Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Automatic skin lesion segmentation by coupling deep fully convolutional networks and shallow network with textons

Lei Zhang
Guang Yang
Xujiong Ye

SPIE.

Lei Zhang, Guang Yang, Xujiong Ye, "Automatic skin lesion segmentation by coupling deep fully convolutional networks and shallow network with textons," *J. Med. Imag.* **6**(2), 024001 (2019), doi: 10.1117/1.JMI.6.2.024001.

Automatic skin lesion segmentation by coupling deep fully convolutional networks and shallow network with textons

Lei Zhang,[a] Guang Yang,[b] and Xujiong Ye[a,*]
[a]University of Lincoln, Laboratory of Vision Engineering, School of Computer Science, Lincoln, United Kingdom
[b]Royal Brompton Hospital, Imperial College London and Cardiovascular Research Centre, National Heart and Lung Institute, London, United Kingdom

Abstract. Segmentation of skin lesions is an important step in computer-aided diagnosis of melanoma; it is also a very challenging task due to fuzzy lesion boundaries and heterogeneous lesion textures. We present a fully automatic method for skin lesion segmentation based on deep fully convolutional networks (FCNs). We investigate a shallow encoding network to model clinically valuable prior knowledge, in which spatial filters simulating simple cell receptive fields function in the primary visual cortex (V1) is considered. An effective fusing strategy using skip connections and convolution operators is then leveraged to couple prior knowledge encoded via shallow network with hierarchical data-driven features learned from the FCNs for detailed segmentation of the skin lesions. To our best knowledge, this is the first time the domain-specific hand craft features have been built into a deep network trained in an end-to-end manner for skin lesion segmentation. The method has been evaluated on both ISBI 2016 and ISBI 2017 skin lesion challenge datasets. We provide comparative evidence to demonstrate that our newly designed network can gain accuracy for lesion segmentation by coupling the prior knowledge encoded by the shallow network with the deep FCNs. Our method is robust without the need for data augmentation or comprehensive parameter tuning, and the experimental results show great promise of the method with effective model generalization compared to other state-of-the-art-methods. [DOI: 10.1117/1.JMI.6.2.024001]

Keywords: melanoma; skin lesion segmentation; fully convolutional networks; textons.

Paper 18210RR received Sep. 19, 2018; accepted for publication Mar. 29, 2019; published online Apr. 15, 2019.

1 Introduction

Maligant melanoma is one of the most aggressive and life-threatening skin cancers in the world and can strike men and women of all ages, races, and skin types.[1,2] The annual incidence has increased dramatically over the past few years, with the estimated of 132,000 melanoma skin cancer occur globally each year, according to Skin Cancer Foundation statistics.[3] Melanoma has different stages ranging from 1 to 3. Early diagnosis and treatment of melanoma is of critical clinical importance for high survival rate, e.g., early stage melanomas can often be cured with surgical excision.[1]

Dermoscopy is commonly used to diagnose skin cancer. A dermoscope is a noninvasive handheld skin imaging device that uses optical magnification and cross-polarized lighting to aid dermatologists in examining diagnostic details under the skin surface.[4] Manual interpretation of dermoscopic images is subjective depending on the experience of the dermatologists. There is a large amount of variation in assessing the probabilities of incidence and the malignancy level of the tissue. Hence, there is a growing demand for efficient computer-aided skin lesion analysis for early melanoma detection.

Standard approaches in computer-aided diagnosis (CAD) of dermoscopic images consist of three steps: dermoscopic image lesion segmentation, feature extraction, and disease classification. Segmentation of skin lesions is an essential step for accurate classification and diagnosis of the skin lesions. However, despite much effort being devoted to skin lesion segmentation, accurate delineation of skin lesions still remains an ongoing challenge. Examples of several major difficulties are low contrast between the lesion and the surrounding skin (different colourings inside the lesion), irregular and fuzzy lesion borders, large amount of artifacts, and intrinsic cutaneous features such as skin lines, blood vessels, air bubbles, hairs, and perspective distortion

In this paper, we describe a fully automatic framework for accurate skin lesion segmentation by coupling a deep fully convolutional network (FCN) with a shallow network with textons derived from domain specific filter kernels.

2 Previous Work

Existing approaches in the literature for skin lesion segmentation can be roughly categorized into traditional histogram-based thresholding, clustering, edge-based detection, region-based detection, morphological detection, model-based, active contours (snakes and their variants), and supervised learning-based methods. Comprehensive surveys on skin lesion segmentation in dermoscopic images are presented in Refs. 5–7. Celebi et al.[8] applied the ensemble of thresholding methods to segment skin lesions in dermoscopic images. The results from threshold fusion presented robust skin lesion segmentation. However, the method may not perform well on images when large amounts of artifacts appear (e.g., caused by hair or bubbles) as these can alter the histogram significantly, which in turn may result in

*Address all correspondence to Xujiong Ye, E-mail: XYe@lincoln.ac.uk

2329-4302/2019/$25.00

biased threshold computations. Barcelos and Pires[9] employed an anisotropic diffusion filter prior to Canny's edge detector to segment lesion edges. The results showed that most of the unwanted edges were removed. However, some regions of the skin lesions were missed. Cavalcanti et al.[10] proposed independent component analysis (ICA)-based active-contours method for skin lesion segmentation. ICA was first used to generate a reliable binary mask for initializing the active contour model implemented using Chan–Vese.[11] This is then followed by morphological operations as a postprocessing step. More recently, Bozorgtabar et al.[12] exploited the contextual information of skin image at the superpixel level and applied Laplacian sparse coding to calculate the probabilities of the skin image pixels to delineate lesion border, this is then followed by a dynamic rule-based refinement. Multiscale superpixel with cellular automata[13] and delaunay triangulation[14] have also been applied for skin lesion segmentation with some degrees of success. However, these methods may fail to accurately segment some skin lesions, such as lesions that touch the image boundary or with significant artifacts appearing in the images.

Recent advances in machine learning, especially in the area of deep learning, such as convolutional neural networks (CNNs), have dramatically improved the state-of-the-art in identifying, classifying, and quantifying underlying patterns in medical images. In particular, exploiting hidden hierarchical feature representations learned solely from data acts at the core of the advances. A recent review of the successes of deep learning in application to medical image registration, segmentation, computer-aided disease diagnosis, or prognosis, is presented in Ref. 15. Mostly recently, much progress has also been achieved by deep learning-based methods in automated skin lesion segmentation.[2,16,17]

Bi et al.[16] exploited FCN to automatically segment skin lesions. To address the limitation of coarse segmentation boundaries produced by the original FCN due to the lack of label refinement and consistency (especially for the skin lesions that have blurry boundaries and/or low variance in the textures between the foreground and the background), multistage FCN was investigated to learn complementary visual characteristics of the different skin lesions. Early stage FCNs learned coarse appearance and localization information while late-stage FCNs learned the subtle characteristics of the lesion boundaries. The complementary information derived from individual segmentation stages was then combined to obtain the final segmentation results. This method has demonstrated promising results on ISBI 2016 skin lesion challenge dataset. Yu et al.[17] improved the segmentation performance via exploiting the CNN net depth, in which a very deep CNNs (ResNet) was employed to increase model capacity that provided promising segmentation. Yuan et al.[2] presented a fully automatic method for skin lesion segmentation using 19-layer deep FCNs. To handle the strong imbalance between the number of foreground and background pixels, instead of using standard cross entropy-based loss function, a loss function based on Jaccard distance was designed to eliminate the need of sample reweighting that is required for the cross entropy-based loss function.

Among these deep learning-based skin lesion segmentation methods, many efforts have been made to focus on designing FCN architectures with specific loss functions to achieve a better performance. However, few attempts are taken to encode clinical valuable prior-knowledge into deep learning architectures to enable accurate segmentation of skin lesions. Toward this direction, this paper aims to develop a general framework, which allows the context information to be modeled and integrated into a deep FCN for fully automated skin lesion segmentation (Fig. 1), where an FCN is designed to fuse information from the intermediate convolutional layers to the output through skip connections and deconvolutional layers so that both low-level appearance information and high-level semantic information can be considered. In addition to the data-driven features derived from the FCN, clinical prior-knowledge of the skin lesions considering the low-level edge and texture features is also taken into account. Those low-level edge and texture features are derived from predefined filter kernels specific to skin lesions using a shallow convolutional network, which is then built into the FCN workflow using convolution operators. The proposed network architecture is trained in an end-to-end manner. In such a way, the domain-specific features can be supplementary to other hierarchical and semantic features learned from the FCN to enhance the fine details of skin lesions for a more accurate segmentation.

Our main contributions are summarized below:

(1) We present a fully automatic framework for accurate skin lesion segmentation by coupling a deep FCN with a shallow network with textons derived from domain specific filter kernels.

(2) We introduce a convolutional shallow network, which allows the clinical prior knowledge (textures) to be modeled and work with a deep FCN complementarily for skin lesion segmentations.

(3) We propose an efficient fusing strategy to combine domain-specific hand-crafted texton features into a deep network that is trained in an end-to-end manner.

(4) Our experimental results suggest that model generalization capability can be improved by introducing the context information into the FCN without the need of data augmentation or comprehensive parameter tuning.

The reminder of this paper is organized as follows. We describe the details of our method in Sec. 3. Experimental settings and results are reported in Sec. 4, with further discussions in Sec. 5. Conclusions are drawn in Sec. 6.

3 Method

3.1 *Overview of the Proposed Method*

The architecture of the proposed segmentation method is composed of two networks: an FCN and a texton-based shallow network. These two networks are integrated complementarily to enable more accurate skin lesion segmentations. Figure 1 illustrates the overall framework of the method. Both data driven and hand-crafted features are taken into account for the segmentation. More specifically, the hierarchical semantic features are learned by the FCN, whereas the context information related to the skin lesion boundaries is modeled by texton derived from the shallow network. These two networks are then integrated by fusing feature maps generated from each network using convolution operators. These two networks interact with each other in the learning stage that enables a more detailed segmentation.

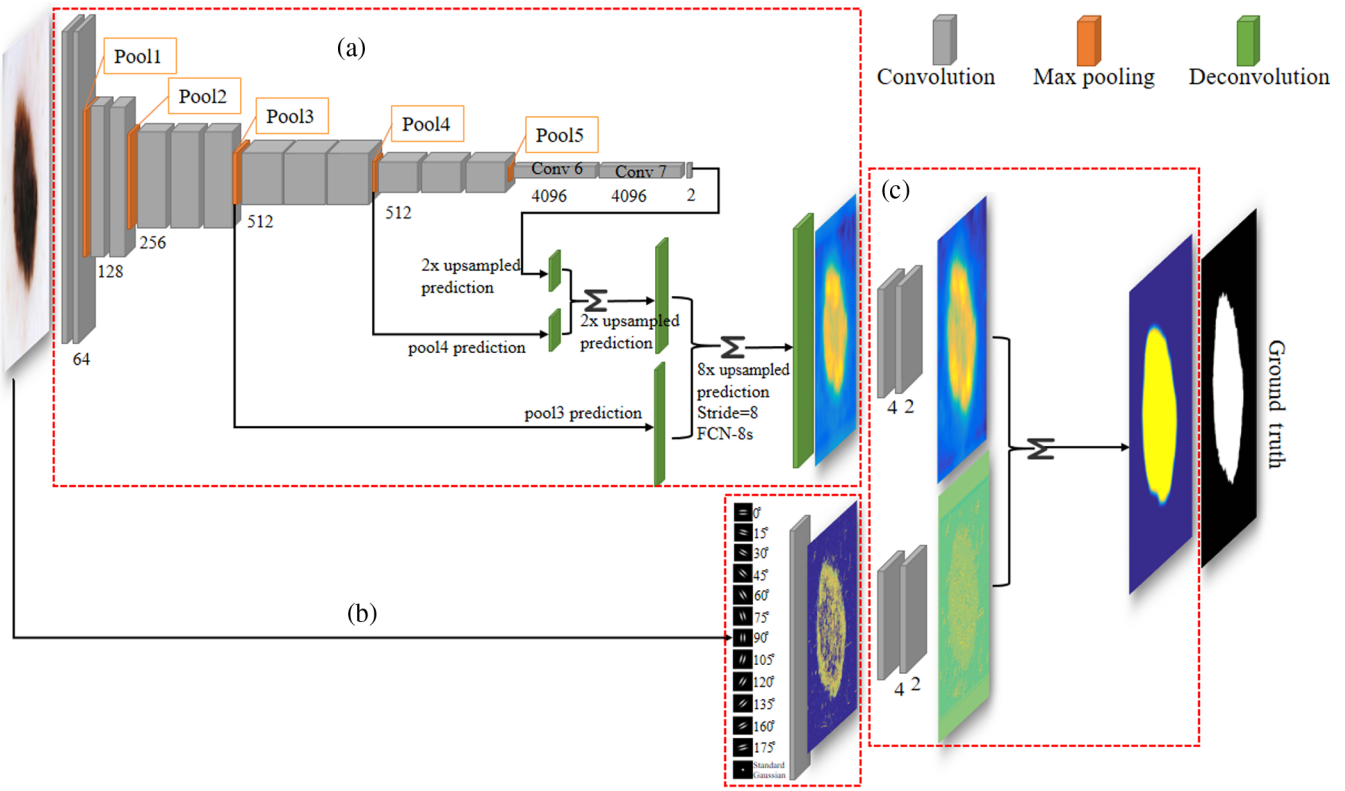

Fig. 1 The framework of the proposed method. The proposed architecture is composed of two networks: (a) an FCN with VGG16 backbone architecture and (b) a texton-based shallow network. These two networks are integrated complementarily in (c) feature maps fusing block to enable more accurate skin lesion segmentations.

In our method, the basis of feature learning process for both networks is derived from the convolution operations. For the deep network, the convolutional filter kernels are represented by weights learned automatically from raw image data, whereas the texton-based shallow network learns primitive elements by manually designing filter kernels based on domain specific knowledge (e.g., skin lesion boundaries and textures). The combination of these two types of networks has two advantages: on the one hand, some important cues derived from clinical prior-knowledge or context information are emphasized during the learning process; on the other hand, the automated weights learning scheme designed in the FCN helps in optimizing the hand-crafted feature map.

3.2 *Fully Convolutional Network*

The FCN architecture has been proven to be the state-of-the-art for semantic segmentation, in which the segmentations are obtained by a pixel-wised prediction. The FCN is trained via mapping an input to its labeled ground truth in an end-to-end supervised learning manner. In this paper, we train the FCN to learn the hierarchical features using the architecture described in Ref. 18. Figure 1 shows the FCN architecture with VGG16 (CNN classification net).[19] The VGG16 architecture is used as a base net in the FCN due to the following reasons: (1) more representative features can be learned *via* a stack of two or three convolutional layers with small filters (3×3), as it increases the complexity of a nonlinear function; (2) the problem of the limited number of training data can be tackled *via* transferring learning,[20–22] namely the pretrained model learned from abundant natural images can be used to train the FCN for skin lesion segmentation; and (3) the deep 16 layers architecture with a large number of weights could encode more complex high-level semantic features.

The VGG16 net is composed of five stacks followed by three transformed convolutional layers, where each block contains several convolutional layers and pooling layers. The convolutional layers can be viewed as filtering-based feature extractors. The filter responses are generated from local receptive fields in the feature maps of the previous layer by discrete convolution operations, which are defined as

$$\mathrm{Y} = \sum_{i=1}^{M} W_i^l X_i^{l-1} + b^l, \tag{1}$$

where W_i^l is the filter kernel in the current layer l, which includes the weights. The input map X_i^{l-1} $(i = 1, \ldots, M)$ is the feature map in the $(l-1)$'th layer. The b^l is the bias in the l'th layer. The result of the local weighted sum is passed to an activation function $f(.)$. Study in Ref. 23 showed that the ReLU can achieve better performance in terms of learning efficiency that results in faster training of very deep neural networks compared to the sigmoid method, we use the ReLU as an activation function in each convolutional layer. The pooling layer provides a mechanism that retains the spatial invariance of the features. However, it reduces the resolution of the feature maps. Typical pooling operations include subsampling and maxpooling. In our experiments, we employ the max-pooling operation

since an experimental study[24] has shown that a maximum pooling operation significantly outperforms subsampling operations.

The FCN is implemented by transforming the last fully connected layer of a regular CNN net (e.g., VGG16)[19] into convolutional layers, and then adding the upsampling and the deconvolutional layer to the converted CNN net. The replacement of the fully connected layer enables the net to accept image inputs with arbitrary sizes. The upsampling and deconvolutional layers produce the output activation map and make the size of the dense feature map to be consistent with the size of input image that enables pixel-wise prediction. As aforementioned, the pooling layer reduces the resolution of the feature map that may result in the coarse predictions, in this case, the skip connection was introduced in Ref. 18 to combine coarse predictions at deep layers and fine scale predictions at shallow layers that improve segmentation details. More specifically, a skip connection fuses 2× upsampled predictions computed on the last layer at stride 32 with predictions from Pool4 at stride 16. The sum of the two predictions is then upsampled back to the image with stride 16. In the same way, the relatively finer prediction maps (the stride 8 predictions) are obtained by fusing predictions of shallower layer (Pool3) with 2× upsampling of the sum of two predictions derived from Pool4 and the last layer.

Although the coarse segmentation issue can be mitigated using the skip connection, some details are still missing in the feature maps recovered from the shallow layer *via* the deconvolutional layers. Moreover, lack of the spatial regularization for the FCN may result in diminishing of the spatial consistency for the segmentations. The local dependency is not sufficiently considered in the FCN may also lead to some prediction disagreements among pixels within the same structure. In the next section, we introduce a shallow network to overcome this problem by integrating the textons-based spatial information into the FCN architecture.

3.3 *Shallow Network with Texton*

Texture is one of the representative spatial information that can provide discriminative features for pattern recognition tasks. The texton has shown its advantages in encoding texture information,[25] where the texture is represented by its responses to a set of filter kernels $(W_1, W_2, \ldots, W_n)$:

$$R = [W_1 * I(x,y), W_2 * I(x,y), \ldots, W_n * I(x,y)], \tag{2}$$

where * indicates the convolution operation, n is the number of filter kernels (W). The texton is defined as a set of feature vectors that are generated by clustering filter responses in R.

The design of an appropriate filter bank is a crucial step to extract specific texture features. Given the clinical prior knowledge that the edge information provides important cues in skin segmentation, we employ the second-order derivative Gaussian filter. The two-dimensional Gaussian is defined as

$$G(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{x^2}{2\sigma_x^2}} \times \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{y^2}{2\sigma_y^2}}. \tag{3}$$

This filter is implemented as an anisotropic filter by introducing an orientation parameter because the edges can be presented at any orientations. The rotated second-order partial derivative of Eq. (3) with respect to the y axis direction is given by

$$\frac{\partial''_{G(x',y')}}{\partial_y} = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{x'^2}{2\sigma_x^2}} \times \frac{1}{\sqrt{2\pi}\sigma_y^5}(y'^2 - \sigma_y^2) e^{-\frac{y'^2}{2\sigma_y^2}}$$
$$x' = x\cos\theta - y\sin\theta$$
$$y' = x\sin\theta - y\cos\theta. \tag{4}$$

This filter is designed as a specific edge detector with $\sigma = 1.5$ and $\theta =$ (0 deg, 15 deg, 30 deg, 45 deg, 60 deg, 75 deg, 90 deg, 105 deg, 120 deg, 135 deg, 150 deg, 165 deg). In addition, we employ standard Gaussian at scale $\sigma = 1$ to extract non-edge structures that also imposes a simple smoothness constrain for feature representations. In order to reduce the computational redundancy and increase the feature representations, for anisotropic filters, the maximal response to the filter kernels across all orientations is considered, whereas the response to the isotropic filter is recorded directly.

Textons are computed from the filter responses, which are generated by applying the filter kernels designed in Eq. (4) to the pixels in m training patches. Then these filter responses are clustered using a k-means clustering algorithm. As a result, the k cluster centroids can be represented as k vectors and the centroids of the clusters form the textons. In order to generate textons for both lesion and nonlesion, two sets of patches related to classes of lesion and nonlesion are prepared using ground truths. More specifically, for lesion class, the patches are cropped from images with original size guided by the ground truth mask, and for nonlesion class, the patches are cropped randomly from nonlesion regions in images. In the training stage, every patch in each set convolutes with the filters, filter responses generated from all patches in the same set are concatenated and clustered to generate textons of one class. In our method, there are $k \times c$ number of textons, where the k (e.g., $k = 8$) is the number of centroids in the k-means and c is the number of classes (i.e., $c = 2$, namely, lesion and nonlesion). All trained textons are stored into a dictionary (D), which will be used to calculate the texton map. The bottom flowchart in Fig. 2 illustrates this process.

Once the texton dictionary has been generated from the training stage, each input image is translated to a texton map using the similar implementation, which is performed to encode spatial context and ensure the intensity consistency of lesion. More specifically, given an input skin image, it first convolves with the filters in the filter bank to produce filter responses, and then each pixel in the image is assigned to one of the texton labels l_i $[l_i \in D\,\forall_i = (1, 2, 3, \ldots, k \times c)]$ from the texton dictionary (D) based on the minimum distance between the texton and the filter responses at the pixel. Through this process, a texton label map is generated, which is further converted into an intensity map. Namely, for the pixels with the same texton label index, mean intensity of those corresponding pixels in the input image is calculated. The label index in the texton label map is then replaced by the corresponding mean intensity. We call this map as a texton map, and this process is shown in the top flowchart in Fig. 2.

This shallow network encodes the global and local spatial information using convolutions with hand-designed filters from domain specific knowledge, which is able to decompose high-order visual features or structures to some primitive elements (here are edges, dots, and spots). Each image can be represented by different distributions of these elements depending on the designed filter bank. In this paper, each skin image is represented by the edges extracted using second-order partial

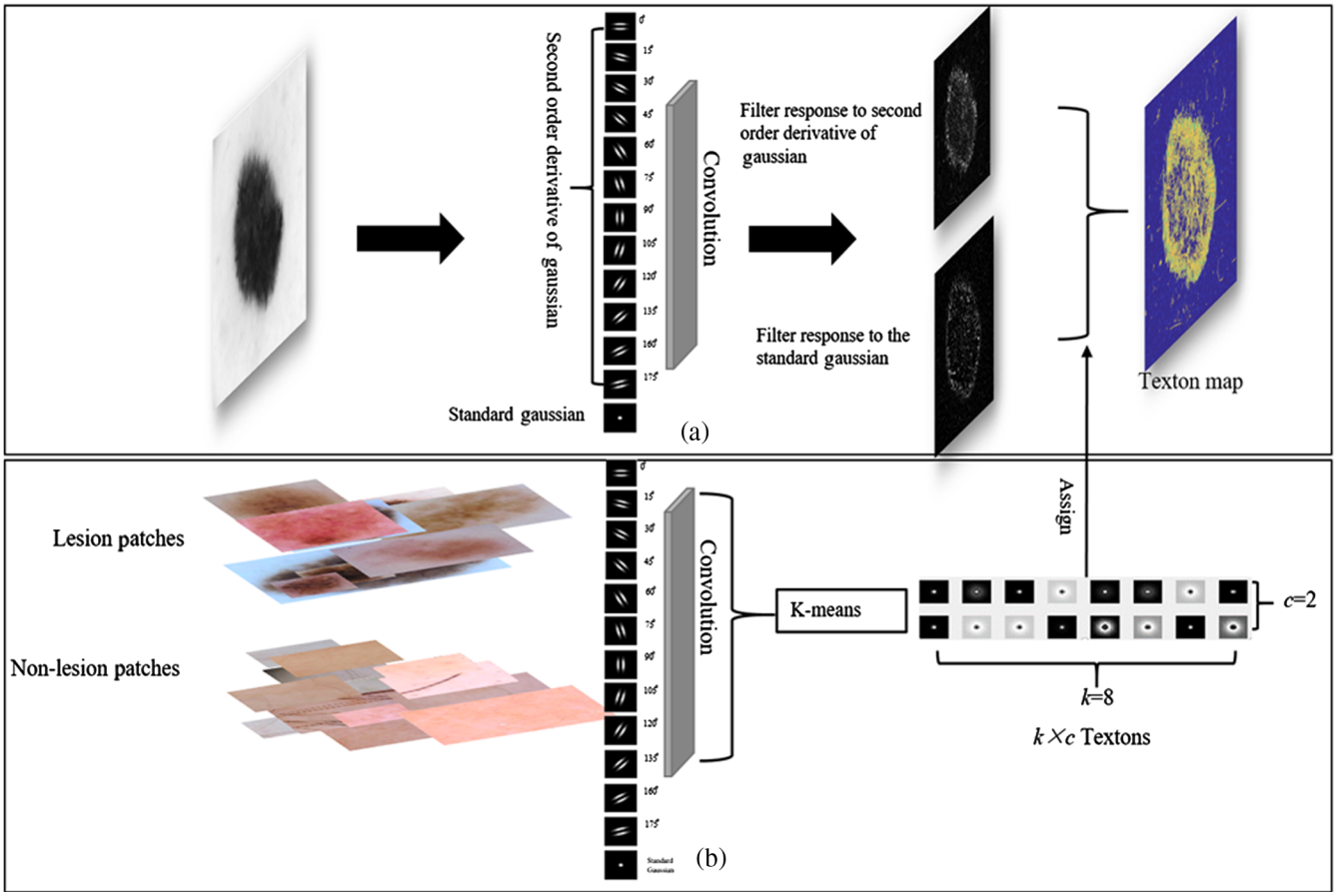

Fig. 2 The framework of the shallow network, where the texton dictionary generation process is shown in (b) bottom flowchart and (a) top flowchart shows the texton map generation process using the learned dictionary.

derivative of Gaussian and blobs extracted using the standard Gaussian. Moreover, depending on the number of k, we are able to discriminate edges with different gradient magnitudes. Therefore, strong boundaries of lesion and weak edges inside lesion can be distinguished. In addition, instead of using pooling-like operation used in CNN, which may reduce the resolution, in the shallow network, the clustering is implemented on the filter response images, which have the same size of the input image. Thus our shallow network may keep more edge details. It is of note that due to the shallow characteristic (e.g., without nonlinear transformation) of the network, some edges in nonlesion region (e.g., hair artifact) may be present, meanwhile, it may also have limitations to work with the case of extremely low-contrast lesions with fuzzy boundaries. However, these nonlesion edges could be suppressed and weak edge cues could be retained and enhanced by fusing the FCN feature map, which will be discussed in the next section.

3.4 *Fusing the Shallow and Deep Network*

The feature maps derived from the FCN and the shallow network with texton described above (red box in Fig. 1) are fused in a single network by an integrating block. Formally, let the desired mapping function as $M(x)$, which indicates a nonlinear transformation from an input x_l to an output x_{l+1}. We hypothesize that this function can be learned more effectively by introducing a model of prior-knowledge in the deep network. Instead of fitting the $M(x)$ directly by the deep neural network, we set the FCN mapping function $F(x)$ as $F(x) = M(x) - T(x)$. The original mapping function thus can be expressed by

$$M(x) = F(x) + T(x), \tag{5}$$

where $T(x)$ is the mapping function, which encodes the prior-knowledge described in Sec. 3.3. Regarding aforementioned mapping function $T(x)$, the output of the transformation is a constant, as the weights of the filter banks are predefined and fixed. In this case, adding $T(x)$ can affect the forward propagation while not influencing the backward propagation to the $F(x)$. This is because, in the backward propagation calculation, the gradient is calculated by (local gradient*upstream gradient) and due to both local gradients are $\frac{\partial M}{\partial F} = 1$ and $\frac{\partial M}{\partial T} = 1$, the upstream gradient derived from loss can be directly passed to the $F(x)$ for weights update while weights of the filter banks [encode the prior-knowledge through $T(x)$] remain unchanged.

In order to fuse the two maps in a complementary manner, we increase the function complexity by introducing two-block convolutional layers. The formulation Eq. (5) can then be expressed as

$$M(x) = C(F(x), \{W_i\}) + C(T(x), \{W_i\})$$
$$C = W_2\lambda(\mu, W_1), \tag{6}$$

where $\{W_i\}$ indicates a set of weights (i is the number of layer) in the convolutional block C with input μ and λ is the activation function ReLU.

In our experiment, each map (e.g., FCN and texton map) is inputted into a separate block, which contains two convolutional layers. The trainable weights in the convolutional layers are considered as filters, which are able to learn functions that allow two feature maps to be fused properly. Namely, in the training process, more details from textons feature map can be supplemented to the feature map of the FCN while the influences of nonlesion edges in the textons feature map are suppressed by the FCN feature map. In our experiments, the filter number and the kernel size are defined empirically as follows: for each block, there are two inputs and four outputs with kernel size of 5×5 pixels in the first convolutional layer. In the second convolutional layer, there are four inputs and two outputs with the same kernel size. Finally, sum of responses to the filters from each block is fed into the whole network, which is then trained by minimizing the softmax cross-entropy loss.

We can visually observe the improvement of our network in Fig. 3, where (a) is an input image (top) and the region with red box is a zoomed-in image (bottom) that we can observe clearly the local details in the image. Fig. 3(b) illustrates the integrated score map (top left in b) derived from our proposed network and its surface (bottom right in b). Fig. 3(c) is the FCN only score map with its surface. As we can see, the map in (b) has finer details than the one in (c), and the region with red box in (a) is predicted with high probability of being lesion in (b) using our proposed network, but it is missed in (c). For a further detailed observation, we can see much more local details in the surface of (b) than the surface shown in (c).

3.5 *Network Training*

Once the texton dictionary is generated, as discussed in Sec. 3.3, the texton map for each image can be calculated based on the minimum distance between a texton in the dictionary and the filter responses at each pixel in the image. This in turns enables our network to be trained in an end-to-end manner. More specifically, the input image goes through the FCN and shallow network in parallel to produce two feature maps, which are further fused using the two-block convolutional layers [shown in the red box in Fig. 1(c)]. The final score map is fed into the softmax with loss layer that forms a complete trainable network.

We employ minibatch stochastic gradient descent (SGD) with momentum[26] to train our network. In our experiment, we set the batch size of 20. The leaning rate for the FCN is set to be 0.0001 and momentum as 0.9. The learning rate in the last integrating layer is set to be 0.001. We initialize the network weights using pretrained VGG16 model for the FCN network and using the initialization as described in Ref. 27 for the integrating layer. We use dropper layers with rate of 0.5 after convolutional layers 6 and 7 in Fig. 1 to reduce the over-fitting.

4 Experimental Settings and Results

4.1 *Experimental Materials and Evaluation Metrics*

In our experiment, two publicly available datasets are used to train and evaluate our segmentation method. These datasets are provided by the International Skin Imaging Collaboration (ISIC) and are widely used for the International Symposium on Biomedical Imaging 2016 (ISBI 2016) and ISBI 2017 challenges, respectively.[28,29]

The ISBI 2016 challenge dataset includes a training set with 900 images and a testing set with 379 images. In the ISBI 2017 challenge dataset, 2000 images are provided as training data, 150 images and 600 images are provided as validation and testing data, respectively. Each image in both datasets is paired with a ground truth labeling in the form of a binary mask, which was obtained from manual delineation by an expert. Although the challenge organizer allows the participants to use some additional external training data, our method is only trained on the training dataset provided by the challenges and evaluated independently on the testing datasets for both challenges. This could verify the robustness and generalization of our

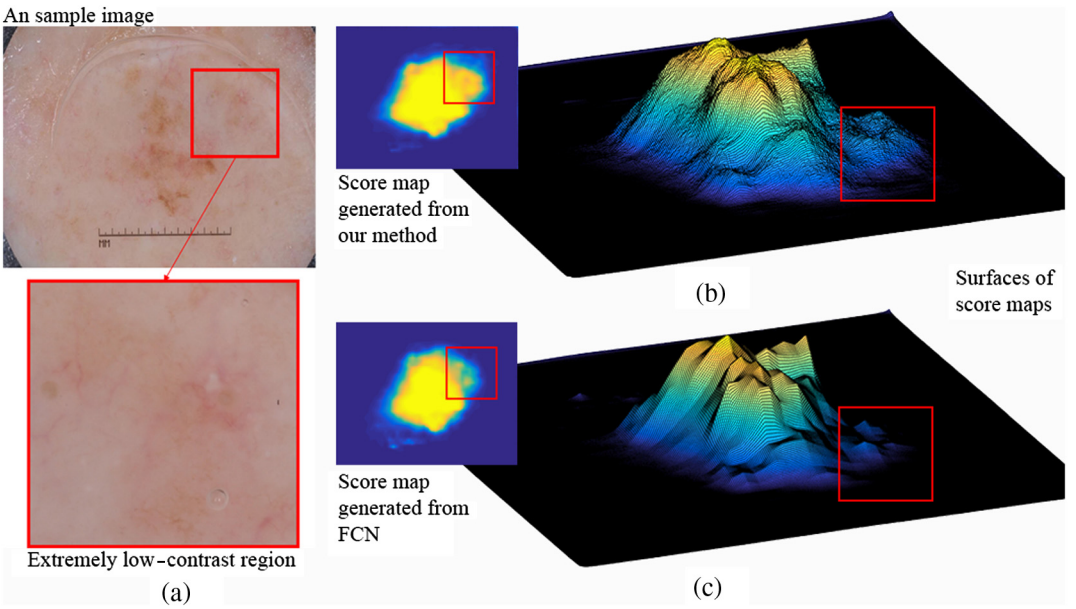

Fig. 3 The effects of fusing feature maps derived from two networks: (a) a sample image with extremely low contrast, particularly in the local region shown in zoomed in box, top image in (b) shows the integrated score map generated from our network with its surface (bottom). (c) The FCN only score map with its surface.

method while providing more comparable results obtained from other participants. The proposed method is evaluated on ISBI2016 and ISBI2017 datasets independently to verify the efficacy of integrating context information modeled by shallow architecture into a deep FCN.

In order to evaluate the performance of our automatic segmentation, we adopted region-based measures suggested in the ISBI challenge 2016 and 2017.[28,29] These evaluation metrics include segmentation accuracy, sensitivity, specificity, dice coefficient, and Jaccard index (JA).

4.2 *Implementation*

Our skin lesion segmentation framework is implemented using MatConvNet[30] with CUDA 7.5. The training and testing are implemented on a PC with a CPU of Inter® i7-4790k at 4.00 GHz and a GPU of Nvidia GeForece GTX 980 Ti with 6 GB GDDR5. All our training and testing data are publicly available from the ISBI challenges website.[28,29]

In our experiments, the proposed network was trained separately using ISBI 2016 and ISBI 2017 training datasets, the validation dataset of ISBI 2017 was adopted to validate the convergence of the proposed network and to determine the maximal number of epochs, batches, size and learning rate, etc. The optimal epoch that yields the best performance on the validation dataset is saved as trained model. Due to the lack of validation dataset in ISBI 2016, we first randomly divided the 900 training data into two sets (800 images for training and 100 images for validation) to obtain the optimal epochs, then our network is trained using the entire ISBI2016 training data using the same settings obtained from the validation set. More specifically, the network was trained with the batch size of 20 and the total number of epochs is set as 200. The leaning rate for the FCN is 0.0001 with momentum of 0.9. The learning rate in the last integrating layer is 0.001. In the training stage, we resize each image to a fixed size (e.g., 384×384 pixels). In order to preserve the original image information, such as the height to width ratio, each image is resized by a factor with the same height to width ratio of the original image and then we perform zero paddings to the image. The number of trainable parameters in the proposed network and the original FCN are 135,066,820 and 135,066,008, respectively. The slightly increased number of parameters in the proposed network comes from introducing the convolutional block to fuse the feature maps from shallow and deep networks. However, more parameters in the proposed network along with the additional computational cost of texton map generation only lead to a little more training time compared to the standard FCN. It takes about 15.25 h to train the proposed network over 200 epochs with the 6 GB GTX 980Ti on the ISIB2016 dataset against about 15 h of training the standard FCN.

In the testing stage, the final score map is fed into a softmax layer that provides a posterior probability map. In order to evaluate our automatic segmentation with respect to the ground truth masks, a simple dual-threshold method[2] is employed to produce a binary output. More specifically, the initial lesion region candidates are obtained by applying the thresholding value of 0.95 on the probability map, and the region that has the maximal number of pixels is kept as lesion candidate. Then a relatively lower thresholding value of 0.5 is applied to the probability map to obtain the whole lesion region.

Table 1 Evaluation results with different input sizes

Image size	Accuracy	Dice	Jaccard	Se[a]	Sp[a]
384×384	94.9471	0.8976	**0.8277**	91.8634	95.2059
256×256	94.9110	0.8970	0.8255	90.0697	95.6065

[a]The Se denotes the sensitivity and Sp is the specificity.
Note: The best Jaccard value is marked as bold.

4.3 *Experimental Results*

4.3.1 *Comparison of different input sizes*

The input size is one of the factors that affect the segmentation performances since a larger size of the input image with relatively higher resolution contains more details. However, with the increasing resolution, it may make the network training more challenging, due to the issues, such as over-fitting and slow convergence. In our experiment, we evaluated and compared the segmentation performances using two input size settings: 256×256 and 384×384. Because of the usage of the VGG16 net, these numbers are the multiples of 32. The evaluation results with different input sizes using ISBI 2016 dataset are summarized in Table 1. We can observe that the JA of segmentations with input size of 384×384 is slightly better than that with input size of 256×256. Figure 4 shows the comparable segmentation results with different input sizes, where the green contour is the ground truth, the red and blue contours are the automated segmentation results with the input of 384×384 and 256×256, respectively. We can observe that segmentations with 384×384 are relatively finer than those generated from the inputs with size of 256×256. For example, comparing red and blue contours, more details with a convex region of red contour are presented at left-hand side of Fig. 4(a) and the red contour with more local details as shown in Fig. 4(d). These results indicate that relatively higher resolution provides more image details, and our proposed network has the model capacity to learn such details.

4.3.2 *Results of different integrating layer settings*

As described in Sec. 3.4, we fuse both feature maps using convolution operators. In our experiments, we compare the segmentation performances using three different settings. Settings (a) and (b): each map is fed into one convolutional layer with different sizes of filter [e.g., 7×7 (a) and 5×5 (b)], and setting (c): each map is fed into two convolutional layers with kernel size of 5×5. The size of all input images is 384×384. The evaluation results for both ISBI 2016 and 2017 datasets are summarized in Table 2. We can see that when one convolutional layer is used, by changing the filter size from 7×7 (a) to 5×5 (b), the sensitivity is increased from 89.7343 to 90.2079 for the ISBI 2016 dataset and increased from 82.3548 to 83.9807 for the ISBI 2017 dataset; however, specificities for both datasets are reduced, from 96.0928 to 95.3334 and from 96.6418 to 96.1468, respectively. For 2016 dataset, the evaluation metric of JA for the setting (b) with filter size of 5×5 is improved to 0.8249 compared with 0.8216 of using setting (a) with filter size of 7×7. A consistent result with the increasing JA can also be found in 2017 dataset, i.e., JA of 0.7252 for the setting (a) against JA of 0.7262 for the setting (b). These results show that increasing the filter size in one convolutional layer may not

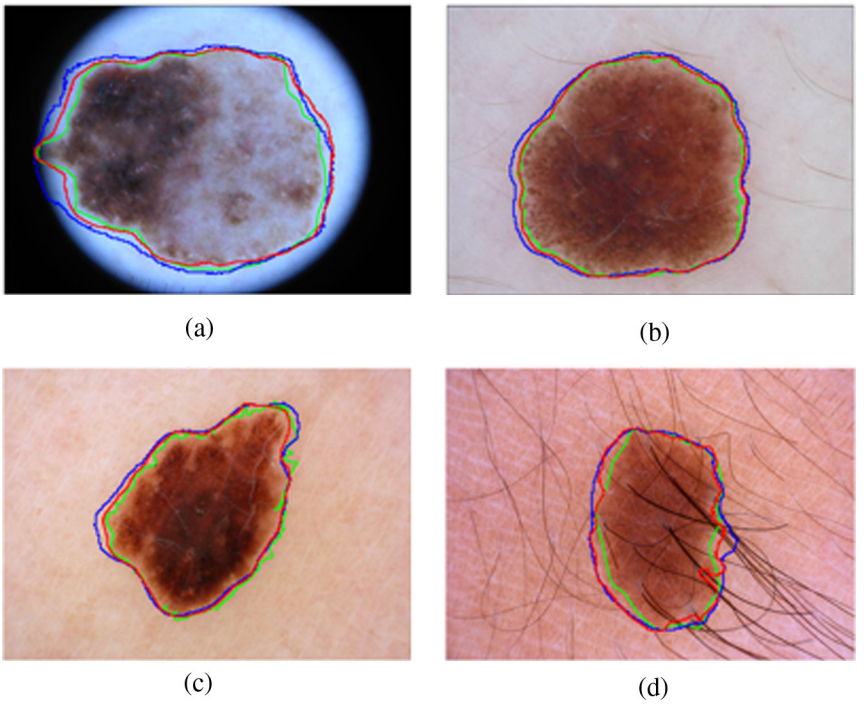

Fig. 4 The comparable segmentation results with different input sizes. (a)–(d) Four examples illustrated, where the green contour in (a)–(d) is the ground truth, the automated segmentation results of our method with input image size of 384×384 and 256×256 are presented in red and blue contours, respectively.

Table 2 Evaluation results with different network settings in the integrating layer

Data sets	N[a]	AC	Dice	Jaccard	Se	Sp
2016	a	94.9641	0.8928	0.8216	89.7343	96.0928
	b	94.8577	0.8958	0.8249	90.2079	95.3334
	c	94.9471	0.8976	**0.8277**	91.8634	95.2059
2017	a	92.5956	0.8142	0.7252	82.3548	96.6418
	b	92.6068	0.8165	0.7262	83.9807	96.1468
	c	92.7284	0.8181	**0.7294**	83.7175	96.3854

Note: N[a] indicates different network settings, where network "a" represents the setting: a convolutional layer with filter size of 7×7 for each feature map, and setting "b" is a convolutional layer with filter size of 5×5, and "c" is two convolutional layers with filter size of 5×5. The best Jaccard values are marked as bold.

Table 3 Results of skin lesion segmentation methods using the ISBI 2016 challenge testing dataset

Method	AC	DI	JA	SE	SP
Yuan[2]	**0.955**	**0.912**	**0.847**	0.918	0.966
EXB	0.953	0.910	0.843	0.910	0.965
CUMED[17]	0.949	0.897	0.829	0.911	0.957
Ours	0.949	0.898	0.828	**0.919**	0.952
Mahmudur	0.952	0.895	0.822	0.880	**0.969**
FCN	0.941	0.886	0.814	0.917	0.949
SFU-mial	0.944	0.885	0.811	0.915	0.955
TMUteam	0.946	0.888	0.810	0.832	0.987

Note: A total of 28 teams participated on the ISBI 2016 challenge; the top five performances on the challenge were collected from the leaderboard and the result of Yuan was collected from their paper. In addition, the results of using FCN only are included for comparison. The best performances are marked as bold.

increase the model generalization. However, it can be seen from Table 2, adding more convolutional layers can improve the segmentation performance, the network with setting (c) achieves relatively better performance in terms of JA comparing to the setting (a) and (b) for both datasets. The performance improvement is due to the fact that the stacks of convolutional layers with 5×5 receptive fields contain more weights compared to a single convolutional layer with 5×5 receptive fields, also the nonlinearities with the stack layers increase the complexity of nonlinear function compared to the settings with a single convolutional layer. Overall, the performances of our method with three sets outperform the FCN on both datasets (Tables 2–4), of which the setting (c) achieved best performance. For ISBI 2016 dataset, our method achieves 94.9471 of accuracy, 0.8976 of Dice, 0.8277 of JA, 91.8634 of sensitivity, and 95.2059 of specificity, whereas the evaluation results for ISBI 2017 dataset reach 92.7284 of accuracy, 0.8181 of Dice, 0.7294 of Jaccard, 83.7175 of sensitivity, and 96.3854 of specificity.

Figure 5 shows the results of our fully automatic segmentation method for the ISBI 2016 testing dataset [Figs. 5(a)–5(d)] and ISBI 2017 testing dataset [Figs. 5(e)–5(h)]. These examples

Table 4 Results of skin lesion segmentation methods using the ISBI 2017 challenge testing dataset

Method	AC	DI	JA	SE	SP
Yading	0.934	0.849	0.765	0.825	0.975
Matt	0.932	0.847	0.762	0.820	0.978
INESC	0.922	0.824	0.735	0.813	0.968
Ours	0.927	0.818	0.729	**0.837**	0.964
FCN	0.923	0.811	0.719	0.816	0.964
Vic	0.922	0.810	0.718	0.789	0.975
Juana	0.915	0.797	0.715	0.774	0.970

Note: The best performances are marked as bold.

cover several challenging cases, which include cases with low contrast [Fig. 5(a)], cases with multiple objects [Figs. 5(a) and 5(e)], cases with other artifacts like hair [Fig. 5(b) and 5(f)], dark mask (d), and other tissues (h), and cases with inhomogeneous surface [Figs. 5(c), 5(d), and 5(g)]. We can observe that our segmentation results (red contours) are almost identical to the corresponding ground truths (green contours) in Figs. 5(a), 5(c), 5(e), 5(g), 5(d), and 5(h) and are slightly under-segmented in Figs. 5(b) and 5(f). These results show that our proposed network has a reasonable model capacity to handle images with various image qualities and can provide accurate lesion segmentation.

4.3.3 *Comparison with other methods on the ISBI 2016 challenge*

We compare the results of our method to the top five ranked teams using the ISBI 2016 challenge testing dataset. The results are listed in Table 3, where a recent work[2] that obtained the best performance is also included for comparison. These methods are ranked according to the JA. It can be seen that our method has shown promising results achieving a competitive performance compared to CUMED team[17] (Table 3). Compared to our relatively shallower architecture (26 layers), the method in Ref. 17 employed a fully convolutional residual network (FCRN) for lesion segmentation that is a very deep network with 50 layers. Therefore, in contrast to the FCRN used in Ref. 17, our network has a similar model capacity but fewer layers to be trained and thus training our framework is much more efficient. The methods proposed by EXB and Yuan[2] achieved better results than our method (Table 3). However, it is of note that our model is only trained on the raw training images provided by the challenge without utilizing any additional dataset and also without applying any data augmentation. Moreover, comparing to the standard FCN, we can observer that our network incorporating the context information with the FCN outperforms the FCN only network on all metrics using the ISBI 2016 challenge dataset.

4.3.4 *Comparison with other methods on the ISBI 2017 challenge*

Table 4 compares the results of our model to the other deep learning-based methods on the ISBI 2017 challenge. Most methods included in Table 4 have adopted the strategy of data augmentation or an additional dataset to train CNN networks such as ResNet and U-net. However, our model is trained using only 2000 training images provided by the ISBI 2017 challenge dataset without applying any data augmentation. The results of our method have shown competitive performance compared to the other teams, and we have achieved the highest sensitivity of 0.837. It is noted that, the main aim of this study is to investigate a general framework while proving its efficacy, which allows the context information to be modeled and integrated into a deep FCN. We did not boost the segmentation performance via comprehensive pre- and postprocessing steps or other ensemble schemes. With the same training protocol and same hyperparameters settings, we can also observe our network outperforms the standard FCN on the ISBI 2017 challenge dataset. These consistent results obtained from both 2016 and 2017 datasets indicate that the generation capability of the network can be improved by introducing the prior-knowledge into the

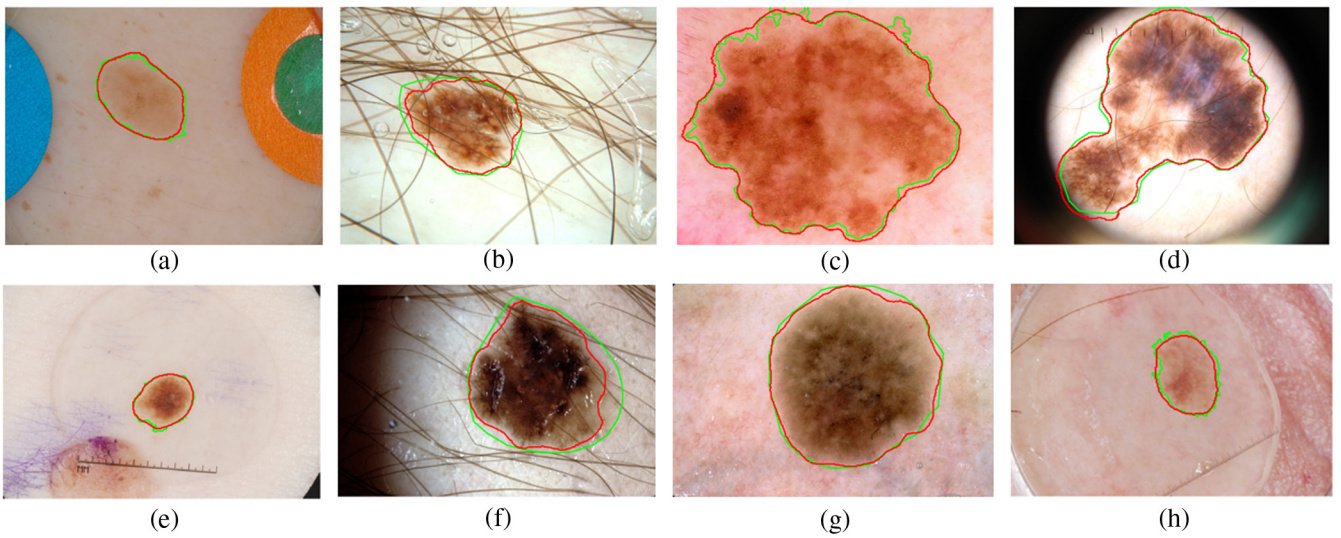

Fig. 5 Segmentation results for two testing datasets: (a)–(d) ISBI 2016 and (e)–(h) ISBI 2017 testing datasets. The green contour is the ground truth and the segmentation results of our methods are presented in red contours.

deep neural network. In addition, our skin lesion segmentation framework is very efficient that takes 0.12 s to infer a input image with the size of 1536×2048pixels.

5 Discussion

In this study, we propose a deep learning framework to couple the FCN derived data driven features with hand-crafted texton features from a shallow network by introducing an integrating block (e.g., two-block convolutional layers) trained in an end-to-end manner. The framework performed well on the skin lesion segmentation without the need of complicated data augmentation or comprehensive parameter tuning.

Although a few studies for skin lesion segmentation have been reported in the literature,[2,13,16,17] effective fully automatic skin lesion segmentation still remains a challenging task due to large variations of skin lesions in the dermoscopic images. This is due to influences of various artifacts, low contrast and illumination, heterogeneous lesion texture, fuzzy boundaries, etc. Among these previous CNN-based methods, two typical networks with different optimized approaches have achieved promising performance. One approach is to design a network, which is accompanied with recent techniques, e.g., reducing the overfitting using batch normalization and speeding up the training using different SGD optimizations, as well as improving the segmentation performance by adapting a specific loss function.[2] The other method is to increase the model capacity via adding more layers in the network or employing a very deep network such as ResNet.[17]

In the field of biomedical image analysis, given the limited number of training samples especially limit reliable ground truths, a common practice for accelerating the network training is initializing the weights with pretrained models learned from abundant nature images, or so-called fine-turning.[20] However, for some networks with recent techniques (e.g., batch normalization and new activation layer), it might be difficult to directly use the well pretrained models (e.g., VGG16) for the weights initialization. Moreover, for the very deep networks, the vanish gradient and degradation problems[31] make the training procedure challenging. Although a series of efforts have been made in CNN-based methods to increase the model capacity and improve the segmentation performance, few attempts are taken to encode clinical valuable prior-knowledge into deep learning architectures to enable accurate segmentation of skin lesion. In those cases, it is worthwhile to investigate different strategies to exploit the potential of the deep CNNs that is beneficial to the different challenging biomedical applications with limited training samples. This study is along this research direction, with an application of skin lesion segmentation in dermoscopic images. In our study, instead of increasing the network depth, designing task specific loss function or using different explicit regularizations, we proposed an alternative strategy to exploit the potential of the deep CNNs, coupling the hand-crafted features modeled from prior-knowledge with data driven features learned from a deep neural network into a single-deep network. More specifically, in our framework, the clinical prior-knowledge is modeled by textons, simultaneously, more abstract features that may be difficult to be encoded by traditional feature engineering techniques are learned automatically by the FCN. It is of note that our framework has significant potential to be extended to other biomedical image segmentation tasks because it is capable of incorporating prior clinical relevant knowledge or domain knowledge from an expert.

With the limited training samples, employing explicit regularization such as data augmentation is one of the commonly

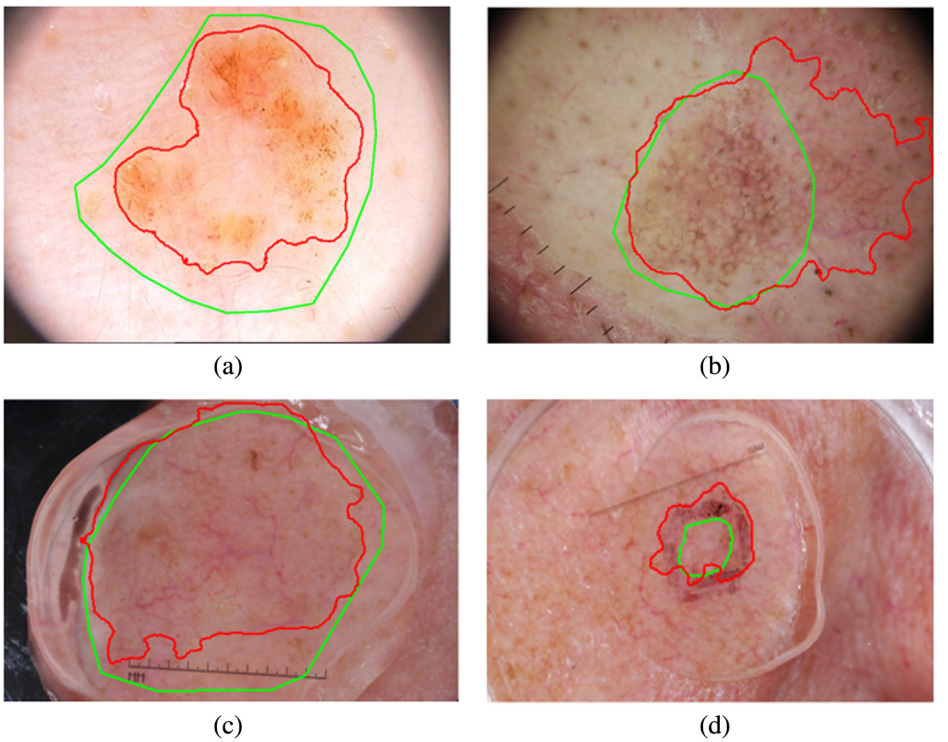

Fig. 6 Some failure cases of our segmentations: (a) and (b) ISBI 2016 testing dataset, and (c) and (d) ISBI 2017 testing dataset. The green contour is the ground truth and the segmentation results of our methods are presented in red contours.

used strategies to improve the model performance. While data augmentation may reduce the generalization error, it is not a major determinant for model generalization.[32] This can be observed from Tables 3 and 4, where we compared our method with the other top-ranked methods, which employed deep CNNs and data augmentation. We can also observe that the absolute differences of accuracy by comparing our method (without data augmentation) to top five performers (with data augmentation) in Table 3 ranged from 0 to 0.006. In some cases, our architecture without data augmentation outperforms those methods with data augmentation in terms of accuracy. The results have demonstrated the generalization capability of our model.

Although our method has achieved promising segmentation results, there are still some poor cases, which are shown in Fig. 6, where (a) and (c) are two under-segmentation examples and (b) and (d) show two over-segmentation examples. We can observe that most of those cases have inhomogeneous appearances and irregular shapes. To further improve the performance, it is worthwhile to exploit the specific local and global patterns of skin lesion via other texture analysis method, e.g., more advanced local dependency and contextual constraints modeled via the Markov random field can be integrated into our framework to further increase the model generalization. It is also worthwhile to employ additional datasets to boost the segmentation performance and to further assess the model generalization. The proposed network presented in this paper may inspire further studies on how to exploit different hand-crafted features and how to integrate them into a deep network to tackle other problems in the field of biomedical image processing and analysis.

6 Conclusion

In this paper, we proposed a deep learning-based method for fully automatic skin lesion segmentation using dermoscopic images. By coupling the FCN with a shallow network, we fused the hierarchal features and hand-crafted texton features efficiently. Experiments on ISBI 2016 and ISBI 2017 skin lesion challenge datasets have demonstrated promising results and effective model generalization compared to other state-of-the-art methods. Our experimental results show that the generation capability of the network can be improved by introducing the prior-knowledge into the deep neural network. Compared to a very deep network (e.g., ResNet), our relatively shallower network can still achieve comparable performance for skin lesion segmentation. The method could also be potentially adapted to other medical image segmentation applications.

Disclosures

The authors declare they have no conflict of interest with regard to the work presented.

Acknowledgments

The authors are grateful to the International Skin Imaging Collaboration, the organizers of International Symposium on Biomedical Imaging 2016 and 2017 (ISBI 2016, 2017) challenge of "Skin lesion analysis toward melanoma detection," who make the data sets publicly available.

References

1. D. S. Rigel, R. J. Friedman, and A. W. Kopf, "The incidence of malignant melanoma in the United States: issues as we approach the 21st century," *J. Am. Acad. Dermatol.* **34**(5), 839–847 (1996).
2. Y. D. Yuan, M. Chao, and Y. C. Lo, "Automatic Skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE Trans. Med Imaging* **36**(9), 1876–1886 (2017).
3. WHO, "How common is skin cancer?," https://www.who.int/uv/faq/skincancer/en/index1.html (2017).
4. S. H. Argenziano et al., *Dermoscopy: A Tutorial*, EDRA Medical Publishing & New Media, Milan (2002).
5. R. B. Oliveira et al., "Computational methods for the image segmentation of pigmented skin lesions: a review," *Comput. Methods Prog. Biomed.* **131**, 127–141 (2016).
6. F. Thompson and M. K. Jeyakumar, "Review of segmentation methods on malignant melanoma," in *Proc. IEEE Int. Conf. Circuit, Power and Comput. Technol. (ICCPCT 2016)* (2016).
7. M. E. Celebi et al., "A state-of-the-art survey on lesion border detection in dermoscopy images," in *Dermoscopy Image Analysis*, pp. 97–129, CRC Press, Boca Raton, Florida (2015).
8. M. E. Celebi et al., "Lesion border detection in dermoscopy images using ensembles of thresholding methods," *Skin Res. Technol.* **19**(1), E252–E258 (2013).
9. C. A. Z. Barcelos and V. B. Pires, "An automatic based nonlinear diffusion equations scheme for skin lesion segmentation," *Appl. Math. Comput.* **215**(1), 251–261 (2009).
10. P. G. Cavalcanti et al., "An ICA-based method for the segmentation of pigmented skin lesions in macroscopic images," in *Annu. Int. Conf. IEEE Eng. Med. and Biol. Soc.*, pp. 5993–5996 (2011).
11. T. E. Chan, B. Y. Sandberg, and L. A. Vese, "Active contours without edges for vector-valued images," *J. Visual Commun. Image Represent.* **11**(2), 130–141 (2000).
12. B. Bozorgtabar, M. Abedini, and R. Garnavi, "Sparse coding based skin lesion segmentation using dynamic rule-based refinement," *Lect Notes Comput. Sci.* **10019**, 254–261 (2016).
13. L. Bi et al., "Automated skin lesion segmentation via image-wise supervised learning and multi-scale superpixel based cellular automata," in *IEEE 13th Int. Symp. Biomed. Imaging (ISBI)*, pp. 1059–1062 (2016).
14. A. Pennisi et al., "Skin lesion image segmentation using Delaunay triangulation for melanoma detection," *Comput. Med. Imaging Graphics* **52**, 89–103 (2016).
15. D. G. Shen, G. R. Wu, and H. I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
16. L. Bi et al., "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Trans. Bio-Med. Eng.* **64**(9), 2065–2074 (2017).
17. L. Q. Yu et al., "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imaging* **36**(4), 994–1004 (2017).
18. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 3431–3440 (2015).
19. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. on Learning Representations* (2014).
20. N. Tajbakhsh et al., "Convolutional neural networks for medical image analysis: full training or fine tuning?" *IEEE Trans. Med. Imaging* **35**(5), 1299–1312 (2016).
21. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *Lect. Notes Comput. Sci.* **8689**, 818–833 (2014).
22. R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 580–587 (2014).
23. X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. and Stat.*, pp. 315–323 (2011).
24. D. Scherer, A. Muller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," *Lect. Notes Comput. Sci.* **6354**, 92–101 (2010).
25. T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vision* **43**(1), 29–44 (2001).
26. N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Networks* **12**(1), 145–151 (1999).
27. K. M. He et al., "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *IEEE Int. Conf. Comput. Vision*, pp. 1026–1034 (2015).

28. D. Gutman et al., "Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)," https://arxiv.org/abs/1605.01397 (2016).
29. G. D. Codella et al., "Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)," https://arxiv.org/abs/1710.05006 (2017).
30. A. Vedaldi and K. Lenc, "MatConvNet convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, pp. 689–692 (2015).
31. K. M. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 770–778 (2016).
32. C. Y. Zhang et al., "Understanding deep learning requires rethinking generalization," in *ICLR*, Toulon, France (2017).

Lei Zhang received his MSc and PhD degrees from the University of East Anglia, UK, in 2008 and 2014, respectively. He is currently a research fellow at the University of Lincoln working in the Laboratory of Vision Engineering. His main research interests are computer vision and machine learning. In particular, medical image processing and analysis, image segmentation, and object detection.

Guang Yang received his MSc degree in vision imaging and virtual environments from the University College London (UCL), Department of Computer Science in 2006 and his PhD jointly from the UCL Centre for Medical Image Computing and Department of Computer Science and Medical Physics in 2012. He is currently an image processing physicist and a senior research fellow working at Cardiovascular Research Centre, Royal Brompton Hospital, and affiliate with the National Heart and Lung Institute, Imperial College London.

Xujiong Ye is a professor of medical imaging and computer vision in the School of Computer Science, University of Lincoln, UK. He has more than 20 years' research and development experience in medical imaging and computer vision from both academia and industry. Her main research is to develop computational models using advanced medical image analysis, computer vision, and artificial intelligence to support clinicians in decision-making.