*, 2007.

[21] I. De , and B. Chanda, “Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure,” *Inf. Fusion*, vol. 14, no. 2, pp. 136–146, 2013.

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

[23] T. A. Runkler , M. Sturm, and H. Hellendoorn, “Model based sensor fusion with fuzzy clustering,” in *Proc. IEEE World Congr. IEEE Int. Conf. Fuzzy Syst.*, 1998.

[24] Y. Liu *et al.*, “Multi-focus image fusion with a deep convolutional neural network,” *Inf. Fusion*, vol. 36, pp. 191–207, 2017.

[25] H. Tang *et al.*, “Pixel convolutional neural network for multi-focus image fusion,” *Inf. Sci.*, 2018, S0020025517311647.

[26] X. Guo *et al.*, “Fully convolutional network-based multifocus image fusion,” *Neural Comput.*, vol. 30, no. 7, pp. 1–26, 2018.

[27] H. Li *et al.*, “Multi-Focus image fusion using U-Shaped networks with a hybrid objective,” *IEEE Sensors J.*, vol. 19, no. 21, pp. 9755–9765, 2019.