Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks

Lequan Yu,* *Student Member, IEEE*, Hao Chen, *Student Member, IEEE*, Qi Dou, *Student Member, IEEE*, Jing Qin, *Member, IEEE*, and Pheng-Ann Heng, *Senior Member, IEEE*

***Abstract*—Automated melanoma recognition in dermoscopy images is a very challenging task due to the low contrast of skin lesions, the huge intraclass variation of melanomas, the high degree of visual similarity between melanoma and non-melanoma lesions, and the existence of many artifacts in the image. In order to meet these challenges, we propose a novel method for melanoma recognition by leveraging very deep convolutional neural networks (CNNs). Compared with existing methods employing either low-level hand-crafted features or CNNs with shallower architectures, our substantially deeper networks (more than 50 layers) can acquire richer and more discriminative features for more accurate recognition. To take full advantage of very deep networks, we propose a set of schemes to ensure effective training and learning under limited training data. First, we apply the residual learning to cope with the degradation and overfitting problems when a network goes deeper. This technique can ensure that our networks benefit from the performance gains achieved by increasing network depth. Then, we construct a fully convolutional residual network (FCRN) for accurate skin lesion segmentation, and further enhance its capability by incorporating a multi-scale contextual information integration scheme. Finally, we seamlessly integrate the proposed FCRN (for segmentation) and other very deep residual networks (for classification) to form a two-stage framework. This framework enables the classification network to extract more representative and specific features based on segmented results instead of the whole dermoscopy images, further alleviating the insufficiency of training data. The proposed framework is extensively evaluated on ISBI 2016 *Skin Lesion Analysis Towards Melanoma Detection* Challenge dataset. Experimental results demonstrate the significant performance gains of the proposed framework, ranking the first in classification and the second in segmentation among 25 teams and 28 teams, respectively. This study corroborates that very deep CNNs with effective training mechanisms can be employed to solve complicated medical image analysis tasks, even with limited training data.**

***Index Terms*—Automated melanoma recognition, fully convolutional neural networks, residual learning, skin lesion analysis, very deep convolutional neural networks.**

I. Introduction

MELANOMA is a type of cancer that mostly starts in pigment cells (melanocytes) in the skin. It is regarded as the most deadly form of skin cancer and accounts for about 75% of deaths associated with skin cancer [1]. According to American Cancer Society, about 76380 new cases of melanomas are estimated to be diagnosed and about 10130 fatalities are estimated in United States in 2016 [2]. Fortunately, if melanoma is detected in its early stages and treated properly, the survival rate is very high [3], [4].

In order to improve the diagnostic performance of melanoma, dermoscopy technique was developed. Dermoscopy is a noninvasive skin imaging technique of acquiring a magnified and illuminated image of a region of skin for increased clarity of the spots on the skin [5]. By removing surface reflection of skin, it can enhance the visual effect of deeper levels of skin and hence provide more details of skin lesions. Dermoscopy assessment is widely used in the diagnosis of melanoma and obtains much higher accuracy rates than evaluation by naked eyes [6]. Nevertheless, the manual inspection from dermoscopy images made by dermatologists is usually time-consuming, error-prone and subjective (even well trained dermatologists may produce widely varying diagnostic results) [5]. In this regard, automated recognition approaches are highly demanded.

Automated melanoma recognition from dermoscopy images is, however, a very challenging task. First, the huge intraclass variation of melanomas in terms of color, texture, shape, size and location in the dermoscopy images as well as the high degree of visual similarity between melanoma and non-melanoma lesions make it difficult to discriminate melanomas from non-melanoma skin lesions. Second, the relatively low contrasts and obscure boundaries between skin lesions (especially at their early stages) and normal skin regions make the automated recognition task even harder. Finally, the presence of artifacts, either natural (hairs, veins) or artificial (air bubbles, ruler marks, color calibration

Manuscript received November 15, 2016; revised December 9, 2016; accepted December 12, 2016. Date of publication December 21, 2016; date of current version April 1, 2017. This work was supported in part by the grants from the Research Grants Council of the Hong Kong Special Administrative Region under Project CUHK 14203115 and Project CUHK 14202514, and the Guangdong Natural Science Foundation (No. 2016A030313047). *Asterisk indicates corresponding author.*

*L. Yu is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: lqyu@cse.cuhk.edu.hk).

H. Chen, Q. Dou, and P.-A. Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: hchen@cse.cuhk.edu.hk; qdou@cse.cuhk.edu.hk; pheng@cse.cuhk.edu.hk).

J. Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong (e-mail: harry.qin@polyu.edu.hk).

Digital Object Identifier 10.1109/TMI.2016.2642839

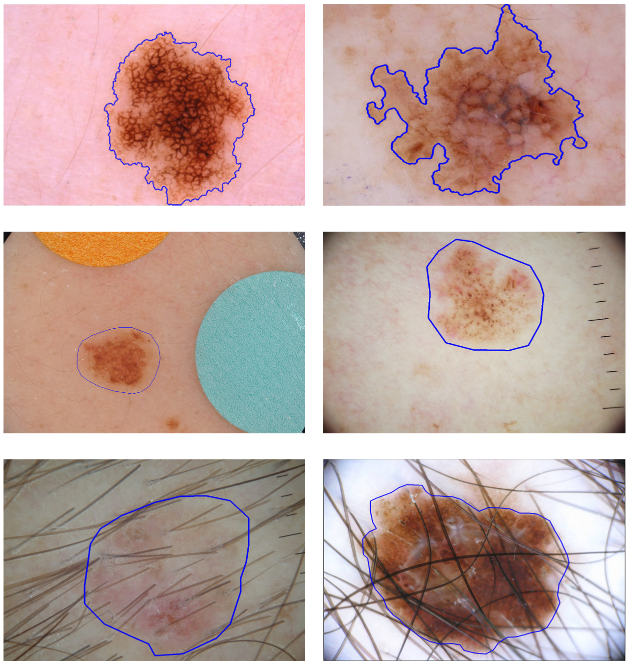


Fig. 1. Automated melanoma recognition from dermoscopy images is a very challenging task. The main challenges include (from top to bottom): high degree of visual similarity between melanoma and non-melanoma lesions, relatively low contrast between skin lesions and normal skin regions, and artifacts in images. The left column images show non-melanomas and right column images show melanomas. Blue contours indicate the skin lesions.

charts, etc.) may blur or occlude the skin lesions and further aggravate the situation. We show these challenges by some examples in Fig. 1.

A lot of efforts have been dedicated to solving this challenging problem. Early investigations attempted to apply low-level hand-crafted features to distinguish melanomas from non-melanoma skin lesions, including shape [7], color [8], [9] and texture [10], [11]. Some researchers further proposed to employ feature selection algorithms to select proper features and utilized combinations of these low-level features to improve the recognition performance [12], [13]. However, these hand-crafted features are incapable of dealing with the huge intraclass variation of melanoma and the high degree of visual similarity between melanoma and non-melanoma lesions, as well as the artifacts existing in dermoscopy images, leading to unsatisfactory results. On the other hand, some researchers also proposed to perform segmentation first and then based on the segmentation results to recognize the melanomas [11]–[14]. The segmentation allows the feature extraction procedure to be conducted only on the lesion regions and thus generate more specific and representative features. But in these methods, both the segmentation and classification procedures are still based on low-level features with limited discrimination capability.

Recently, convolutional neural networks (CNNs) with hierarchical feature learning capability have led to breakthroughs in many medical image analysis tasks, including classification [15], [16], detection [17]–[20] and segmentation [21], [22]. Some researchers started to employ CNNs for melanoma classification, aiming at taking advantage of their discrimination capability to achieve performance gains. Codella *et al.* proposed to integrate CNNs, sparse coding and support vector machine (SVM) for melanoma recognition [23]. Kawahara *et al.* presented a fully convolutional neural network based on AlexNet [24] to extract representative features of melanoma [25]. But these methods either just rely on the features trained from natural image dataset (such as ImageNet [26]) without sufficiently considering the characteristics of melanoma or utilize CNNs with quite shallow architecture. They can not well deal with the challenges of melanoma recognition. There is still much room to tap the potentials of CNNs to further improve the accuracy of melanoma recognition.

Many theoretical investigations [27], [28] and practical studies [29], [30] have demonstrated that network depth is a major factor of model expressiveness. The discrimination capability of features learned from CNNs can be enriched by increasing the number of stacked layers (network depth). The performance gain of very deep networks in natural image processing tasks has been exploited by recent works [29]–[31]. A straightforward thought is that if we can harness very deep networks to solve challenging medical image analysis problems, such as melanoma recognition. However, finding a good solution is not that straightforward. One of the main concerns is that, compared with natural image processing problems, the training data of medical applications is usually quite limited. This makes it difficult to effectively train very deep networks with a large amount of parameters. Another challenge is that the interclass variation in medical image analysis tasks is usually much smaller than that in natural image processing tasks (e.g., the interclass variation between melanoma and non-melanoma lesions is much smaller that interclass variation between person and car).

In this paper, we propose a novel method based on very deep CNNs with a set of effective training schemes in order to meet the challenges of automated melanoma recognition. Similar to some previous works, we propose to first segment the skin lesions from dermoscopy images and then classify them into melanoma ones and non-melanoma ones so that the classification stage can extract more specific and representative features within the lesion regions instead of performing it in the whole dermoscopy images. We employ very deep networks (more than 50 layers) for both the segmentation and the classification stages in order to obtain more discriminative features for more accurate recognition. To overcome the degradation problem [32] when a network goes deeper, we utilize residual learning technique [31] in our framework. For effective and accurate skin lesion segmentation, we further construct a fully convolutional residual network (FCRN) incorporating a multi-scale contextual information integration scheme. We extensively evaluate the proposed framework on ISBI 2016 *Skin Lesion Analysis Towards Melanoma Detection* Challenge dataset. Experimental results demonstrate the significant performance gains of the proposed framework, ranking the first in classification and the second in segmentation among 25 teams and 28 teams, respectively.

The main contributions of our work can be summarized as follows:

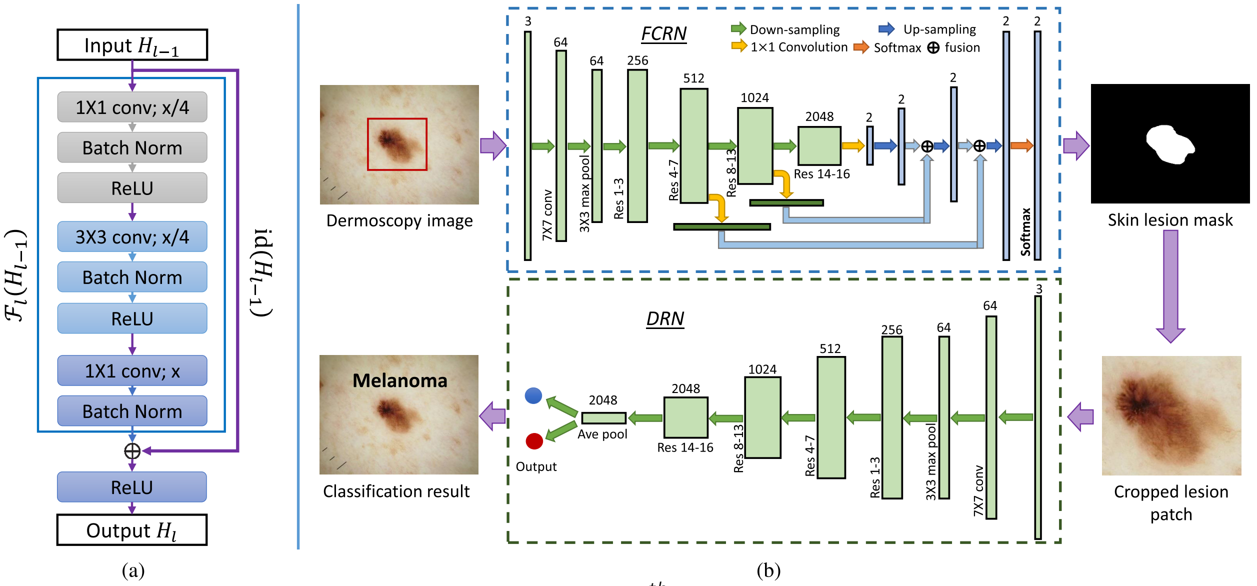


Fig. 2. The flowchart of the proposed framework. (a) Illustration of the ℓ^{th} residual block. x is the number of feature maps of output. We use 1×1 convolutional layers to reduce (and restore) the dimensions. (b) Architectures of fully convolutional residual network (FCRN) for skin lesion segmentation and very deep residual network (DRN) for skin lesion classification. Both of the two networks are trained using the residual learning technique. The numbers above residual blocks represent the number of feature maps. The down-sampling operator is performed by 7×7 convolutional layer (stride is 2), max pooling layer (stride is 2) and convolutional layers (stride is 2) appearing in the first convolutional layers of block 4, 8 and 14.

1) We propose a novel and comprehensive two-stage approach based on very deep CNNs with a set of effective training schemes to ensure the performance gains of increasing network depth with limited training data for automated melanoma recognition. To the best of our knowledge, we are not aware of any previous work that employs such substantially deeper networks (more than 50 layers) in medical image analysis field. Experiments demonstrate that, compared with much shallower counterparts, the very deep CNNs are capable of acquiring richer and more discriminative features and achieving better performance.
2) We propose a very deep fully convolutional residual network (FCRN) for accurate skin lesion segmentation, and further enhance its capability by incorporating a multi-scale contextual information integration scheme. The network is general enough and can be easily extended to solve other medical image segmentation tasks with targeting objects having large variations.
3) We compare the performance of networks with different depths and corroborate that very deep CNNs with effective training mechanisms can be employed to solve complicated medical image analysis tasks, even with limited training data. This may inspire more studies to tap the potentials of network depth of CNNs to solve challenging medical image analysis problems.

The remainder of this paper is organized as follows. We introduce the details of our method in Section II. Experiments and results are reported in Section III. We further discuss our method in Section IV and conclusions are drawn in Section V.

II. Methods

As mentioned above, melanomas have huge intraclass variation and there is a high degree of visual similarity between melanoma and non-melanoma lesions, which will severely influence the recognition performance if we directly perform skin lesion classification on original dermoscopy images, especially considering the limited training data in our hand. In this regard, we propose to meet these challenges using a two-stage framework, as shown in Fig. 2 (b). We first construct a very deep fully convolutional residual network, which incorporates multi-scale feature representations, to segment skin lesions. Based on the segmentation results, we employ a very deep residual network to precisely distinguish melanomas from non-melanoma lesions. In this section, we first briefly introduce the background and basics of residual networks (Section II-A) and then detail the proposed FCRN for segmentation (Section II-B) and the deep residual network for classification (Section II-C), respectively.

A. Background on Very Deep Residual Networks

Although some progress has been achieved by deep learning based methods in automated melanoma recognition [23], [25], there is still a gap between the results of expert assessment and the results obtained from automated methods due to the challenges mentioned above [7]. In this regard, how to learn more discriminative representations of skin lesions for more robust analysis with limited training data is still an open problem. Computational theory has evidenced that the network depth is a key factor of model expressiveness for a long time [27], [28]. Recent research works on large-scale image recognition tasks further demonstrate that increasing network depth can achieve significant quality gains [29], [30]. Inspired by these works, we propose to exploit very deep CNNs, aiming at extracting more discriminative features, to cope with this challenging task.

However, exploiting very deep networks is not as easy as stacking more layers. Training a very deep network for effective skin lesion analysis is quite difficult. When a network goes deeper, its accuracy gets saturated and even degrades rapidly aimed converging to a solution [31]. In other words, it is usually more difficult for a deeper network to find an optimal solution than its shallower counterparts. In this regard, we have to effectively deal with the *degradation problem* [31] when training a deeper network to hold on the performance gains achieved by increasing the network depth. In addition, the *vanishing gradient* problem may become more obvious when training a deeper network, making it difficult to tune the parameters of the early layers in the network [33], [34]. Furthermore, in skin lesion analysis, as well as many other medical image analysis tasks, limited quantity training dataset further exacerbate the difficulties in training a very deep network, as deeper networks usually have more parameters than shallower counterparts and require more training samples to retain their generalization capability.

Recent years, some training schemes have been proposed to effectively train a network when its layers increases, including careful initialization [34], deep supervision in hidden layers [35], and batch normalization [36]. While these schemes can largely alleviate the vanishing gradients problem, they are incapable of handling the degradation problem and do not experimentally demonstrate accuracy gains with substantially increased depth [37]. In order to ease the training of very deep networks and take full advantage of its performance gains, we exploit the newly developed residual learning technique [31] to train our skin lesion analysis framework. This technique introduces extra skip connections to improve the information flow within the network and explicitly reformulate the layers as learning residual functions with reference to the layer inputs and, by this way, addresses the degradation problem of deeper network. Meanwhile, we also employ some widely used mechanisms, such as batch normalization, to deal with the vanishing gradients problem. We briefly introduce the fundamentals of residual leaning here and readers can refer to [31], [37] for more details.

A deep residual network is composed of a set of residual blocks, each of which consists a few stacked layers (e.g., convolutional layers, rectified linear unit layers and batch normalization layers). Given the ℓ-th residual block B_l, if we denote the input and output of B_l as $H_{\ell-1}$ and H_ℓ respectively, and employ $\mathcal{H}_\ell(\mathbf{x})$ to represent the underlying mapping of these stacked layers, in traditional way, we can obtain:

$$H_\ell = \mathcal{H}_\ell(H_{\ell-1}). \tag{1}$$

However, when leveraging residual learning, instead of making these stacked layers to approximate underlying mapping $\mathcal{H}_\ell(\mathbf{x})$, we want them to fit another residual mapping function $\mathcal{F}_\ell(\mathbf{x}) := \mathcal{H}_\ell(\mathbf{x}) - \mathbf{x}$. In this case, the output of this residual mapping is $H_\ell - H_{\ell-1}$ and Eq. (1) can be rewritten as:

$$H_\ell = \mathcal{F}_\ell(H_{\ell-1}) + H_{\ell-1}, \tag{2}$$

where $\mathcal{F}_\ell(\mathbf{x})$ is the residual mapping function that residual block B_l learns. Note that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping [31].

Fig. 2 (a) illustrates a typical residual block, which consists of convolutional (conv), Rectified Linear Unit (ReLU) and batch normalization (Batch Norm) layers. In practice, the operation of residual learning can be performed by shortcut connections and element-wise additions. Note that the dimensions of input and output of residual block B_ℓ should be equal when using shortcut connections. However, in most cases, we need to change the dimensions of feature maps (such as downsampling operations). In this regard, a linear projection W_s is employed to match dimensions of input and output. Specifically, the Eq. (2) can be further converted to:

$$\begin{aligned} H_\ell &= \mathcal{F}_\ell(H_{\ell-1}) + \mathrm{id}(H_{\ell-1}) \\ \mathrm{id}(\mathbf{x}) &= W_s\mathbf{x}, \end{aligned} \tag{3}$$

where id(·) represents the identity transformation and it is a linear projection. After constructing the residual block, we can build very deep networks by stacking residual blocks.

B. FCRN for Skin Lesion Segmentation

1) Fully Convolutional Residual Network: The networks proposed in [31] are designed for classification. In order to achieve accurate and efficient skin lesion segmentation, we further construct a fully convolutional residual network (FCRN) based on residual blocks, which can take an arbitrary-sized image as input and output an equal-sized prediction score mask. After successive down-sampling operations in the original residual network, the dimensions of feature maps are gradually reduced and become much smaller than that of the original input image. To bridge the resolution gap so that both the learning and inference procedures can be performed in an efficient end-to-end way, we exploit deconvolutional layers as the up-sampling operation to connect coarse predict maps and dense pixels predictions [38], [39]. Specifically, we use deconvolutional layers to upsample small prediction maps and get the equal-sized prediction maps with input images. Note that the weights within the deconvolutional layers are also trainable during the learning process.

Compared with the original residual network, the advantage of the proposed FCRN is that it can make pixel-wise predictions, which is of valuable significance for skin lesion segmentation task. In such a full convolutional architecture, we can efficiently obtain accurate segmentation mask of an input dermoscopy image with a single forward propagation in the testing procedure. More importantly, with the per-pixel-wise error back-propagation in the training procedure, each single pixel can be considered as an independent training sample. Consequently, the number of equivalent training samples is greatly boosted, which is helpful to train very deep networks with a large amount of parameters under limited training data.

By harnessing the discriminative features learned from the very deep network, the proposed FCRN can produce good prediction maps of skin lesions. However, when carefully probing the prediction maps, we find that the prediction maps neglect some detailed local information. The underlying reason of this phenomenon is that the deconvolutional layers, albeit being essential to construct an efficient end-to-end network, have large strides (32 pixels in our implementation) and hence only exploit object-level features in upper layers without sufficiently leveraging the low-level spatial information in bottom layers of the network. Actually, when the FCRN goes deeper, the size of receptive field is becoming larger and the feature maps can capture more global and abstract contextual features of skin lesions. In contrast, the features from lower layers with smaller receptive fields can reflect the local structure information of skin lesions, which is also quite important for effective segmentation but discarded by the single deconvolutional layer. In this regard, we propose to integrate multi-scale contextual information in the proposed FCRN. Specifically, the network generates several skin lesion prediction maps by employing different levels of features in FCRN, and then fuse these prediction maps with a summing operation by the following deconvolutional layers. Please refer to the shallow gray arrows and the *fusion* symbols in Fig. 2 (b). As a result, the generated prediction maps encode both global and local features of skin lesions, making the prediction more accurate and robust.

2) Network Architecture: Fig. 2 (b) shows the architecture of the FCRN. This proposed FCRN contains 16 residual blocks in down-sampling path. Each residual block consists of two 1×1 convolutional layers, one 3×3 convolutional layer, three batch normalization layers and three ReLU layers, as shown in Fig. 2 (a). Besides these residual blocks, it also contains one 7×7 convolutional layer and one 3×3 max pooling layer (both with stride 2) as prelayers.

As for the up-sampling path, we generate three kinds of prediction maps by employing different levels of features. They are 8-pixel stride prediction map, 16-pixel stride prediction map and 32-pixel stride prediction map, respectively. To do this, we add three 1×1 convolutional layers on top of the residual block 7, the residual block 13 and the residual block 16 to produce the 8-, 16- and 32-pixel stride prediction maps, respectively. In order to obtain the final prediction map, we first fuse the 32-pixel stride prediction map with the 16-pixel stride prediction map by adding a 2×up-sampling deconvolutional layer on the top of 32-pixel one. Similarly, we then add a 2×up-sampling deconvolutional layer on the top of fused 16-pixel one to fuse the new 16-pixel stride prediction map and the 8-pixel stride prediction map. Finally, the network produces the final prediction map with the same size as the input image by adding a 8×up-sampling deconvolutional layer on top of the fused 8-pixel stride prediction map. In this way, the skin lesion probability map incorporating multi-scale contextual features can be generated after the Softmax classification layer. Note that in the above fusion scheme, all of the fusions are operated by pixel-wise additions. After acquiring the skin lesion probability map, the final skin lesion masks are generated by setting a threshold T (empirically set as 0.5 in our experiments).

3) Training Procedure: To train the FCRN, we first crop an sub-image from every original dermoscopy image with ground truth by automatically figuring out the smallest rectangle containing the lesion region and enlarging its length and width by $1.1 - 1.3$ times in order to include more neighboring pixels for training. Then we randomly crop another sub-image with the same length and width on the dermoscopy image to increase the negative training samples. Note that as the objective of the FCRN is for segmentation, every pixel can be considered as a training sample. We take all the pixels on these cropped sub-images as training samples to train the FCRN. In the testing phase, we do not perform the sub-image cropping procedure or other detection-like processes. We directly segment the whole dermoscopy images and the prediction masks of the whole image were produced with an overlap-tile strategy.

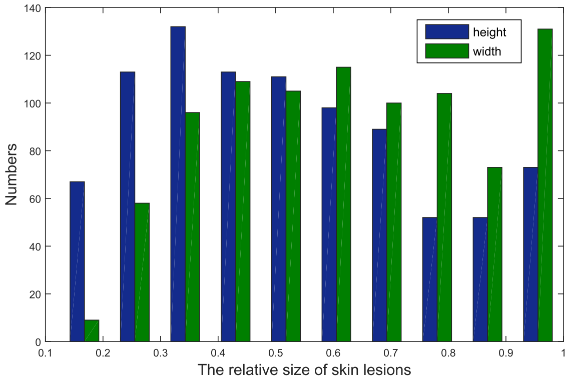


Fig. 3. Distribution of the relative size of skin lesions in our training dermoscopy images.

C. Skin Lesion Classification

1) Integration of the Two Stages: The skin lesions have large variation in size.We exploit *relative size*, the ratio between the size of smallest rectangle containing the lesion region and the size of original image, to illustrate the variation. Fig. 3 shows the distribution of the relative size of skin lesions in our training dataset (900 dermoscopy images). In this circumstance, if we directly perform skin lesion classification on original dermoscopy images, the variation of lesion size will severely influence both the training and testing performance, especially when the number of training samples is quite limited. A straightforward way to solve this problem is employing multi-scale models, where we can train several networks with different scales and then fuse the prediction results. However, this scheme is not optimal in our case, as the range of the relative size is quite wide (from 0.15 to 0.95) and there are no dominant values in the distribution, making it difficult to select appropriate scales to design multi-scale models. In this regard, we propose to first segment the skin lesions from dermoscopy images and then resize the segmented lesions into a fixed size. Finally, we conduct the classification on the post-processed lesions. Compared with training the classification network directly on original dermoscopy images, training it on segmented results can help it extract more representative features specific to the lesion for better recognition, especially under limited training data.

2) Classification Network Architecture: We construct a very deep residual network to classify skin lesions based on the segmentation results. The architecture of the network is almost the same with that of the down-sampling path of the proposed FCRN for segmentation, as shown in Fig. 2 (b). The difference is that we add a 7×7 average pooling layer followed by the $16th$ residual blocks to extract the global deep residual features. Some researchers found a small but consistent advantage of replacing the Softmax layer with a linear support vector machine layer when training CNNs [40]. Inspired by these findings, we exploit two classifiers, Softmax classifier and support vector machine (SVM) classifier, to obtain two predictions and then average them to get the final results. Note that the Softmax classifier and the SVM classifier are trained independently in an end-to-end way with the proposed residual networks.

3) Training Procedure: When training the classification network, we automatically crop an image patch containing the whole skin lesion from each dermoscopy image and then resize these image patches into a fixed size (250 × 250 in our implementation). When testing, we also automatically crop the image patch containing the whole segmented skin lesion and feed the resized image patch into classification networks. To increase robustness and reduce overfitting, we further utilized the strategy of data augmentation to enlarge the training dataset. The augmentation operators include rotation (90, 180 and 270 degrees), translation and adding random noise into cropped image patches.

III. Experiments and Results

A. Dataset

We performed extensive experiments to evaluate our method on a public challenge dataset of *Skin Lesion Analysis Towards Melanoma Detection* released with ISBI 2016 [41]. This dataset is based on the International Skin Imaging Collaboration (ISIC) Archive,[1] which is the largest publicly available collection of quality controlled dermoscopic images of skin lesions. The challenge employs a subset of representative images with 900 images as training data and 350 images as testing data. The ground truth is held out by the organizer for independent evaluation. After releasing the challenge result, the organizer released the ground truth to encourage further investigations. In this case, we can perform extensive experiments to comprehensively evaluate our method. In this section, we present the challenge results and ranking provided by the organizer as well as other extensive experiment results conducted by ourselves.

B. System Implementation

The proposed method was implemented with C++ and Matlab based on Caffe library [42] on a computer equipped with a NVIDIA TITAN X GPU. The networks were trained with Stochastic gradient descent (SGD) method (we set batch size as 4, momentum as 0.9, weight decay as 0.0005, the learning rate as 0.001 initially and reduced it by a factor of 10 every 3000 iterations). We adopted batch normalization (BN) [36] right after each convolutional layer to accelerate the training speed except the three convolutional layers for generating prediction scores. Specifically, according to the recommendation configuration of Caffe library,[2] we added the *BatchNormLayer* and *ScaleLayer* after each convolutional layer and froze the learning parameters of the *BatchNormLayer*. In order to enhance the training efficiency, we used a pretrained model (trained on ImageNet dataset) [31] to initialize the weights of both the segmentation and classification networks. For the deconvolutional layers equipped for the FCRN, we initialized them with bilinear interpolation weights and set the learning rate of these deconvolutional layers 1/10 of other layers, as we only need to slightly tune the weights of deconvolution layers to obtain satisfactory results. With the inference of fully convolutional architecture, our method was very efficient; it averagely took 0.84s to process one dermoscopy image with size of 1024×768 (0.52s for segmentation and 0.32s for classification). In order to encourage other researchers to further investigate the potential of deep residual networks on medical image analysis tasks, we released our implementation and network architectures in the project website: http://www.cse.cuhk.edu.hk/~lqyu/skin/.

C. Evaluation Metrics

We applied the challenge evaluation metrics to evaluate both the segmentation and classification performance of our method. For the segmentation, the evaluation criteria include sensitivity (SE), specificity (SP), accuracy (AC), Jaccard index (JA) and Dice coefficient (DI). The organizer first calculated these criteria for each test dermoscopy image and then averaged each criterion on the whole testing dataset to get the final results. The criteria are defined as:

$$\begin{aligned} AC &= \frac{N_{tp} + N_{tn}}{N_{tp} + N_{fp} + N_{fn} + N_{tn}}, \\ SE &= \frac{N_{tp}}{N_{tp} + N_{fn}}, \quad SP = \frac{N_{tn}}{N_{tn} + N_{fp}}, \\ JA &= \frac{N_{tp}}{N_{tp} + N_{fn} + N_{fp}}, \quad DI = \frac{2 \cdot N_{tp}}{2 \cdot N_{tp} + N_{fn} + N_{fp}}, \end{aligned} \tag{4}$$

where N_{tp}, N_{tn}, N_{fp} and N_{fn} denote the number of true positive, true negative, false positive and false negative, respectively, and they are all defined on the pixel level. A lesion pixel is considered as a true positive if its prediction is lesion; otherwise it is regarded as a false negative. A non-lesion pixel is considered as a true negative if its prediction is non-lesion; otherwise it is regarded as a false positive. Participants are ranked based on the results of JA, as it is generally considered as the most important criterion for segmentation.

As for the classification, there are four evaluation criteria, including sensitivity (SE), specificity (SP), accuracy (AC) and average precision (AP). The definition of SE, SP and AC is the same as the metrics for segmentation, but here they are measured at image level instead of pixel level. The detailed definition of AP can be found in [41]. In the classification task, the numbers of melanoma and non-melanoma lesions in testing dataset are quite imbalanced (melanoma/non-melanoma = 75/304). In this case, the false positive rate should be relatively small and the true negative rate should be relatively large, resulting in most points fall in the left part of the receiver operating characteristic (ROC) curve. Therefore,

[1] https://isic-archive.com

[2] http://caffe.berkeleyvision.org/doxygen/classcaffe_1_1BatchNormLayer.html\#details

TABLE I
ARCHITECTURES OF DOWN-SAMPLING PATH IN FCRN-38, −50 AND −101

38-layer	50-layer	101-layer
7×7, 64, stride 2		
3×3 max pool, stride 2 ResBlock 1-3		
ResBlock 4-7	ResBlock 4-7	ResBlock 4-7
ResBlock 8-9	ResBlock 8-13	ResBlock 8-30
ResBlock 10-12	ResBlock 14-16	ResBlock 31-33

TABLE II
COMPARISON OF ARCHITECTURES WITH DIFFERENT DEPTHS

Networks	AC	DI	JA	SE	SP
VGG-16	0.903	0.794	0.707	0.796	0.945
GoogleNet	0.916	0.848	0.776	0.901	0.916
FCRN-38	0.929	0.856	0.785	0.882	0.932
FCRN-50	**0.949**	**0.897**	**0.829**	**0.911**	**0.957**
FCRN-101	0.937	0.872	0.803	0.903	0.935

TABLE III
COMPARISON OF FCRNs WITH DIFFERENT SCHEMES OF MULTI-SCALE CONTEXTUAL FEATURE INTEGRATION

Network	AC	DI	JA	SE	SP
FCRN8	**0.949**	**0.897**	**0.829**	**0.911**	**0.957**
FCRN16	0.930	0.855	0.785	0.882	0.934
FCRN32	0.923	0.848	0.773	0.899	0.914

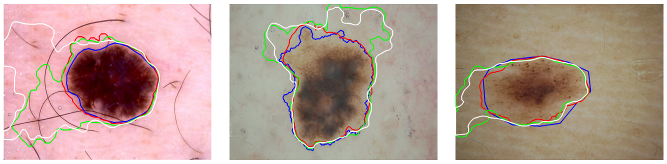


Fig. 4. The segmentation results of FCRN8, FCRN16 and FCRN32. The blue, red, green and white contours indicate the segmentation results of ground truth, FCRN8, FCRN16 and FCRN32, respectively.

the area under the ROC curve (AUC) has a high risk of being noisy in this task. In this case, the organizer employed the AP to rank the participants and measured AUC as a reference in the challenge [41].

D. The Performance of Our Method in Segmentation

1) Experiments on Segmentation Network Depth: In order to investigate if the increase of network depth can enhance the discrimination capability of convolutional networks and thus make them better deal with the challenges of melanoma recognition, we compared the performance of the proposed FCRN with different depths (38, 50, and 101 layers, respectively), fully convolutional VGG-16 network [29] and fully convolutional GoogleNet [30]. Table I illustrates network architectures of the down-sampling path in FCRNs with different layers. All of the above five networks harnessed the same up-sampling strategy and multi-scale contextual information integration scheme. The difference is the different architectures of the down-sampling path. As for the three FCRNs, the difference is the different residual block numbers at each scale. Not that for all the FCRNs with 38 layers, 50 layers and 101 layers, we performed extensive experiments with different architectures (maintain the total number of layers while adjusting the number of residual blocks at different scales) and selected the architectures with best results for comparison. The results are listed in Table II.

It is observed that the FCRNs with 38 layers, 50 layers and 101 layers all achieve better performance in all five metrics than the 16-layer VGG and 22-layer GoogleNet, except that the SP of the 38-layer and 101-layer FCRN is lower than that of the 16-layer VGG and the SE of the 38-layer FCRN is lower than that of 22-layer GoogleNet, demonstrating the increase of network depth can effectively enhance the discrimination capability of convolutional networks. The better performance of 50-layer FCRN than 38-layer FCRN further verifies that network depth is a key factor of model expressiveness and we can get better performance with a deeper network. However, when the FCRN goes as deep as 101 layers, its performance is worse than that of the 50-layer FCRN. One of the underlying reasons of this phenomenon may be that the 101-layer FCRN has about two times as much parameters as the 50-layer FCRN, and it is difficult to effectively train such a deep network with so many parameters using the limited training data that we can acquire in our application. Nevertheless, the 101-layer FCRN achieves better performance than 38-layer FCRN and 16-layer VGG, demonstrating the residual learning technique can contribute the training process of such a deep network with 101 layers. Note that as the insufficiency of training data is a common problem in many medical image analysis applications, we should carefully study the trade-off between network depth and network performance under such a situation for each application. In the following experiments, we employ our 50-layer FCRN for comparisons as it exhibits a good balance between network depth and performance. Note that such a network is still substantially deeper than existing networks in medical image analysis field.

2) Experiments on the Multi-Scale Contextual Integration Scheme: We further conducted a set of experiments to demonstrate the importance of the multi-scale integration scheme equipped for the proposed FCRN on avoiding the disregard of local information in the fully convolutional architecture. We constructed three variants of the proposed FCRN for the experiments. The first one only utilized the 32-pixel stride predictions and we referred it as FCRN32. The second one fused the 32-pixel stride predictions and the 16-pixel stride predictions and we referred it as FCRN16. The third one fused the 32-, 16-, and 8-pixel stride predictions and we referred it as FCRN8. Table III shows the results of these variants of FCRN. It is observed that the FCRN8 achieves the best results among these three variants on all the five metrics, demonstrating the effectiveness of the proposed multi-scale integration scheme, which can be easily extended to other semantic segmentation tasks. Fig. 4 further provides some typical examples to demonstrate the advantage of the multi-scale integration scheme, where the segmentation results of FCRN8 are much better than other two networks.

3) Qualitative Evaluation: Fig. 5 shows some segmentation results of the proposed FCRN on some challenging cases of both melanoma and non-melanoma lesions, including cases with low contrast (Fig. 5 (a), (b), (c) and (h)), cases with irregular shapes (Fig. 5 (e), (f), (g)), and cases with severe artifacts (Fig. 5 (g) and (h)). Our method achieves satisfactory results in all of these challenging cases, demonstrating the very deep network with effective learning mechanisms is a promising way to meet the challenges of skin lesion analysis.

4) Comparison with Other Methods in the Challenge: There were totally 28 teams participating the ISBI challenge of skin lesion segmentation and the challenge results (only top ten teams) are listed in Table IV. Note that each team was only allowed for one submission and the teams were ranked according to the Jaccard index (JA) metric. Our method ranked the second among the 28 teams. In fact, most participants in the top ten employed CNNs to perform the segmentation, demonstrating the popularity, as well as performance gains, of CNNs in skin lesion analysis, surpassing traditional methods based on hand-crafted features. However, as we know, most of them exploited AlexNet [24], VGG-16 [29] or other shallower networks for this challenging task, whereas we leveraged a substantially deeper network with 50 layers. Experimental results showed that our very deep network outperformed most of our shallower counterparts, demonstrating the discrimination capability gained from substantially increasing the network depth. The EXB team was the only team that achieved better results than our method. This may be because this team focused on the segmentation task and employed some pre- and post- processing schemes to refine the results [43], whereas the main aim of our FCRN was to provide a good basis for the following recognition task and we did not employ any compelling refinement step.

TABLE IV
RESULTS OF SKIN LESION SEGMENTATION CHALLENGE ON ISBI 2016

Method	AC	DI	JA	SE	SP
EXB	**0.953**	**0.910**	**0.843**	0.910	0.965
CUMED (ours)	0.949	0.897	0.829	0.911	0.957
Mahmudur	0.952	0.895	0.822	0.880	0.969
SFU-mial	0.944	0.885	0.811	**0.915**	0.955
TMUteam	0.946	0.888	0.810	0.832	**0.987**
UiT-Seg	0.939	0.881	0.806	0.863	0.974
IHPC-CS	0.938	0.879	0.799	0.910	0.941
UNIST	0.940	0.867	0.797	0.876	0.954
Jose Luis	0.934	0.869	0.791	0.870	0.978
Marco romelli	0.936	0.864	0.786	0.883	0.962

Note: There are totally 28 submissions. The top 10 entries are shown here and the ranking (from top to bottom) was made according to the JA.

TABLE V
RESULTS OF CLASSIFICATION WITH AND WITHOUT SEGMENTATION

Method	AC	AUC	AP	SE	SP
Without segmentation	0.828	0.782	0.560	0.427	0.927
With segmentation	**0.855**	**0.783**	**0.624**	**0.547**	**0.931**

TABLE VI
COMPARISON OF DIFFERENT NETWORKS ON THE CLASSIFICATION TASK

Method	AC	AUC	AP	SE	SP
VGG-16	0.826	**0.826**	0.529	0.413	0.928
GoogleNet	0.847	0.801	0.581	0.507	**0.931**
DRN-50	**0.855**	0.783	**0.624**	**0.547**	**0.931**

E. The Performance of Our Method in Classification

1) Classification With and Without Segmentation: We employed a two-stage framework for automated melanoma classification and recognition. In order to validate the necessity of the two-stage scheme, we compared the classification performance of our very deep residual network with and without the segmentation stage. Both of these two experiments were with the same network architecture, training parameters and data augmentation strategies. Table V lists the experimental results. It is observed that the two-stage scheme achieves much better results than directly employing the very deep residual network on the original dermoscopy images without segmentation with 11.4% relative improvement on the AP metric (the official ranking metric). This is because the variation of lesion size is very large (see Fig. 3). Training the classification network based on the segmentation results instead of the original dermoscopy images can effectively prevent the learning process being distracted by other structures and artifacts in images and hence can generate more discriminative features for better recognition. It is worthwhile to point out that the segmentation and classification stages are seamlessly integrated in our framework and the whole recognition process is performed in an automated way without any manual interactions.

2) Experiments on Classification Network Depth: We also investigated if the increase of network depth can enhance the discrimination capability of convolutional networks on the classification task. Table VI lists the classification performance of our very deep networks with 50 layers, VGG-16 network [29] and GoogleNet [30] based on segmentation ground truth. As we can see, the 50-layers residual network gets the best performance on the AP metric than 16-layers VGG network and 22-layers GoogleNet, which demonstrates that increasing network depth can also improve the discrimination capability of networks on the classification task.

3) Experiments of Model Fusion: In our skin lesion classification stage, we trained two networks with different Softmax and SVM classifiers in an end-to-end way and we found the simple average fusion can further improve the classification performance. Table VII lists the skin lesion classification performance for different classifiers in the testing dataset. Noting that in order to better verify the performance gain of this fusion model, we cropped the skin lesion regions as the segmentation ground truth not as our segmentation results. We have observed that the fusion model achieved better performance on AP (the ranking metric), AC and SE metrics, which demonstrates the effectiveness of this simple fusion scheme.

4) Quantitative Evaluation and Comparison With Other Methods: We participated the challenge of skin lesion classification without providing any segmentation results. For every input image, we first utilized the proposed FCRN

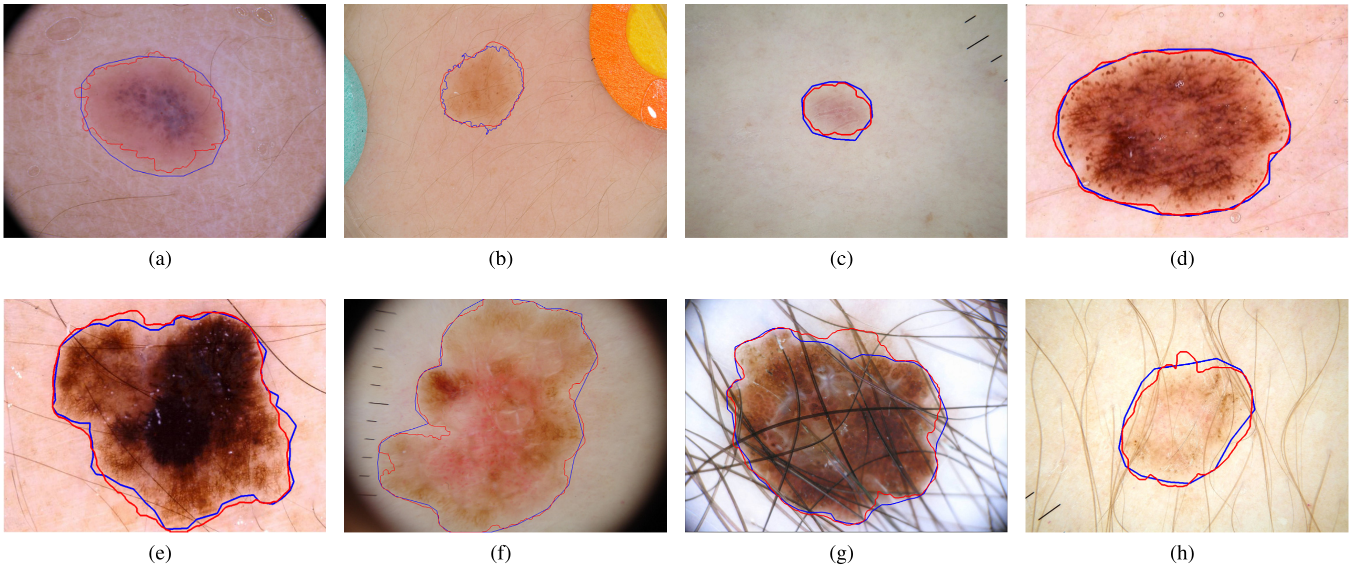


Fig. 5. Examples of some skin lesion segmentation results from test images. The first and second rows are non-melanoma and melanoma lesions, respectively. The red and blue contours indicate the segmentation results of our method and ground truth, respectively.

TABLE VII
RESULTS OF SKIN LESION CLASSIFICATION FOR DIFFERENT CLASSIFIERS

Method	AC	AUC	AP	SE	SP
Softmax	0.850	**0.790**	0.609	0.507	**0.934**
SVM	0.844	0.779	0.616	0.520	0.824
Fusion	**0.855**	0.783	**0.624**	**0.547**	0.931

TABLE VIII
RESULTS OF SKIN LESION CLASSIFICATION CHALLENGE ON ISBI 2016

Method	AC	AUC	AP	SE	SP
CUMED (ours)	**0.855**	0.804	**0.637**	0.507	0.941
GTDL	0.813	0.802	0.619	0.573	0.872
BF-TB	0.834	**0.826**	0.598	0.320	0.961
ThrunLab	0.786	0.796	0.563	0.667	0.816
Jordan Yap	0.844	0.775	0.559	0.240	**0.993**
Haebeom Lee	0.821	0.793	0.555	0.200	0.974
GT-DL1	0.815	0.813	0.552	0.467	0.901
GT-DL2	0.681	0.800	0.545	0.787	0.655
Sebastien PARIS	0.731	0.793	0.542	0.773	0.720
USYD-BMIT	0.599	0.780	0.537	**0.853**	0.536

Note: There are totally 25 submissions. The top 10 entries are shown here and the ranking was made according to the AP scores.

model to obtain the skin lesion segmentation results and then harnessed the very deep residual classification network to produce the possibilities of melanoma. Note that, albeit having two steps, our method produces the results in an automated way. There were totally 25 teams submitting their results for this challenge. The results were evaluated by using the above-mentioned metrics and the teams were ranked according to the average precision (AP). We list the top ten results in Table VIII. We rank *the first place* in the challenge with the AP value 0.637, demonstrating the advantages of the proposed method in dealing with the challenges of the skin lesion recognition. Actually, our network outperforms most of its shallower counterparts by a large margin, which evidences that increasing network depth with effective learning mechanism can improve the discrimination capability of CNNs for challenging medical image analysis tasks, even if the training data are limited. Both the segmentation and classification results demonstrate the effectiveness of our very deep residual networks for automated skin lesion analysis.

IV. DISCUSSION

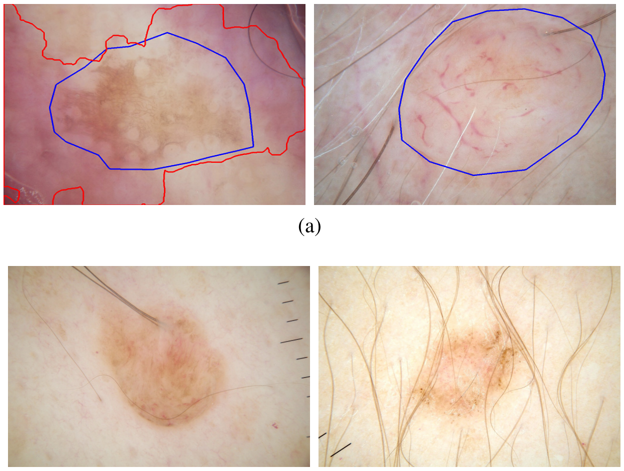
While recent years have witnessed the remarkable success of deep convolutional neural networks in medical image analysis tasks, there still exists a gap between manual assessment of experts and automated evaluation of computers in many clinical applications where the targeting objects have large intraclass variation and small interclass variation [17], [44]. One encouraging news from recent studies in computer vision field is that we still have much room to exploit both the network width [32], [45] and depth [29]–[31] of CNNs to improve their performance. In this paper, we focus on tapping the potential of *network depth* and endeavor to investigate if very deep CNNs are capable of dealing with complicated medical image analysis tasks and achieving more performance gains than their shallower counterparts. To the best of our knowledge, we are not aware of any previous work that explores or verifies the efficiency of very deep networks in medical image analysis field.

Although network depth has been proved to be a major determinant of model expressiveness, both in the theory [27], [28] for a long time and in practice [29], [30] recently, it is difficult to train an effective very deep network because of the degradation problems, which will become more and more severe when a network goes deeper. For medical image analysis tasks, besides the above problems, another main obstacle to hinder the application of a very deep CNN is the limited quality training dataset. In this sense, whether very deep CNN is beneficial to challenging medical applications with limited training data is still an open problem and worthwhile to be explored and verified. We, aiming at improving the performance of automated melanoma recognition in dermoscopy images, propose a novel two-stage framework based on very deep CNNs by leveraging a set of effective training schemes. It is worthwhile to point out that these training schemes are general enough that can be easily extended to other medical image analysis tasks sharing the similar challenges of skin lesion analysis.

One of issues of our framework is that if we can share the weights between the segmentation and classification networks. We did not share the weights in our implementation and we think it is not beneficial to do that. The weights of segmentation and classification networks are very similar since the two network weights are both initialized by the pretrained model on ImageNet [31]. However, the two networks should not share the completely same weights. The segmentation task was performed in the original scale and we did not adopt image resize operations, whereas we resized the image patches into a fixed size (250 × 250) when performing classification (we zoomed out the images in most of cases). Therefore, the segmentation and classification networks were performed in different image scales and should adopt different features although their weights can be similar.

One of the main concerns of employing deep CNNs in medical image analysis tasks is the insufficiency of quality training data. While the techniques employed in this work can effectively alleviate this problem, we still encounter performance degradation when the network goes deeply to more than 100 layers (see the results reported in Table II). Compared with the natural image processing tasks, which usually have millions of training samples (e.g., ImageNet dataset [26] having 1.2 million images for classification and MS COCO dataset [46] having about 120K images for segmentation) to support networks with hundreds of layers [31], [47], we face the difficulty in fully exploiting the discrimination capability gains of very deep CNNs under the circumstance of limited training data. In addition, while the proposed framework can achieve satisfactory results for both segmentation and classification in most cases, there are still some failure cases, as shown in Fig. 6. Fig. 6 (a) shows two failure cases of segmentation, whereas Fig. 6 (b) shows some failure cases of classification. It is observed that most of these failure cases have low contrast, irregular shapes and artifacts around the lesions. In the future, we shall investigate to integrate Bayesian learning, especially probabilistic graphical models [48], [49] into our networks to further enhance the discrimination capability of the very deep CNNs to tackle the limited training data problem. On the other hand, the segmented results generated by the proposed FCRN provide us a good basis for combining hand-crafted clinical features and features learned from CNNs to further improve the recognition performance.

Although this work, to the best of our knowledge, is the first to apply very deep CNNs to solve a complicated medical image analysis problem, i.e., automated melanoma recognition, we believe very deep CNNs can find more and more applications in medical domain. The techniques exploited in this work may inspire more studies on how to take full advantages of network depth to break the performance bottleneck of many other complex medical image analysis problems.


(a)

(b)

Fig. 6. Some failure cases of our framework: (a) failure cases of segmentation, where red and blue contours indicate the segmentation results of our method and the ground truth, respectively, and (b) melanoma lesions neglected by our framework.

V. Conclusion

In this paper, we propose a novel method based on very deep CNNs to meet the challenges of automated melanoma recognition in dermoscopy images, which consists of two steps: segmentation and classification. We seamlessly connect the two steps and form an automated framework without need of manual interaction. Compared with much shallower counterparts, the very deep CNNs can generate features with high discrimination capability, and hence improve the performance of both segmentation and classification tasks. We further construct a novel FCRN incorporating a multi-scale contextual information integration scheme for accurate skin lesion segmentation. Extensive experiments conducted on the open challenge dataset of *Skin Lesion Analysis Towards Melanoma Detection* in ISBI 2016 demonstrated the effectiveness of the proposed method. Our study corroborates that very deep CNNs with effective training mechanisms can be employed to solve complicated medical image analysis problems, even with limited training data. Further investigations include integrating probabilistic graphical models into our networks to further enhance the discrimination capability and exploring our method on more applications.

References

[1] A. F. Jerant *et al.*, “Early detection and treatment of skin cancer,” *Amer. Family Phys.*, vol. 62, no. 2, pp. 357–386, 2000.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2016,” *CA, Cancer J. Clinicians*, 2015.

[3] K. A. Freedberg, A. C. Geller, D. R. Miller, R. A. Lew, and H. K. Koh, “Screening for malignant melanoma: A cost-effectiveness analysis,” *J. Amer. Acad. Dermatol.*, vol. 41, no. 5, pp. 738–745, 1999.

[4] C. M. Balch *et al.*, "Final version of the American joint committee on cancer staging system for cutaneous melanoma," *J. Clin. Oncol.*, vol. 19, no. 16, pp. 3635–3648, 2001.

[5] M. Binder *et al.*, "Epiluminescence microscopy: A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists," *Arch. Dermatol.*, vol. 131, no. 3, pp. 286–291, 1995.

[6] M. Silveira *et al.*, "Comparison of segmentation methods for melanoma diagnosis in dermoscopy images," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 1, pp. 35–45, Feb. 2009.

[7] N. K. Mishra and M. E. Celebi. (Jan. 2016). "An overview of melanoma detection in dermoscopy images using image processing and machine learning." [Online]. Available: https://arxiv.org/abs/1601.07843

[8] R. J. Stanley, W. V. Stoecker, and R. H. Moss, "A relative color approach to color discrimination for malignant melanoma detection in dermoscopy images," *Skin Res. Technol.*, vol. 13, no. 1, pp. 62–72, 2007.

[9] Y. I. Cheng *et al.*, "Skin lesion classification using relative color features," *Skin Res. Technol.*, vol. 14, no. 1, pp. 53–64, 2008.

[10] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*. New York, NY, USA: Springer, 2013, pp. 63–86.

[11] T. Tommasi, E. La Torre, and B. Caputo, "Melanoma recognition using representative and discriminative kernel classifiers," in *Proc. Int. Workshop Comput. Vis. Approaches Med. Image Anal.*, 2006, pp. 1–12.

[12] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 233–239, Mar. 2001.

[13] M. E. Celebi *et al.*, "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 31, no. 6, pp. 362–373, 2007.

[14] G. Schaefer, B. Krawczyk, M. E. Celebi, and H. Iyatomi, "An ensemble classification approach for melanoma diagnosis," *Memetic Comput.*, vol. 6, no. 4, pp. 233–240, 2014.

[15] A. A. A. Setio *et al.*, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.

[16] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207–1216, May 2016.

[17] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[18] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[19] H. R. Roth *et al.*, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. New York, NY, USA: Springer, 2014, pp. 520–527.

[20] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.

[21] 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.

[22] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "Dcan: Deep contour-aware networks for accurate gland segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2016.

[23] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," in *Machine Learning in Medical Imaging*. New York, NY, USA: Springer,, 2015, pp. 118–126.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[25] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Aug. 2016.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[27] J. Håstad and M. Goldmann, "On the power of small-depth threshold circuits," *Comput. Complex.*, vol. 1, no. 2, pp. 113–129, 1991.

[28] J. Håstad, "Computational limitations of small-depth circuits," Tech. Rep., 1987.

[29] K. Simonyan and A. Zisserman. (Apr. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[30] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[31] K. He, X. Zhang, S. Ren, and J. Sun. (Dec. 2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385

[32] S. Zagoruyko and N. Komodakis. (Nov. 2016). "Wide residual networks." [Online]. Available: https://arxiv.org/abs/1605.07146

[33] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[34] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[35] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. (2014). "Deeply-supervised nets." [Online]. Available: https://arxiv.org/abs/1409.5185

[36] S. Ioffe and C. Szegedy. (Feb. 2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167

[37] K. He, X. Zhang, S. Ren, and J. Sun. (Mar. 2016). "Identity mappings in deep residual networks." [Online]. Available: https://arxiv.org/abs/1603.05027

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[39] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1395–1403.

[40] Y. Tang. (Jun. 2013). "Deep learning using linear support vector machines." [Online]. Available: https://arxiv.org/abs/1306.0239

[41] D. Gutman *et al.* (May 2016). "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)." [Online]. Available: https://arxiv.org/abs/1605.01397

[42] Y. Jia *et al.* (Jun. 2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[43] K. Sirinukunwattana *et al.* (Mar. 2016). "Gland segmentation in colon histology images: The GlaS challenge contest." [Online]. Available: https://arxiv.org/abs/1603.00275

[44] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016.

[45] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 473–480.

[46] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV*. New York, NY, USA: Springer, 2014, pp. 740–755.

[47] J. Dai, K. He, and J. Sun. (Dec. 2015). "Instance-aware semantic segmentation via multi-task network cascades." [Online]. Available: https://arxiv.org/abs/1512.04412

[48] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1529–1537.

[49] H. Wang and D.-Y. Yeung. (Apr. 2016). "Towards Bayesian deep learning: A survey." [Online]. Available: https://arxiv.org/abs/1604.01662