IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 17, 2020, accepted May 4, 2020, date of publication May 11, 2020, date of current version May 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993625

Convolutional Neural Network for Extracting 3D Point Clouds of Fibrous Web From Multi-Focus Images

JUE HOU[1,2,3], WENBIN OUYANG[3], BUGAO XU[3], AND RONGWU WANG[1,2]

[1]Key Laboratory of Textile Science and Technology, Ministry of Education, Donghua University, Shanghai 201620, China
[2]College of Textiles, Donghua University, Shanghai 201620, China
[3]Department of Merchandising and Digital Retailing, University of North Texas, Denton, TX 76203, USA

Corresponding author: Rongwu Wang (wrw@dhu.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2232020G-01, and in part by the National Natural Science Foundation of China Grant under Grant 61771123.

ABSTRACT This paper presents a new method for extracting 3D point clouds from multi-focus images of a fibrous web acquired on an optical microscope to analyze microscopic structures of a fibrous web. The algorithm consists of two major parts: (1) utilizing a convolutional neural network (CNN) to extract in-focus objects from multi-focus images, and (2) a depth identification module (DIM) which is a frequency domain-based model used to identify the depths of object points. The network, namely the multi-focus image deblurring network (MIDN), was designed by introducing gradient features into the network to deblur images and generate the ranges of focal depths of object points. Based on the results of MIDN, DIM was constructed to calculates the focal plane depth for each point. The experiments show that the combination of MIDN and DIM provides a practical way to generate complete, accurate 3D structures of nonwoven.

INDEX TERMS Image reconstruction, optical microscopy, artificial neural networks, machine vision.

Three dimensional (3D) reconstruction is an important technique that can be used for multi-focus microscopic images analysis. Confocal microscopes have been widely used for 3D reconstruction of the microscopic structures [1], because it can acquire depth information directly and filter out background noise. However, the point-by-point imaging principle of a confocal microscope leads to low-speed scanning and possible damage on samples [2], and its high price also limits widespread applications [3]. Hence, it is valuable to retrieve 3D information from 2D images which are acquired on regular light microscopes. The optical coherence tomography is the most common way that uses the sequential images of an object captured on various focal planes/depths to reconstruct the 3D surface image [4]–[6]. Quantitative phase imaging (QPI) can developed to deal with transparent and translucent objects in optical microscopy [7], [8]. Based on QPI, LED matrix illumination was utilized to capture images under different beam angles by controlling the LED arrays [9]–[11]. However, the aforementioned methods require specific modifications on microscopes.

To retrieve 3D information from sequential images, a point spread function (PSF) and statistical characteristics were also used to restore non-degraded images from the captured images. Classical image restoring algorithms, such as Wiener filter [12] and Kalman filter [13], were proposed to eliminate the degradation function through linear iterations. However, if the input signals are interfered by random noise, the linear iteration-based algorithms cannot acquire stable results. Some approaches utilized the model of illumination patterns to create sophisticated mathematical representations [14], [15], but they relied on high-precision priori knowledge [16]. Lately, computational optical sectioning microscopy (COSM) became a popular 3D microscopy method because of its high accuracy [17]. In COSM, an image sequence is collected as a series of microscopic images that are focused at different planes on the specimen [18], and a classical method, the nearest neighbor deconvolution (NND), is used to remove the blurriness of the current image. The core idea of NND is that the current image is influenced by its adjacent images, and its blurry information can be eliminated by subtracting the product of two adjacent images from the interlayer PSF. In addition, the frequency components of the specimen can be obtained by using frequency components of images to divide

The associate editor coordinating the review of this manuscript and approving it for publication was Yunjie Yang.

the Fourier transform of the PSF. As implemented in the Jassan-Van Cittert method [18] and the maximum likelihood estimation method [19], [20], frequency-based deconvolution is another viable approach used in COSM [21]. However, the actual PSF is not invariant in the 3D space, and most of these methods assume that the PSF is a variant model. In addition, using the estimation of PSF to recover the 3D information of image sequence is not suitable for applications which need high-speed calculations [21].

Deep learning has been widely used in microscopic image segmentation and restoration [22]. Rivenson *et al.* [23] elaborated a deep learning model for improving the resolution of optical microscopic images without hardware adaptation. Ronneberger *et al.* [24] proposed a pioneering model of deep learning in microscopy, called U-Net, and took full advantages of feature maps in the contracting path to increase the accuracy of pixels localization. Weigert *et al.* [25] explored a U-net based network to eliminate the influence of noise and the need for the PSF, and performed unsupervised and end-to-end training through the peak signal to noise ratio loss function. Compared with the traditional deconvolution method, i.e., Richardson-Lucy deconvolution algorithm [26], the proposed CNN model achieved higher quality restoration with faster speed. The CARE network, which is another U-net based network [27], utilized the synthetic ground-truth and fluorescence microscopic images as the training dataset to optimize the model to raise the efficiency of the fluorescence microscopic images restoration. Other CNN models, such as Residual Network [28], [29] and Generative Adversarial Network [30], were also reported for enhancing the quality of microscopic images. For the application of finding focal planes from image sequence, Li *et al.* [31] developed a three-layer network to generate the clearest layer map from multi-focus images, and Conchello and Lichtman [18] designed a two-input network to generate the probability map of fusion. However, these methods have not been used for the 3D reconstruction of an examined sample whose thickness is far beyond the depth of view of a microscope.

Nonwoven materials have a wide range of applications, particularly in filtering devices and medical masks. The filtering performance of a nonwoven depends on its 3D structure and important parameters such as porosity and filling ratio of fibers. Because a nonwoven is constituted by massive crossing fibers and its thickness significantly surpasses the depth of view of a light microscope, retrieving the 3D structure of a nonwoven from its multi-focus images remains challenging. Normally, this transformation requires to separate in-focus pixels i.e., object points in each image from the background and to determine the best focal plane for each object point. Park *et al.* [44] proposed a patch-level CNN model to extract high-dimensional features from hand-crafted features, and used another CNN model to localize the in-focus regions. However, the experiments showed that this method cannot distinguish the low-contrast focus regions. Zhao *et al.* [43] designed a multi-stream network (BTBNet) to detect in-focus regions. The BTBNet combines multiple convolutional layers to compose streams, and utilizes the streams to extract features in different scales. At the end of BTBNet, the features are input into a decision network. Although the BTBNet can detect the in-focus regions accurately, the sophisticated network structure has high computational costs.

In this paper, we present a two-step approach to reconstruct 3D image of fibrous webs by utilizing sequential microscopic images captured at different focal planes. In the first step, a CNN model, which is named as the multi-focus image deblurring network (MIDN), is used to extract in-focus/sharp pixels from optical sections. The MIDN can extracts features from the optical sections and generates a feature map by the encoder-decoder structure. To improve the performance of the network, the gradient features are introduced into the network, and generate a probability map. A modified Conditional Random Field is used to connect the feature map and the probability map and utilized convolutional layers to generate the map of in-focus objects. In the second step, a depth identification module (DIM) is utilized to select an optimal depth for each objects points from the results of MIDN with the frequency domain information. The DIM, inspired by the NND algorithm, focuses on the power spectrum changes between adjective layers and uses Gaussian kernels to smooth the distribution of power spectrum changes. Nonwovens are selected as examples for acquiring multi-focus images on an optical microscope and used for training and validating the proposed 3D reconstruction algorithm. The major tasks performed in the research include: (1) the introduction of the activation path, which is derived from Conditional Random Field, into the CNN; (2) the creation of a microscopic multi-focus image dataset of nonwovens; and (3) the design of a depth identification module (DIM) for the optimal focal plane determination in a high speed.

I. MIDN FOR IMAGE IN-FOCUS POINTS EXTRACTION

A. ARCHITECTURE OF CONVOLUTIONAL NEURAL NETWORK

The well-known network U-net is composed of an encoder path and a decoder path, and the strategy of U-net utilizes rich features to generate higher accuracy outputs. Hence, we take advantages of U-net and design a light weight network, called Multi-focus Image Deblurring Network (MIDN), as shown in Figure 1. In the MIDN, the captured image $f(x, y, z)$ is fed into the network, and the network generates in-focus point candidates on current image $\hat{g}(x, y, z)$. As the ground truth, the sets of in-focus pixels $g(x, y, z)$ are used to supervise the optimization of MIDN. The difference between output of network and the ground truth, $|\hat{g}(x, y, z) - g(x, y, z)|$, is the objective function of MIDN, and the function approaches to the minimum in the training. The architecture of MIDN is shown in Figure 1 and the specific setup of the MIDN is listed in Table 1.

The MIDN consists of feature extracting path, activation path and output path. Although the feature extracting path inherits the similar structure of U-net, it needs less float

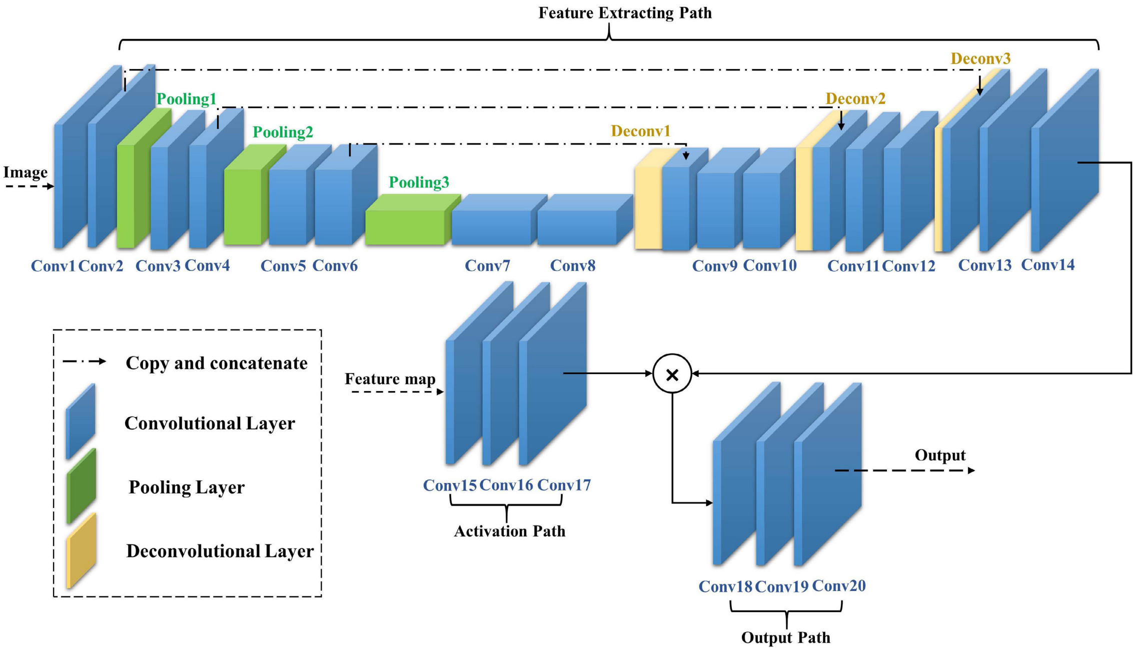


FIGURE 1. The architecture of MIDN. Each box indicates a layer. The original images are fed into the Conv1 layer and the gradient maps are fed into the Conv15 layer. The dash lines represent the feature maps of convlutional layers are copied and concatenated with the feature maps of deconvolutional layers. The outputs of Conv14 layer are activated by the outputs of Conv17 layer, then fed into the Conv18 layer.

point operations (FLOPs) to generate results. The feature extracting path involves 14 convolutional layers, 3 pooling layers and 3 deconvolutional layers. The first 8 convolutional layers and 3 pooling layers are adopted to extract the features from inputs, the last 6 convolutional layers and 3 deconvolutional layers are utilized to build the high quality outputs. To increase the output resolution and accuracy, the feature maps are copied and combined with the feature maps of the deconvolutional layers as shown in Figure 1. Besides the feature extracting path, a branch of three convolutional layers is added to the network. The branch introduces gray gradients of pixels before generating a feature map. The generated feature maps give coefficients to all the pixels of the feature extracting path results, and thus the branch is called the activation path in this paper. At the end of the network, the output path is designed to generate the results according to the products of the activation path outputs and the feature extracting path outputs. Compared to the U-Net structure, the float point operations (FLOPs) of MIDN are about 1.1×10^{11} and it obviously less than the original U-Net whose FLOPs are about 1.7×10^{11}.

B. ACTIVATION PATH OF MIDN

Generally, different kinds of objects have various features which can be utilized as clues to classify objects. However, in in-focus object detection, clear objects (on the focal plane) and blurry objects (out of the focal plane) often have similar characteristics such as the topological structures and the colors. To distinguish objects in a low-contrast region, intensity gradients, a degree of clearness, can be introduced into the network to increase the accuracy of the output. The gradient magnitude, $|\nabla f|$, at pixel (x, y) is defined as follows:

$$|\nabla f| = \sqrt{(\frac{\partial f}{\partial x})^2 + (\frac{\partial f}{\partial y})^2} \qquad (1)$$

where the $|\nabla f|$ indicates the gray value distribution over an image, the $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ indicate the partial derivatives of the f. As reported in [32], [33], Conditional Random Field (CRF) is an effective method to feed additional features into a CNN and exampled by DeepLab [34] such as a recurrent neural network (RNN) in [33]. In the CRF as RNN model, the regular CNN outputs are regarded as priori probability and used to calculate the energy of label assignments, E(x), with pairwise potential. E(x) can be calculated as the [33] reported:

$$E(x) = \sum_{i} \varphi_u(x_i) + \sum_{i<j} \varphi_p(x_i, x_j) \qquad (2)$$

where φ_u denotes the unary potential which measures the probability of assigning label x_i to pixel i, and $\varphi_p(x_i, x_j)$ is the pairwise potential which measures the cost of assigning labels x_i and x_j to pixels i and j respectively. In Eq2, the pairwise potential supplies a penalty mechanism to the label assignment, in which the energy decreases when pixels get inappropriate labels. Generally, the pairwise potential of fully connected CRF is calculated as an RNN or a post-processing model. Incorporating such a fully connected CRF into a CNN is rather time-consuming. In this paper, the concerned regions center on the edges of objects, and we focus on building the

TABLE 1. The setup details of MIDN.

feature extracting path					
layer	Conv1	Conv2	Pooling1	Conv3	Conv4
kernel	3×3×1×64	3×3×64×64	2×2	3×3×64×128	3×3×128×128
stride	1	1	2	1	1
activation method	relu	relu	-	relu	relu
layer	Pooling2	Conv5	Conv6	Pooling3	Conv7
kernel	2×2	3×3×128×256	3×3×256×256	2×2	3×3×256×512
stride	2	1	1	2	1
activation method	-	relu	relu	-	relu
layer	Conv8	Deconv1	Conv9	Conv10	Deconv2
kernel	3×3×512×256	4×4×256×256	3×3×512×256	3×3×256×128	4×4×128×128
stride	1	2	1	1	2
activation method	relu	-	relu	relu	-
layer	Conv11	Conv12	Deconv3	Conv13	Conv14
kernel	3×3×256×128	3×3×128×64	4×4×64×64	3×3×128×64	1×1×64×1
stride	1	1	2	1	1
activation method	relu	relu	-	relu	relu
activation path					
layer			Conv15	Conv16	Conv17
kernel			3×3×1×64	3×3×64×64	1×1×64×1
stride			1	1	1
activation method			sigmoid	relu	relu
output path					
layer			Conv18	Conv19	Conv20
kernel			3×3×1×256	3×3×256×64	1×1×64×1
stride			1	1	1
activation method			relu	relu	sigmoid

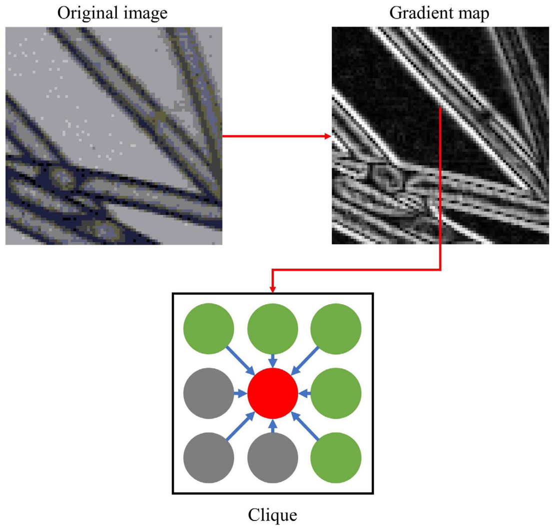


FIGURE 2. Illustration of Clique. According to the gradient map, the green pixels have higher gradient values and the gray pixels have lower gradient values, and all the neighbors give activation coefficients to the center pixel.

relationship in edges rather than the whole image. So we design an activation path to calculate the local potentials in a novel method. The pairwise potential in Eq2 allows different features, such as gradient, color, to be brought into the energy calculation, and different pixels to be paired up. To connect a pixel with its neighbors, the pixel and its eight neighbors are treated as a clique as Figure2.

In Figure2, the center pixel is influenced by its surrounding neighbors, and the center pixel receives a penalty coefficient c defined as:

$$c = \sum_{i=1}^{i=8} \sigma(\omega_i \times f_i) \tag{3}$$

where f_i and ω_i denote that the gradient intensity and the coefficient of neighbor pixel i, and $\sigma(\ldots)$ is the symbol of the sigmoid function. Different from the theorem of [36], the pairwise term of Eq.2 is converted to the clique potential as Eq.3, and it describes that the compatibility between focus degree and gradient value. Since the gradient values of edge points are non zero, c is also greater than zero, the neighbors intend to encourage rather than penalize the center pixel. Hence, we propose to change the form of Eq2 to the following form:

$$Q(x) = \varphi_u(x) + c_x = ln(e^{\varphi_u(x)} \times e^{\epsilon c_x}) \tag{4}$$

where $Q(x)$ is the energy of labeling pixel x, $\varphi_u(x)$ indicates the unary potential and ϵc_x denotes the normalized penalty coefficient of pixel x. Compared with the Eq2, Eq4 takes the natural log and converts the clique potential to the coefficient of unary term. In Eq4, ϵ is a normalized coefficient of c_x, which is used to allow ϵc_x to be negative numbers that impose penalty to the center pixel. The Eq3 is easily considered as the convolution operation, so we design the Compatibility Layer (Conv15) to calculate the penalty coefficient c. The Compatibility Layer includes 64 kernels, and the kernels are initialized with constants. The kernels can be optimized during the network training, so the Compatibility Layer can give appropriate coefficient ω to the center pixel under different situations. Besides the Compatibility Layer, the activation path also involves Gaussian Layer (Conv16) and Fusion Layer (Conv17). The traditional CRF models utilize Gaussian kernels to eliminate the isolated points in an image. In our model, the Gaussian kernels are considered as a Gaussian Layer whose kernels are are initialized with Gaussian distribution whose mean value and variance are 0 and 1 respectively, and the Gaussian kernel also be optimized by the training. The Gaussian Layer not only smooths the feature maps, but also extracts the high dimensional features. At the end of the activation path, 1 × 1 kernels of Fusion Layer are used to weight different features and calculate the c of each pixel.

Figure 3 shows a few multi-focus images of a nonwoven sample and their corresponding maps of in-focus points which are generated by MIDN. In the outputs of MIDN, each point value indicates the probability of the pixel in the focal plane. A multi-focus image contains both focused and defocused fiber pixels captured at one focal plane or layer, and the layer number indicates the depth position, z, in the imaging system. After MIDN, only in-focus pixels of fibers are filtered out and selected as candidate points for voxels of fibers in the 3D space. Some of these pixels may remain focused in several consecutive multi-focus images and will appear in the outputs as well.

II. DEPTH DEFINITION BY FREQUENCY MODEL

In the maps of in-focus points, each candidate point that appears in multiple MIDN results is associated with different depths z. For the same (x_0, y_0), we need to identify an optimal z_0 to form a voxel, (x_0, y_0, z_0), that builds the 3D structure of the nonwoven. Since the intensity of a pixel in an image is always associated with its neighboring pixels, it can be regarded as a spread-out region/patch centered at the current pixel. This intensity distribution of this patch can be approximated by a two-dimensional Gaussian function, $g(r, \sigma)$, with a radius of r and a spatial constant σ. Denote the intensity distributions of the same circular region (r) or patch in two adjacent multi-focus images as $f_i(r)$ and $f_{i+1}(r)$. The patches include the corresponding points on the original sample as $f_0(r)$. As an invariant linear system, the imaging process of adjacent images in sequential can be represented as:

$$\frac{f_i(r)}{f_{i+1}(r)} = \frac{f_0(r) \otimes g(r, \sigma_i) + o_i(r)}{f_0(r) \otimes g(r, \sigma_{i+1}) + o_{i+1}(r)} \tag{5}$$

Here, the function $g(\ldots)$ indicates the PSF and the $o(r)$ is the defocus term. According to the Fourier transform, f_i and f_{i+1} can be described as follows:

$$\frac{\mathcal{F}_i(\lambda)}{\mathcal{F}_{i+1}(\lambda)} = \frac{\mathcal{F}_0(\lambda) \times \mathcal{G}(r, \sigma_i) + \mathcal{O}_i(r)}{\mathcal{F}_0(\lambda) \times \mathcal{G}(r, \sigma_{i+1}) + \mathcal{O}_{i+1}(r)} \tag{6}$$

where F and f, G and g, O and o are the Fourier pairs, and σ_i equals to $\frac{1}{\sqrt{2\pi\sigma_i}}$. Subscripts i and $i+1$ refer to two consecutive focal planes. According to the [37], the images of in-focus objects tend to have more high frequency components and higher power in the frequency domain. The patches in this paper are around the candidate points which are selected by MIDN, and most of the pixels in the patches are near the focal planes. Hence, the low frequency components in Eq6 occupy a small percentage of patch areas. To simplify computation, we assume the low frequency term, O, as a constant, and rewrite the Eq6 as:

$$C_i = ln\mathcal{F}_i - ln\mathcal{F}_{i+1} = 2\pi^2(\sigma_i^2 - \sigma_{i+1}^2) \tag{7}$$

The difference between lnF_i and lnF_{i+1} is C_i. Different from the in-focus points, the out-focus points change slowly and the low frequency components between connected layers are eliminated approximately through Eq7.

Assume that the depth range of candidate point $P(x_0, y_0)$ is $[z_{min}, z_{max}]$ which is defined by MIDN. As the [38] reported, the frequency domain, gradients and variance are generally used to measure the clearness degree of $P(x_0, y_0)$. Among different measurements, the variance is the most efficient method to identify the focal depth of pixel. However, some overlapped objects have more than one focal plane, and the variance also cannot reflect the multiple focal planes as shown in Figure 4. Compared to the variance, the distribution of C_i has an obvious valley in the scope of $[z_{min}, z_{max}]$, and the C_i is more sensitive to the change of focus.

During the procedure of focus, a pixel can remain in-focus in multiple layers when the corresponding object point near the focus plane, and the patch of this point becomes blur as the sample away from the focal plane. So the difference in power spectrum C_i keeps low level near focal plane. Since the accurate ground truth for $ln\mathcal{F}_i$ is not available, the optimization of C_i can only be solved by a non-supervised method. Thus, we build a model (Figure 5) which is created to identify the optimized depth of object points. The MIDN gives the z scopes of candidate pixels, and the patches (in the scopes) whose center at the candidate pixels are fed into the cells. Each cell involves a Gaussian kernel and Fourier transform. The Gaussian kernels (mean value is 0 and variance is 1) are used to eliminate the noise of patches, and the Fourier transform is used to convert the patch into the frequency domain and calculate the integral of the power spectrum. After that, we use two one-dimensional convolutional layers whose kernels are set to be 5 × 1 × 1 to smooth the distribution of the integrals of the power spectrum. The illustration of this algorithm and the smoothing results are shown in Figure 5.

The Figure 5 shows that the distribution of integrals is converted as a smooth curve, and the layer of minimum points between two peaks of curve is the optimized depth of the current point.

III. DATASET FOR DETECTING IN-FOCUS OBJECTS IN MICROSCOPIC IMAGES

Most of the approaches of clear regions are evaluated on the public blur detection dataset [45], which involves over 1000 natural scene images and provides human annotations for blur region detection. The labels of the dataset mark the whole object regions. But in this application, we only need to extract fine structures of the objects such as fiber contours from microscopic images. No suitable dataset is available for detecting in-focus regions in a microscopic image.

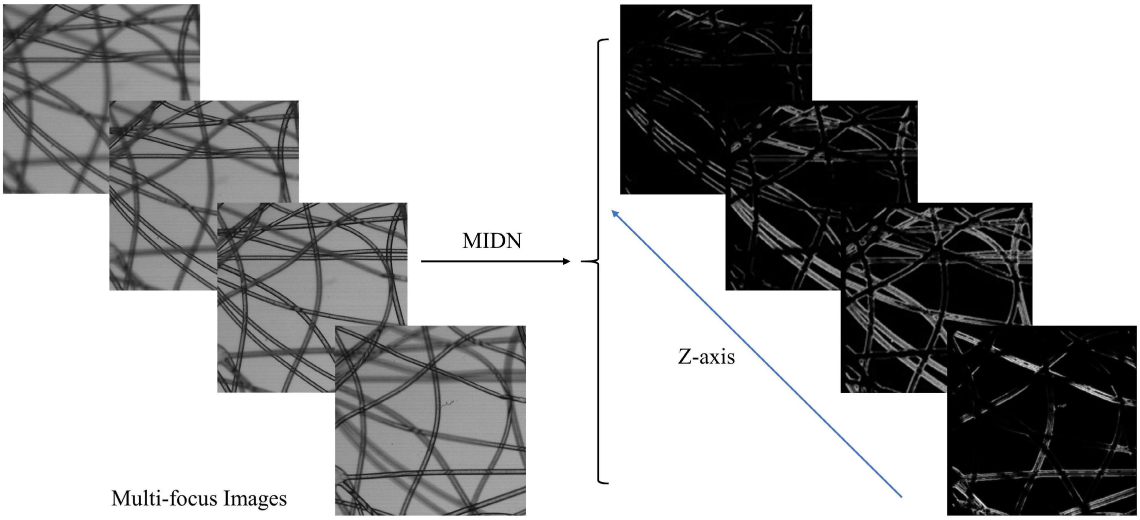
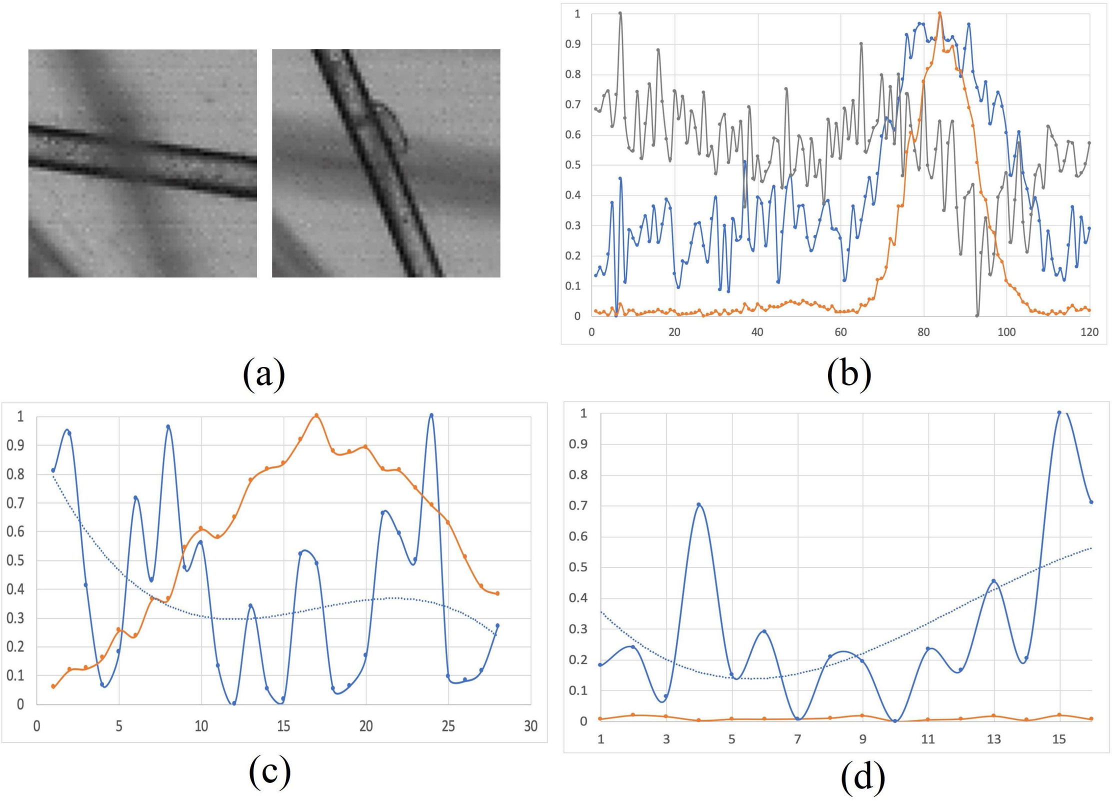


FIGURE 3. The MIDN results of multi-focus images. The left image sequence involves the microscopic images, and the right image sequence involves the maps of in-focus points.



(a)

(b)

(c)

(d)

FIGURE 4. Illustration of different measurements. (a) is the objects overlap areas. (b) is the illustration of different clear degree measurements which are supplied by [38], where the blue line is frequency domain change, the gray line is the gradients change, the orange line is the variance change. (c) is the variance and *C* changes of first fiber in (a), where the z scope of this fiber is defined by MIDN. (d) is the variance and *C* changes of second fiber in (a), where the z scope of this fiber is defined by MIDN. The blue lines and the orange lines in (c) and (d) are *C* distribution and variance distribution respectively.

We captured a set of nonwoven microscopic images using a motorized microscope equipped with a JAI BM-141GE camera and a UPLSAPO 10X object lens to create a new dataset. Since the thickness of nonwoven sample was larger than the microscopic depth of view, 100 layer/sections were captured at each (x,y) position (see Figure 6). The size of the

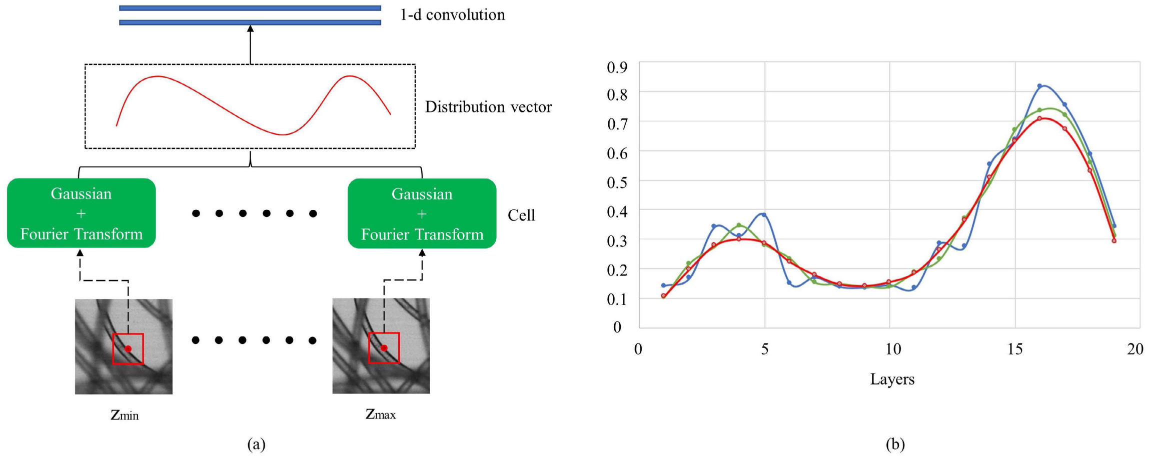


FIGURE 5. Illustration of depth definition model. (a) is the model structure. The z scope is defined by the results of MIDN. The green boxes indicate the cell, and each cell involves Gaussian operation and Fourier transform. The cells output the integral of the power spectrum in frequency domain. The cell outputs compose a vector which is illustrated as the dash line box. We input the vector into the 1-d convolutional layers, and output the smooth curve as (b). (b) is the C distribution curves, where the blue curve indicates the original C distribution.The green curve and red curve are the curves which are smoothed by one convolutional kernel and two convolutional kernel respectively. Comparing with one 1-d convolutional layer, using two 1-d convolutional layers can generate a smooth curve which can be detected the local minimum points easily.

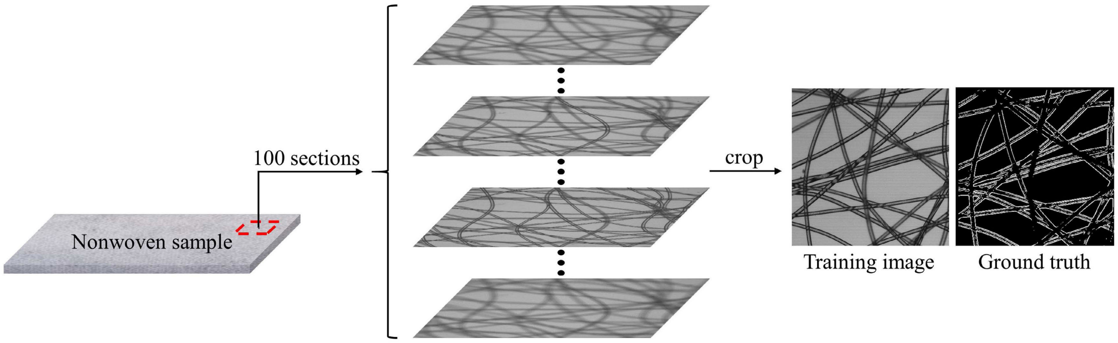


FIGURE 6. The illustration of nonwoven sample images capturing. The red box indicates the acquisition point of sections. In one acquisition point, 100 sections were captured by optical microscope. Hence, the captured sections can cover the thickness of nonwoven samples. We collected more than 10000 raw images, and selected 6400 images from raw images to build the dataset.

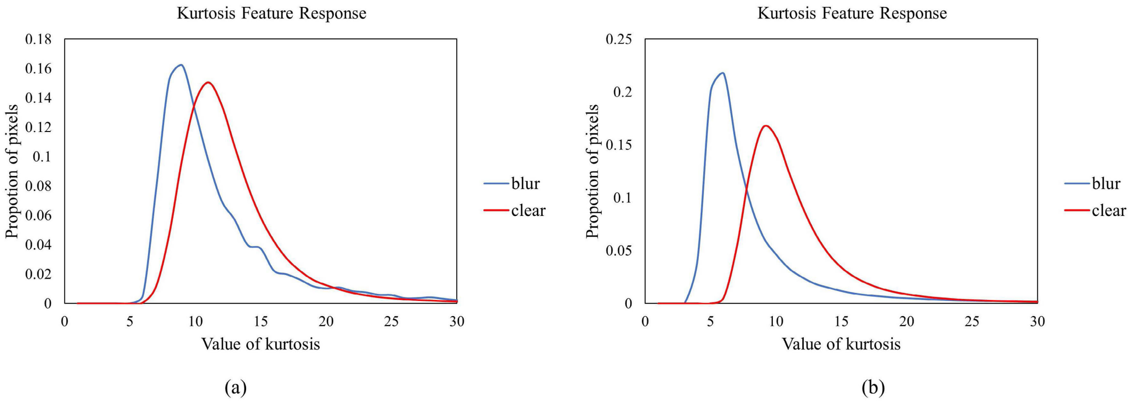


FIGURE 7. The curves of kurtosis responses. The blue curves indicate the kurtosis distributions of blur pixels and the red curves indicate the kurtosis distributions of clear pixels. (a) illustrates the kurtosis distribution of clear pixels and blur pixels which are marked by one observer. (b) illustrates the kurtosis distribution of clear pixels and blur pixels which are marked by two observers. Compare with the curves of (a), the curves of (b) have smaller overlapped regions.

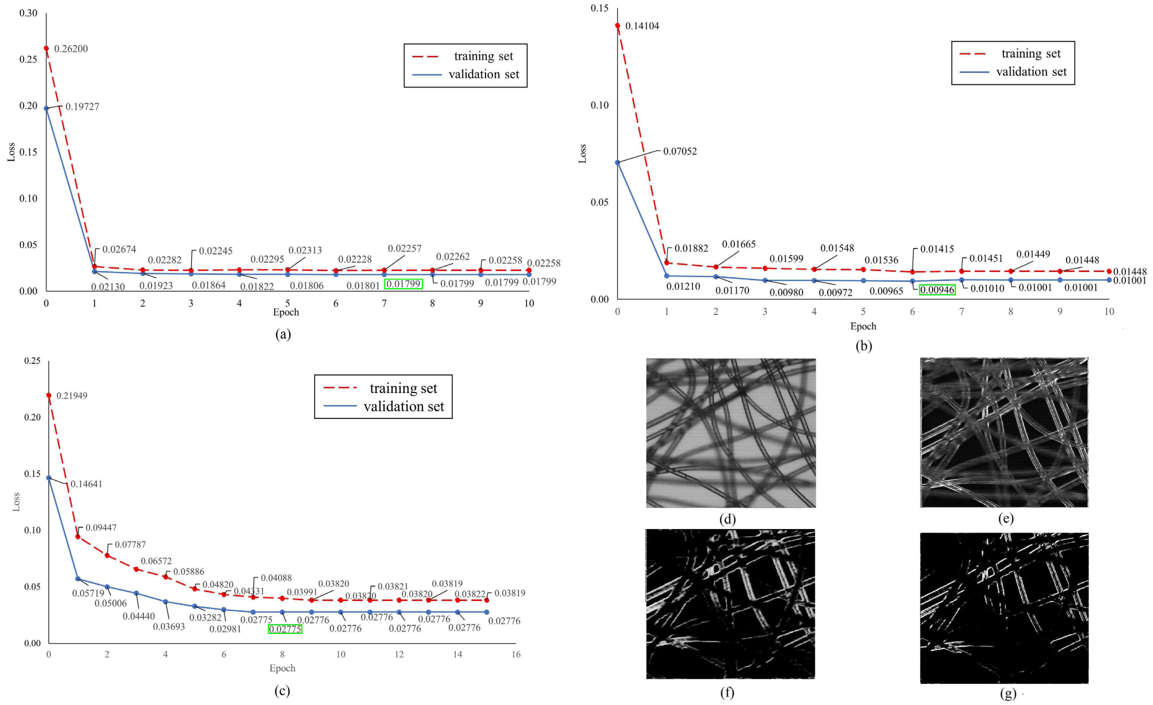


FIGURE 8. The learning curves of model training. The green boxes indicates the minimum losses of validation set. (a) shows the learning curves of training the feature extracting path in two-step strategy. (b) shows learning curves of the model after unlocked parameters. (c) shows that the learning curves of training the network as a whole. (d) is the sample of testing set. (e) is the sample of network output (8 epochs) under the strategy of training the network as a whole. (f) is the sample of feature extracting path output, where training the model 7 epochs and froze other parts. (g) is the sample of the final result after 13 epochs training (two-step strategy).

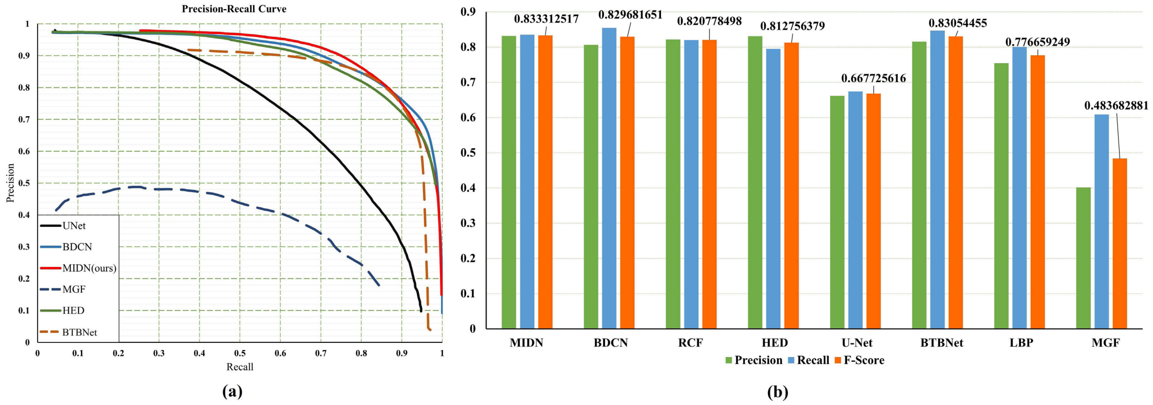


FIGURE 9. The testing results of models. (a) includes the Precision/Recall curves of different networks. (b) represents the comparisons of Precision, Recall and the F-score between different state-of-the-art models. The results show that our model achieves the highest accuracy between models.

captured images is 1024 × 768 pixels. An area of 320 × 320 pixels was randomly cropped out from each image for the training. Our dataset contains 7000 images, 5000 images were used as the training set, 1000 images as the validation set and the rest 1000 images as the testing set. The ground truth images of fiber edges in the dataset were labeled by two human observers manually. According to the reference [45], the distribution of kurtosis values is an effective measurement for clearness. The Figure 7 shows that the kurtosis distribution of in-focus pixels (marked by "1") and kurtosis distribution of blur pixels (marked by "0") have small overlap part between each other. In contrast, the in-focus pixels and blur pixels on the ground truths which are labeled by one observer have similar kurtosis distributions. Although more observers can lead to more labeling reliability for the ground truths, the human annotations are expensive and time consuming. The strategy of labeling by two observers is the best trade-off between accuracy and time cost. According to the ground truths, the proposed dataset is divided into two parts: in-focus points (positive samples) and out-focus points (negative

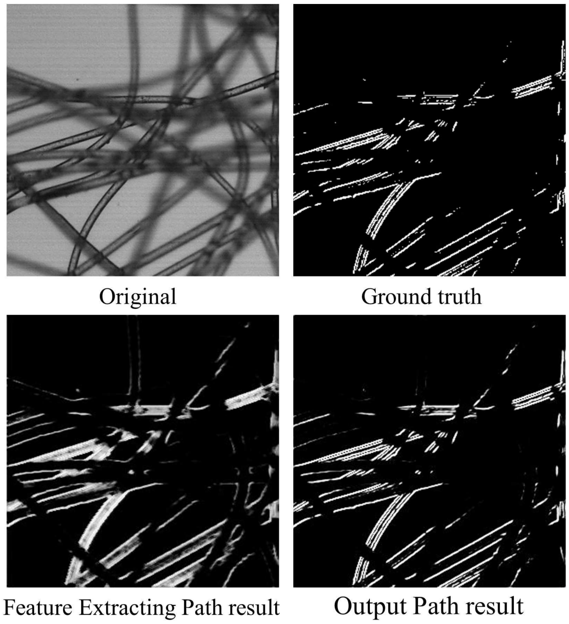


FIGURE 10. The comparison of feature extracting path result and output path result. Compare with the feature extracting path, the output path can extract details of the objects and generate clear result.

samples). Compared to the large number of negative samples, the positive samples only occupy a small proportion of all the samples. To overcome the imbalance losses between positive samples and negative samples, we applied the class-balanced cross-entropy loss function as follows, [40]

$$l(X, W) = -\beta \times \sum_{i \in positive} logPr(y_i = 1|X_i, W) - (1-\beta) \times \sum_{i \in negative} logPr(y_i = 0|X_i, W) \quad (8)$$

where $\beta = N_{negative}/(N_{positive} + N_{negative})$. The $N_{negative}$ and $N_{positive}$ indicate the points number of out-focus pixels and in-focus pixels respectively. The CNN output value which is activated by sigmoid function (Pr) and the in-focus probability at pixel i are represented by X_i and y_i respectively. The $l(X, W)$ denotes the loss value, in which the W indicates all the parameters of the network.

IV. EXPERIMENTS

A. MIDN FOR IMAGE IN-FOCUS POINTS EXTRACTION

The MIDN was implemented on a deep learning framework – Pytorch. The network training was performed on a single graphics card – GTX 2080TI. The convolutional kernels and deconvolutional kernels of the network were initialized by Xavier algorithm [39], and the max pooling was adapted in the experiments. During the training, the original images and the gradient maps were fed into the feature extracting path and activation path respectively. Of the parameters in the network optimizer, the stochastic gradient decent algorithm was used to optimize the network and the learning rate was fixed as 1×10^{-5}. In addition, the momentum of learning was set to 0.99. We rotated the images to 4 different angles, and augmented dataset,which includes 20000 training images, was 4 larger than the unaugmented set. The training strategy of the MIDN was divided into two steps: the feature extracting path training and the whole network training. The Figure 8 shows that the learning curves of training the network in different strategies. The experiments of training the MIDN as a whole show that the loss values decreased to the local minimum points after 8 epochs training, and the loss of the validation set decreased to 0.02775. The loss values could not decrease in the following training, and the images show that the results still involved a large number of out-focus points. Hence, we adopted a two-step training strategy. At the beginning of training, we utilized the ground truths to optimize the feature extracting path, and froze the parameters of other parts of MIDN. Here, the outputs of feature extracting path were activated by sigmoid function, and calculated the losses with ground truths by the class-balanced cross-entropy loss function. The loss values of the feature extracting path decrease to about 0.0179 after 7 epochs, and the changes of loss values tended to stability. Then, we unlocked all the parameters of network and continued to train the whole network. The training was terminated after 10 more epochs. It is very important that the final model is selected when the least validation loss is reached, so we selected the model which was trained by 6 epochs as shown in Figure 8(b). Compared to training the network as a whole, the two-step strategy offered lower training losses (0.01415) and validation losses (0.00946), and the samples showed that the outputs reserved higher accurate results. In this paper, the standard measurements, precision, recall and F-Score. In this paper, the standard measurements: precision, recall and F-Score ($\frac{2 \times Precision \times Recall}{Precision + Recall}$) of the models were evaluated on our datasets. In addition, we also added the mean absolute error(MAE) which describes the pixel-wise differences between ground truth and the results. The MAE is calculated as:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |G(x, y) - M_{final}(x, y)| \quad (9)$$

where W and H represent the width and height of the images, and x and y indicate the coordinates of the pixels respectively. The smaller MAE, the higher the accuracy. To compare the performance of different algorithms, we used three state-of-the-art in-focus region detection models: BTBNet [43], LBP [47] and MGF [46]. Because most of the objects are fiber edges in the dataset, the state-of-the-art edge detection models, such as BDCN [42], RCF [41], HED [40] and U-Net [24], were also selected to compare with the MIDN.

The precision and recall results are shown on the Figure 9. The results show that the F-score of our proposed model MIDN reaches 0.833, which is the highest when compared with the in-focus region detection model BTB-Net (F-score = 0.830) and the edge detection model BDCN (F-score = 0.829). We also observed a phenomenon in which the precision/recall curve of MIDN is not as long as the

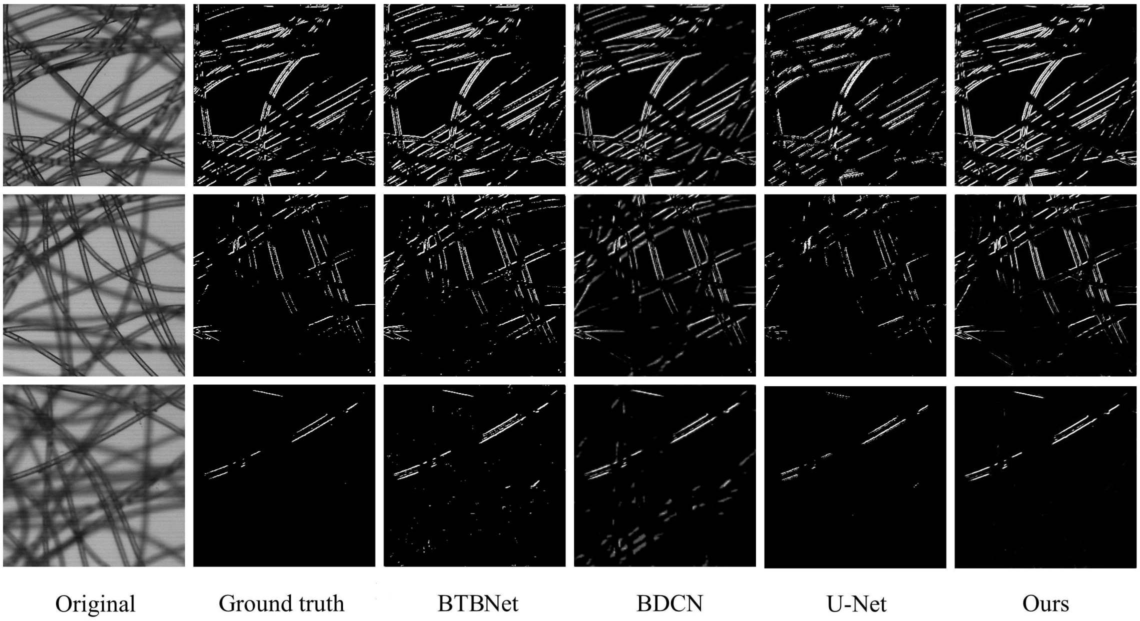


FIGURE 11. The comparison of test results.

TABLE 2. The MAE results and processing single image time.

Model	MIDN(ours)	BDCN	RCF	HED	U-Net	BTBNet	LBP	MGF
MAE	21.501	21.292	21.747	22.038	28.925	21.722	24.143	35.634
Time(units:s)	0.094	0.158	0.102	0.095	0.124	0.097	0.072	-

other curves. This is because the activation path supplies encouragements to the confident points and eliminates the false points by the penalties. Although the U-Net involves more FLOPs than the MIDN and the feature extracting path is inherited by the U-Net, the MIDN increases the F-score by 24.8% from the original U-Net. Figure 10 shows that the results of the feature extraction path result and the output path. The feature extraction path involves most of the true points, but the topological structures of objects are coarse. In contrast, the output path has higher accuracy, and most of the false points are filtered. Thereby, introducing effective features by the activation path can effectively improve the performances of the deep learning model.

The Table 2 reports the MAE values and the speeds. The BDCN has the lowest MAE among all the evaluated models. Although our model (MIDN) has a slight higher MAE than the BDCN, the MIDN takes 40.5% less time to process a single image than the BDCN. The average speed of processing one image by the MIDN is 0.094 second and it is fastest among all the deep learning models. The LBP is a non-deep learning method. Although the LBP has a shorter process time per image (0.072s) than the MIDN (0.094s), its MAE value (24.143) is much higher than that of the MIDN (21.501) and F-score of LBP is much lower than our model. Figure11 provides the visual comparisons of several deep learning models with the ground truths. The MIDN demonstrates the closest results to the ground truths among the evaluated models.

B. DIM FOR DEPTH IDENTIFICATION

After the training of the MIDN, the testing dataset and training dataset are converted to the two sets of candidate points maps. Here, the outputs of the MIDN supply the candidate points and their coordinates (x, y, z), in which z refers to the layer index, which can be converted to depth D by $D = z \times d_\delta$, where the d_δ denotes the distance between two layers. In this paper, the radius of patches was set to 7, and the high-pass filter was the Gaussian filter whose variance and radius were set to 1 and 7, respectively. We generated a 3D point cloud of the testing set to evaluate the performance of the DIM. To verify the performance, we proposed a Euclidean distance based measurement to evaluate the accuracy of 3D point clouds. Because of the continuity of fibers, each section of fibers has similar depth. It can be summarized that the fibers in non-crossing regions can be cut into multiple short sections and the variance of depth of points in the same fiber section is defined by following:

$$S^2 = \frac{\sum^n (D - D_i)}{n} \tag{10}$$

where the S^2 denotes the variance between points in same section, n indicate the points number of the section. D_i and $\bar{D}$ are represent the depth of point and average depth of this section respectively. A smaller value of S^2 represents higher continuity and higher quality of the 3D point cloud.

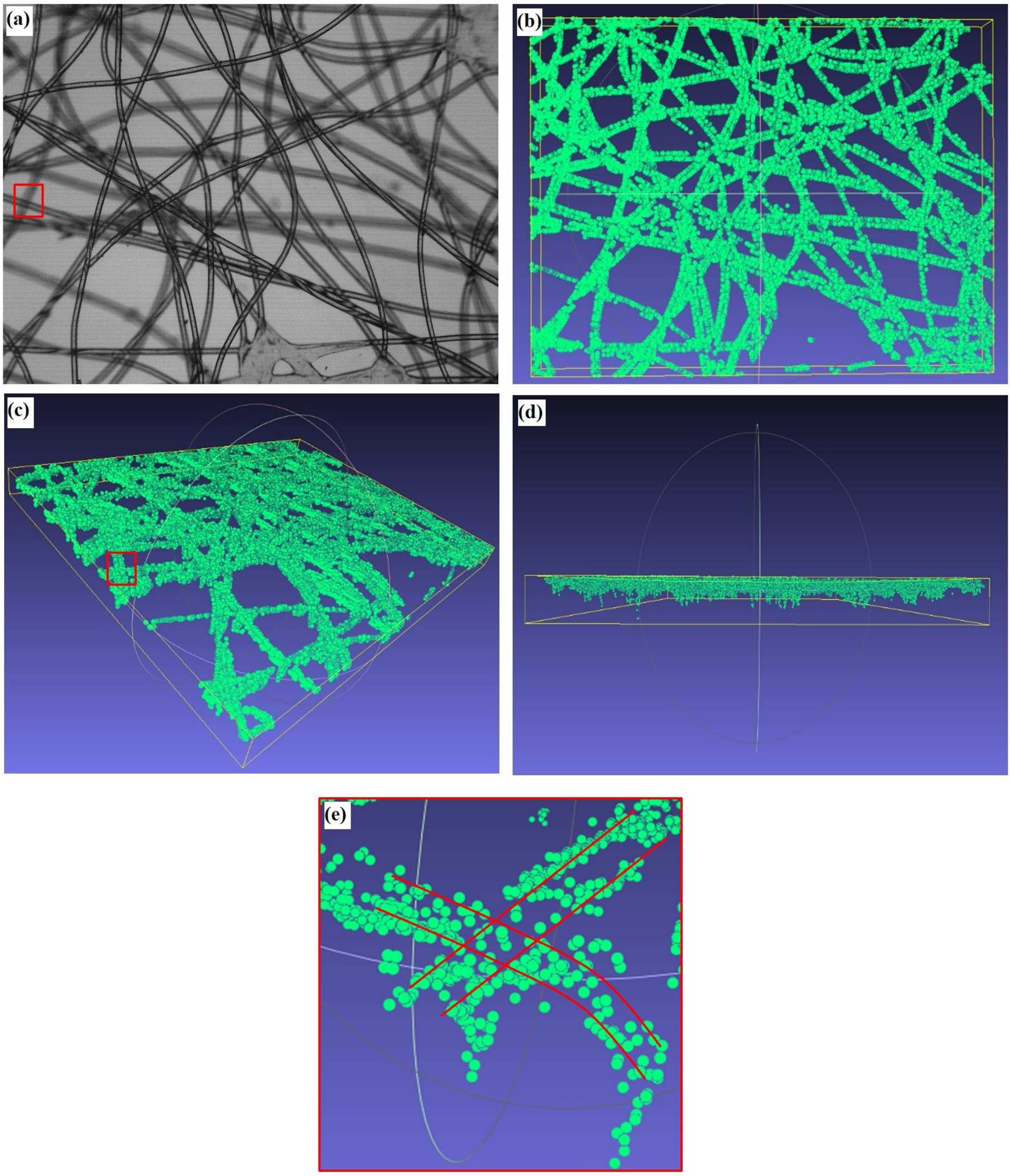


FIGURE 12. The point clouds of a test nonwoven. (a) one multi-focus image, (b)-(d) different views of the point clouds, and (e) enlarged view of a red box area in (c).

We marked the non-crossing regions in an image manually and cut fibers into 5-pixel length sections.

To compare the performance of the DIM with other method, we used the sharpness calculation strategy which is presented in [38] to find the most optimal to build the comparison model. The coefficients S^2 of 3D models which are built by the sharpness calculation and the DIM converge to 25.3 and 22.7, respectively. The S^2 of the DIM model is 10% lower than that of the sharpness calculation model, meaning that the DIM can generate more continuous point clouds with less noise than the comparison model. Figure 12 shows the 3D point clouds which are extracted by the DIM. The 3D point clouds exhibit a complete 3D structure of the nonwoven, even in those fiber crossing regions, and permits more accurate quantitative analysis on porosity and fiber orientations. Thus, the proposed model is effective in extracting

3D point clouds and filtering noise from multi-focus images.

V. CONCLUSION

In this research, we designed a multi-focus image deblurring network (MIDN) and a depth identification module (DIM) to extract 3D point clouds from microscopic images of a nonwoven web. The proposed network MIDN takes advantage of U-net and introduces the gradient maps to facilitate in-focus pixel selection. A new dataset of microscopic images was collected to build human annotations of ground truths. The experiments show that MIDN has better performance than the state-of-the-art networks on the testing dataset. The DIM combines the Fourier Transform and Gaussian kernels to find the minimum energies among multi-focus images. The experiments also demonstrate that DIM can deter the noise in point clouds effectively. The hybrid MIDN and DIM method can generate a complete, accurate 3D structures of a nonwoven web from its microscopic multi-focus images.

REFERENCES

[1] A. Santamaría-Pang, P. Hernandez-Herrera, M. Papadakis, P. Saggau, and I. A. Kakadiaris, "Automatic morphological reconstruction of neurons from multiphoton and confocal microscopy images using 3D tubular models," *Neuroinformatics*, vol. 13, no. 3, pp. 297–320, Jul. 2015.

[2] S. Wäldchen, J. Lehmann, T. Klein, S. van de Linde, and M. Sauer, "Light-induced cell damage in live-cell super-resolution microscopy," *Sci. Rep.*, vol. 5, no. 1, Dec. 2015, Art. no. 15348.

[3] J. Jonkman and C. M. Brown, "Any way you slice it—A comparison of confocal microscopy techniques," *J. Biomol. Techn.*, vol. 26, no. 2, pp. 54–65, Jul. 2015.

[4] B. Tamadazte, N. Le Fort-Piat, S. Dembélé, and G. Fortier, "Robotic micromanipulation for microassembly: Modelling by sequencial function chart and achievement by multiple scale visual servoings," *J. Micro-Nano Mechatronics*, vol. 5, nos. 1–2, pp. 1–14, Mar. 2009.

[5] X. M. Zhang, R. W. Wang, H. B. Wu, and B. Xu, "Automated measurements of fiber diameters in melt-blown nonwovens," *J. Ind. Textiles*, vol. 43, no. 4, pp. 593–605, Apr. 2014.

[6] Y. Jia, S. T. Bailey, T. S. Hwang, S. M. McClintic, S. S. Gao, M. E. Pennesi, C. J. Flaxel, A. K. Lauer, D. J. Wilson, J. Hornegger, J. G. Fujimoto, and D. Huang, "Quantitative optical coherence tomography angiography of vascular abnormalities in the living human eye," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 18, pp. E2395–E2402, May 2015.

[7] H. Majeed, S. Sridharan, M. Mir, L. Ma, E. Min, W. Jung, and G. Popescu, "Quantitative phase imaging for medical diagnosis," *J. Biophoton.*, vol. 10, no. 2, pp. 177–205, Feb. 2017.

[8] P. Nellist, H. Yang, L. Jones, G. Martinez, R. Rutte, B. Davis, T. Pennycook, M. Simson, M. Huth, H. Soltau, L. Strüder, R. Sagawa, Y. Kondo, and M. Humphry, "Efficient and quantitative phase imaging in two- and three-dimensions using electron ptychography in STEM," in *Proc. Eur. Microsc. Congr.*, 2016, pp. 517–518, doi: 10.1002/9783527808465.EMC2016.6511.

[9] C. Zuo, J. Sun, J. Zhang, Y. Hu, and Q. Chen, "Lensless phase microscopy and diffraction tomography with multi-angle and multi-wavelength illuminations using a LED matrix," *Opt. Express*, vol. 23, no. 11, pp. 14314–14328, Jun. 2015.

[10] V. Poher, H. X. Zhang, G. T. Kennedy, C. Griffin, S. Oddos, E. Gu, D. S. Elson, J. M. Girkin, P. M. W. French, M. D. Dawson, and M. A. A. Neil, "Optical sectioning microscopes with no moving parts using a micro-stripe array light emitting diode," *Opt. Express*, vol. 15, no. 18, pp. 11196–11206, Sep. 2007.

[11] D. F. Albeanu, E. Soucy, T. F. Sato, M. Meister, and V. N. Murthy, "LED arrays as cost effective and efficient light sources for widefield microscopy," *PLoS ONE*, vol. 3, no. 5, 2008, Art. no. e2146.

[12] J. S. Goldstein, I. S. Reed, and L. L. Scharf, "A multistage representation of the Wiener filter based on orthogonal projections," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2943–2959, Nov. 1998.

[13] A. V. Balakrishnan, "The Kalman filter," *Math. Intelligencer*, vol. 1, no. 2, pp. 90–92, 1978.

[14] D. S. C. Biggs, "3D deconvolution microscopy," *Current Protocols Cytometry*, vol. 52, no. 1, pp. 12.19.1–12.19.20, Apr. 2010.

[15] N. Dey, L. Blanc-Feraud, C. Zimmer, P. Roux, Z. Kam, J.-C. Olivo-Marin, and J. Zerubia, "Richardson–Lucy algorithm with total variation regularization for 3D confocal microscope deconvolution," *Microsc. Res. Technique*, vol. 69, no. 4, pp. 260–266, 2006.

[16] E. Mudry, K. Belkebir, J. Girard, J. Savatier, E. Le Moal, C. Nicoletti, M. Allain, and A. Sentenac, "Structured illumination microscopy using unknown speckle patterns," *Nature Photon.*, vol. 6, no. 5, pp. 312–315, May 2012.

[17] C. Preza and J.-A. Conchello, "Depth-variant maximum-likelihood restoration for three-dimensional fluorescence microscopy," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 21, no. 9, pp. 1593–1601, Sep. 2004.

[18] J. Conchello and J. W. Lichtman, "Optical sectioning microscopy," *Nature Methods*, vol. 2, no. 12, pp. 920–931, 2005.

[19] J.-A. Conchello, "Image estimation for structured-illumination microscopy," *Proc. SPIE*, vol. 5701, pp. 34–41, Mar. 2005.

[20] J.-A. Conchello and E. W. Hansen, "Enhanced 3-D reconstruction from confocal scanning microscope images 1: Deterministic and maximum likelihood reconstructions," *Appl. Opt.*, vol. 29, no. 26, pp. 3795–3804, Sep. 1990.

[21] J. G. McNally, C. Preza, J.-A. Conchello, and L. J. Thomas, "Artifacts in computational optical-sectioning microscopy," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 11, no. 3, pp. 1056–1067, Mar. 1994.

[22] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. Int. Conf. Learn. Represent.*, 2013, pp. 1–10

[23] Y. Rivenson, Z. Gorocs, H. Gunaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," *Optica*, vol. 4, no. 11, pp. 1437–1443, 2017.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[25] M. Weigert, L. Royer, F. Jug, and G. Myers, "Isotropic reconstruction of 3D fluorescence microscopy images using convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 126–134.

[26] Y. A. Bunyak, O. Yu. Sofina, and R. N. Kvetnyy, "Blind PSF estimation and methods of deconvolution optimization," 2012, *arXiv:1206.3594*. [Online]. Available: http://arxiv.org/abs/1206.3594

[27] M. Weigert *et al.*, "Content-aware image restoration: Pushing the limits of fluorescence microscopy," *Nature Methods*, vol. 15, no. 12, pp. 1090–1097, Dec. 2018.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] A. Fakhry, T. Zeng, and S. Ji, "Residual deconvolutional networks for brain electron microscopy image segmentation," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 447–456, Feb. 2017.

[30] T. Nguyen, Y. Xue, Y. Li, L. Tian, and G. Nehmetallah, "Deep learning approach for Fourier ptychography microscopy," *Opt. Express*, vol. 26, no. 20, pp. 26470–26484, Oct. 2018.

[31] S. Li, J. T. Kwok, and Y. Wang, "Multifocus image fusion using artificial neural networks," *Pattern Recognit. Lett.*, vol. 23, no. 8, pp. 985–997, Jun. 2002.

[32] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning, "A conditional random field word segmenter for Sighan bakeoff 2005," in *Proc. 4th SIGHAN Workshop Chin. Lang. Process.*, 2005, pp. 168–171. [Online]. Available: https://www.aclweb.org/anthology/I05-3027

[33] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.

[34] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[35] A. C. Muller and S. Behnke, "Learning depth-sensitive conditional random fields for semantic segmentation of RGB-D images," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 6232–6237.

[36] P. Krahenbuhl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Neural Inf. Process. Syst.*, 2011, pp. 109–117.

[37] C. Bui-Thu, T. Do-Hong, T. Le-Tien, and H. Nguyen-Duc, "An efficiently phase-shift frequency domain method for super-resolution image processing," in *Proc. Int. Conf. Adv. Technol. Commun.*, Oct. 2009, pp. 68–73.

[38] Q. Zhao, B. Liu, and Z. Xu, "Research and realization of an anti-noise auto-focusing algorithm," in *Proc. 5th Int. Conf. Intell. Hum.-Mach. Syst. Cybern.*, vol. 2, Aug. 2013, pp. 255–258.

[39] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.-Proc. Track*, vol. 9, pp. 249–256, Jan. 2010.

[40] S. Xie and Z. Tu, "Holistically-nested edge detection," *Int. J. Comput. Vis.*, vol. 125, nos. 1–3, pp. 3–18, 2015.

[41] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1939–1946, Aug. 2019.

[42] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3828–3837.

[43] W. Zhao, F. Zhao, D. Wang, and H. Lu, "Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3080–3088.

[44] J. Park, Y.-W. Tai, D. Cho, and I. S. Kweon, "A unified approach of multi-scale deep and hand-crafted features for defocus estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1736–1745.

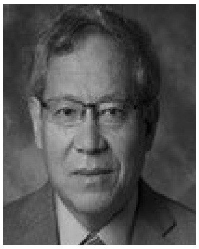
[45] J. Shi, L. Xu, and J. Jia, "Discriminative blur detection features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2965–2972.

[46] B. Su, S. Lu, and C. L. Tan, "Blurred image region detection and classification," in *Proc. 19th ACM Int. Conf. Multimedia (MM)*, Nov. 2011, pp. 1397–1400.

[47] X. Yi and M. Eramian, "LBP-based segmentation of defocus blur," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1626–1638, Apr. 2016.

JUE HOU is currently pursuing the Ph.D. degree in digital textile technology with Donghua University, Shanghai, China. He was a Visiting Student with the University of North Texas, USA. His research interests include computer vision, pattern recognition, and deep learning.

WENBIN OUYANG received the B.S. and M.S. degrees in electrical engineering from Donghua University, Shanghai, China, and the Ph.D. degree in computer science and engineering from the University of North Texas, USA, in 2018. His research interests include image processing, deep learning, and 3D technology.



BUGAO XU received the Ph.D. degree from the University of Maryland, College Park, in 1992. He joined the faculty of The University of Texas at Austin, in 1993. Since 2016, he has been a Professor and the Chair with the Department of Merchandising and Digital Retailing, and a Professor with the Department of Computer Science and Engineering, University of North Texas. His research interests include high-speed imaging systems, image and video processing, and 3D imaging and modeling.

RONGWU WANG received the Ph.D. degree from Donghua University, China, in 2008. He was a Postdoc with the Information Science and Technology School, Donghua University, China, from 2008 to 2010. He has been the faculty of the Donghua University, since 2010. Since 2014, he has been a Professor and the Chair with the Department of Nonwoven Material Science and Technology, Donghua University. His research interests include image processing, machine learning, textile imaging, and analysis.