Dermoscopic Image Segmentation via Multi-Stage Fully Convolutional Networks

Lei Bi, *Student Member, IEEE*, Jinman Kim*, *Member, IEEE*, Euijoon Ahn, *Student Member, IEEE*, Ashnil Kumar, *Member, IEEE*, Michael Fulham, and Dagan Feng, *Fellow, IEEE*

***Abstract*— *Objective:* Segmentation of skin lesions is an important step in the automated computer aided diagnosis (CAD) of melanoma. However, existing segmentation methods have a tendency to over- or under-segment the lesions and perform poorly when the lesions have fuzzy boundaries, low contrast with the background, inhomogeneous textures, or contain artifacts. Furthermore, the performance of these methods are heavily reliant on the appropriate tuning of a large number of parameters as well as the use of effective pre-processing techniques such as illumination correction and hair removal. *Methods:* We propose to leverage fully convolutional networks (FCNs) to automatically segment the skin lesions. FCNs are a neural network architecture that achieves object detection by hierarchically combining low-level appearance information with high-level semantic information. We address the issue of FCN producing coarse segmentation boundaries for challenging skin lesions (e.g., those with fuzzy boundaries and/or low difference in the textures between the foreground and the background) through a multi-stage segmentation approach in which multiple FCNs learn complementary visual characteristics of different skin lesions; early-stage FCNs learn coarse appearance and localization information while late-stage FCNs learn the subtle characteristics of the lesion boundaries. We also introduce a new parallel integration method to combine the complementary information derived from individual segmentation stages to achieve a final segmentation result that has accurate localization and well-defined lesion boundaries, even for the most challenging skin lesions. *Results:* We achieved an average Dice coefficient of 91.18% on the ISBI 2016 Skin Lesion Challenge dataset and 90.66% on the PH2 dataset. *Conclusion and Significance:* Our extensive experimental results on two well-established public benchmark datasets demonstrate that our method is more effective than other state-of-the-art methods for skin lesion segmentation.**

***Index Terms*— Segmentation, Melanoma, Dermoscopic, Fully Convolutional Networks (FCN)**

L. Bi, J. Kim*, E. Ahn, A. Kumar, M. Fulham and D. Feng are with the School of Information Technologies, The University of Sydney, Australia. (*corresponding author, e-mail: jinman.kim@sydney.edu.au).

M. Fulham is also with the Department of Molecular Imaging, Royal Prince Alfred Hospital, Australia and Sydney Medical School, The University of Sydney, Australia.

D. Feng is also with the Med-X Research Institute, Shanghai Jiao Tong University, China.

This work was supported in part by Australia Research Council (ARC) grants.

I. Introduction

Malignant melanoma has one of the most rapidly increasing incidences in the world and has a considerable mortality rate [1]. Early diagnosis is particularly important since melanoma can be cured with prompt excision [2]. Dermoscopy is a non-invasive dermatology imaging technique for the in vivo observation of pigmented skin lesions [3] and plays an important role in the early diagnosis of malignant melanoma [2]. It uses optical magnification and either liquid immersion and low angle-of-incidence lighting or cross-polarized lighting to make the contact area translucent, increasing the visibility of subsurface structures when compared to conventional clinical images. Identifying melanoma in dermoscopic images using human vision alone can be inaccurate, subjective, and irreproducible even among experienced dermatologists [4, 5]. This is attributed to the challenges in interpreting images with diverse characteristics (Figure 1) including lesions of varying sizes and shapes, lesions that may have fuzzy boundaries, different skin colors, and the presence of hair [4, 6]. These significant challenges have motivated the development of computer aided diagnosis (CAD) systems that can assist the dermatologists' clinical diagnosis [7-9].

A. Related Work

Lesion segmentation is a fundamental requirement for melanoma CAD. A number of segmentation methods have been recently proposed to segment skin lesions, divided across three main categories: (1) semi-automatic, which attempts to segment the skin lesions in an interactive manner; (2) un-supervised fully automatic, which attempts to segment the skin lesions automatically without using training data; and (3) supervised fully automatic, which attempts to segment the skin lesions automatically using trained classifiers. For a more detailed discussion of the field, readers can refer to the two comprehensive skin lesion segmentation survey papers written by Celebi et al. [10, 11].

Semi-automatic methods require user initialization of the segmentation process, such as through seed selection [12] or contour placement [12, 13]. These seeds and contours can then be grown or morphed to the skin lesion boundaries according to predefined functions. However, the manual initializations are usually subjective, time-consuming, and non-reproducible. As a consequence, such methods are unreliable for wide adoption in clinical environments.

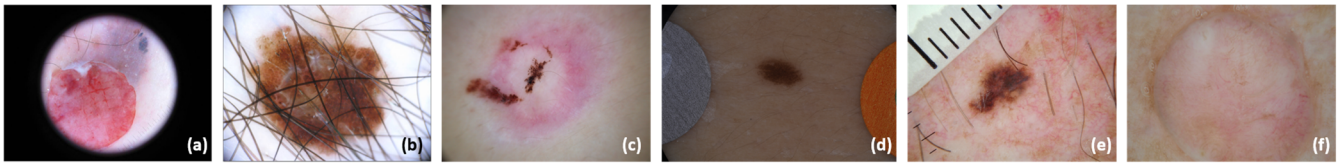

Figure 1. Examples of skin lesions with various visual characteristic such as fuzzy boundaries (a, b, c, e), presence of hair (b, e), inhomogeneity (a, c) and low-contrast to the background (d, f) which adds complexity to automated image analysis.

Unsupervised fully-automatic skin lesion segmentation methods mainly focus on thresholding [12, 14-18], energy functions [19-21] and iterative/statistical region merging [22, 23]. Thresholding methods attempt to separate the skin lesions based on a threshold value, which is generally calculated by analyzing pre-defined image features e.g., intensity histogram. Methods based on energy functions attempt to identify skin lesion boundaries by minimizing a well-defined cost (energy) function defined on image characteristics such as edges, smoothness, and statistical distributions. Iterative/statistical region merging based methods recursively merge pixels or regions together in a hierarchical manner. More recently, saliency [24], multi-scale superpixel with cellular automata (MSCA) [25], sparse coding with dynamic rule based refinement (SCDRR) [26] and delaunay triangulation [27] have also been applied for skin lesion segmentation. However, unsupervised methods have a limited capacity to accurately segment challenging skin lesions, such as lesions that touch the image boundary and those with artifacts nearby (Figure 1d and Figure 1e). Thresholding based methods are further limited by the intensity distribution of the lesion and may fail if the distribution contains multiple peaks (e.g., inhomogeneous lesions, Figure 1a).

There have been a limited number of studies that investigated the segmentation of skin lesions in a fully automatic supervised manner. These methods usually extract pixel or region features such as pixel-level Gaussian features [28, 29], RGB color features [30] and texture features [31] and then use various classifiers, such as Bayes classifier [29], wavelet network [30] or support vector machines [31], to separate the skin lesions from the surrounding healthy skin. However, all these methods rely on using low-level features, such as color and texture features, which do not capture image-wide variations. In addition, their performance depends heavily on correctly tuning a large number of parameters and effective pre-processing techniques such as illumination correction and hair removal, which thereby restricting its generalizability.

Deep learning methods based on convolutional neural networks (CNN) have recently achieved great success in image classification, object detection and segmentation problems [32-34]. This success is primarily attributed to the capability of CNNs to learn image feature representations that carry a high-level of semantic meaning [35, 36]. More recently Yu et al. [37] used a 50-layer deep residual network for segmentation, where the residual blocks proposed by He et al. [38] were used to increase the overall depth of the networks (number of layers) and enable segmentation based upon more meaningful image features. However, this resulted in much lower segmentation performance for melanoma studies, which are usually more challenging for segmentation due to severe inhomogeneity.

B. Our Contribution

In order to overcome the challenges in skin lesion segmentation and the limitations of the existing methods, we propose a new automatic skin lesion segmentation method for dermoscopic images, which we have named multi-stage fully convolution networks (FCN) with parallel integration (mFCN-PI). Our method is based on the state-of-the-art object detection method of FCN [32] adapted for skin lesion segmentation with multiple key improvements. The great success of FCN on object detection is primarily attributed to the capability of FCN to capture feature representations that contain a high-level of semantic information. While FCN may be generally applied to skin lesion segmentation, it is unable to accurately delineate the skin lesion boundary and produces inconsistent outcomes for challenging inhomogeneous skin lesions, which results in coarse and noisy segmentation results. In contrast, our proposed method learns and refine the skin lesion segmentation results across multiple stages, and then integrates these complementary multi-stage segmentation results. In addition, our method behaves like an ensemble of many deep learners, where each deep learner in the ensemble learns distinct additional information and their fusion allows capabilities that may not be captured via a single deep learner. Once trained, our method provides end-to-end segmentation at interference time with no pre- or post-processing required. When compared with FCN and all other skin lesion segmentation methods mentioned above, our method introduced the following contributions:

Our method harnesses high-level semantic information with multi-stage learning in an end-to-end way for effective skin lesion segmentation. Leveraging the fully convolutional networks (FCN) [32], our method can take an image of arbitrary size and output the segmentation label directly without any pre-processing, e.g., illumination correction, filtering and hair removal, or manual intervention, e.g., seed selections, contour placement. Hence, our method is particularly effective on images with complex artifacts that are difficult to pre-process.

We propose a multi-stage FCN (mFCN) approach to train and to predict the segmentation in multiple stages and therefore minimized the segmentation errors for challenging skin lesions. During training, mFCN is capable of learning from both the training data (images and the manual annotations) and the estimated results derived from the previous mFCN stage. The ability to learn from previous stages has the advantage in not

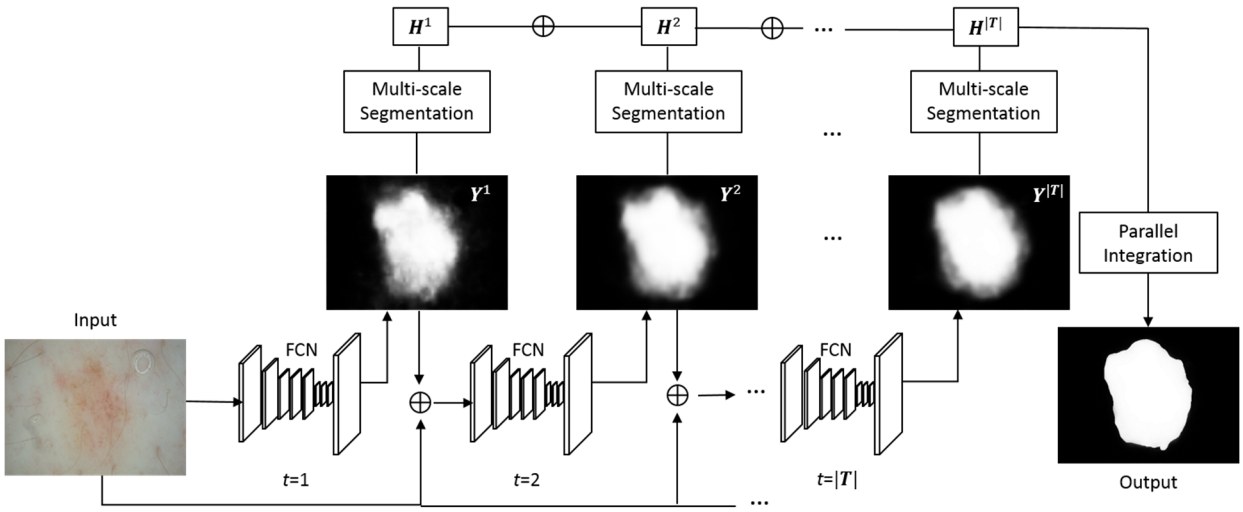

Figure 2: Flow diagram of our proposed multi-stage fully convolutional networks with parallel integration (mFCN-PI) method.

only boosting the training data but also in optimizing the learning of the lesion boundaries, which are usually difficult to segment. During prediction, mFCN uses dermoscopic images and the estimated probabilities derived from previous stage to iteratively and gradually improve the segmentation accuracy.

We propose a parallel integration (PI) approach to further refine the segmentation of the skin lesion boundaries. We integrated the complementary segmentation results produced at individual stages of the mFCN to encourage agreement between the labels of neighboring pixels, which ensured that the appearance of segmented lesion was spatially consistent and resulted in better segmentation of the boundaries.

The rest of the paper is organized as follows. Section 2 describes our method. Sections 3 presents the experimental results on two well-established public datasets, comparing our method to the existing state-of-the-art methods and also the conventional FCN architecture. This is followed by discussions in Section 4. Finally, conclusions are made in Section 5.

II. Methods

A. Overview of the Framework

The outline of our proposed segmentation method is shown in Figure 2. The mFCN was applied to the input dermoscopic images to obtain a probability map $\boldsymbol{Y}^1$, $\boldsymbol{Y}^2$,…,$\boldsymbol{Y}^{|\boldsymbol{T}|}$ of the lesion area. This estimated probability map together with the input dermoscopic image were then fed into the following mFCN networks. Finally, we used a multi-scale segmentation with parallel integration approach to integrate the segmentation results generated at different stages $\boldsymbol{H}^1$, $\boldsymbol{H}^2$,…,$\boldsymbol{H}^{|\boldsymbol{T}|}$ and to produce the final segmentation results.

B. Multi-stage Fully Convolutional Networks

The traditional FCN architecture contains downsampling and upsampling components [32]. The downsampling part has convolutional and max-pooling layers to extract high-level abstract information and has been widely used in convolutional neural networks (CNN) for image classification related tasks [39]. The upsampling part has convolutional and deconvolutional layers that upsample the feature maps to output the score masks [40].

Convolutional layers are defined on a translation invariance basis and have shared weights across different spatial locations. Both the input and the output of convolutional layers are feature maps and are calculated by convolving convolutional kernels:

$$f_s(\boldsymbol{X};\boldsymbol{W},\boldsymbol{b}) = \boldsymbol{W} *_s \boldsymbol{X} + \boldsymbol{b} \quad (1)$$

Where $\boldsymbol{X}$ is the input feature map, $\boldsymbol{W}$ denotes the kernel, $\boldsymbol{b}$ is the bias, $*_s$ represents convolution operation with stride s. As a result, the resolution of the output feature map $f_s(\boldsymbol{X};\boldsymbol{W},\boldsymbol{b})$ is downsampled by a factor of s. Convolutional layers are usually interleaved with max-pooling layers. Max-pooling layers are form of non-linear downsampling, which is usually used to further improve translation invariance and representation capability [34]. In addition, max-pooling layers have the capability to partition the input into non-overlapping sub-regions, which minimizes the computation cost of the upper layers and also reduces over-fitting [32]. The FCN network can be defined as:

$$\boldsymbol{Y} = \boldsymbol{U}_{\boldsymbol{S}}(\boldsymbol{F}_{\boldsymbol{S}}(\boldsymbol{I};\boldsymbol{\theta});\boldsymbol{\varphi}) \quad (2)$$

where $\boldsymbol{Y}$ is the output prediction, $\boldsymbol{I}$ is the input image, $\boldsymbol{F}_{\boldsymbol{S}}$ denotes the feature map produced by the stacked convolutional layers with a list of stride $\boldsymbol{S}$, $\boldsymbol{U}_{\boldsymbol{S}}$ denotes the deconvolution layers that upsamples the feature map by a list of factors $\boldsymbol{S}$ to ensure both the output $\boldsymbol{Y}$ and input $\boldsymbol{I}$ have the same size (height and width). $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are the learned parameters for convolutional and deconvolution layers.

For skin lesion segmentation, the FCN takes an image of arbitrary size and outputs a probability map of the same size which indicates the lesion area. Our multi-stage FCN embeds the probability map produced at the previous FCN for training and testing and this can be defined as:

$$\begin{cases} \boldsymbol{Y}^t = \boldsymbol{U_S}(\boldsymbol{F_S}(\boldsymbol{I}, \boldsymbol{Y}^{t-1}; \boldsymbol{\theta}^t); \boldsymbol{\varphi}^t), & if\ t > 1 \\ \boldsymbol{Y}^t = \boldsymbol{U_S}(\boldsymbol{F_S}(\boldsymbol{I}; \boldsymbol{\theta}^t); \boldsymbol{\varphi}^t), & if\ t = 1 \end{cases} \quad (3)$$

where t is the stage, and $\boldsymbol{Y}^1$ is the output of the original FCN.

Each stage of our mFCN can be trained individually by minimizing the overall loss between the predicted results and the ground truth annotation of the training data:

$$\underset{\boldsymbol{\theta}^t, \boldsymbol{\varphi}^t}{arg\min} \sum \mathcal{L}(\boldsymbol{Y}^t, \boldsymbol{Z} | \boldsymbol{\theta}^t, \boldsymbol{\varphi}^t) \quad (4)$$

where $\mathcal{L}$ calculates the loss (per-pixel multinomial logistic loss) of the ground truth annotation $\boldsymbol{Z}$ and the predicted results. The mFCN network parameters $\boldsymbol{\theta}^t$ and $\boldsymbol{\varphi}^t$ can then be iteratively updated using stochastic gradient descent (SGD) [41] algorithm.

C. Parallel Integration

To further enhance the segmentation results such as the contour of the lesions, we used a multi-scale segmentation approach on each individual stage thereby improving the robustness of our algorithm when segmenting lesions of various sizes and contrasts. We scaled down the input image at the testing time for a number of times with/without flip (a flipped image is mirrored across the vertical axis). The probability maps of the flipped inputs were flipped back to return to their original orientation and all images were scaled up to return to the same size as the input. We then averaged these results to produce the multi-scale probability map $\boldsymbol{H}^t$ at stage t, defined according to:

$$\boldsymbol{H}^t = \frac{1}{|\boldsymbol{G}| \times |\boldsymbol{L}|} \sum_{\sigma \in \boldsymbol{G}} \sum_{l \in \boldsymbol{L}} \boldsymbol{Y}^t(\boldsymbol{I}, \sigma, l) \quad (5)$$

Where $\boldsymbol{G}$ represents different scales and $\boldsymbol{L}$ represents the flip operation where $\boldsymbol{L} = \{with, without\}$. Scale down was used to reduce the production of coarse images with many isolated regions, which is common when FCN's are applied to upsampled (scale up) images. We used $|\boldsymbol{G}| = 11$ different scales ranging from 0.5 to 1 of the original size of the input image with an increment of 0.05 at the testing time.

In general, the mFCN produced results at different stages were complementary to each other, in which the early stages produced strong skin lesion detection results (detect most of the skin lesion area) while later stages generated finer lesion boundary definitions. Therefore, we produce the final segmentation map by leveraging a parallel integration approach based on cellular automata (CA) [42-44]. CA is an evolving model and has the capability to optimize a single probability map via exploiting local similarity (neighborhood pixels). For each iteration of CA, individual pixels of the probability map propagate according to their neighborhood pixels and constrains the boundary of the probability map. To take advantage of the different stages on different aspects of the segmentation process, we assumed that the spatially corresponding pixels of the probability maps (having the same coordinates) on different stages would have equal influence in determining the pixels' probability value in the following iteration. Therefore, we treat these corresponding pixels the same as the neighborhood pixels and evolved via CA, which can be calculated as:

$$\boldsymbol{H}^t(i+1) = \boldsymbol{H}^t(i) + \sum_{c \in \boldsymbol{T}/t} r \cdot sign\big(\boldsymbol{H}^c(i) - \boldsymbol{\mu}\big(\boldsymbol{H}^c(k)\big) \cdot [1, \ldots, 1]^{\mathrm{T}}\big), k = 0 \quad (6)$$

Where i represents number of iterations, $\boldsymbol{T}$ represents all different stages ($t \in \boldsymbol{T}$), and $\boldsymbol{\mu}$ represents a threshold value for binarizing the original probability map. For simplicity, we set $\boldsymbol{\mu}$ to the Otsu threshold [45]. r is a constant weight that controls the importance of the foreground (skin lesion) and background, and was empirically set to 0.15 to encourage the pixel to follow the neighborhood skin lesion pixels. i was set to 5 iterations to ensure convergence and when $i = 0$, we used the results of the multi-scale probability map at different stages. After the convergence, we averaged the CA produced results of different stages to produce the final integrated probability map.

The final integrated probabilistic map was converted into a binary segmentation result via thresholding at 50% of the maximum value of the map. We refined the segmentation, using the process in Celebi et al.'s prior work [10, 11]: a morphological erosion process was used to smooth the boundary and fill holes, while connected thresholding was used to remove small isolated single pixels.

D. Training mFCN

There is a scarcity of medical image together with annotations for use as training data due to the cost and complexity of the acquisition procedures [46]. In contrast to the limited data in the medical domain, there are much more data available in the field of general images [47]. Existing works have shown evidence that the problem of insufficient training data can be alleviated by fine-tuning, where the lower layers of the fine-tuned network are more general filters (trained on general images) while those in the higher layers are more specific to the target problem [40, 46, 48]. Therefore, we used the off-the-shelf MatConvNet [49] version of FCN trained on the PASCAL VOC 2011 dataset. In order to achieve more precise details of pixel prediction, we fine-tuned a stride-8 FCN architecture (FCN-8s) on the ISBI 2016 Skin Lesion Challenge training dataset (more details about the dataset can be found in Section III); Data augmentation techniques including random crops and flips were used to improve robustness [33, 50]. For the first stage $t = 1$, we fine-tuned the pre-trained FCN model using the RGB dermoscopic images. In the following stages, there was an additional fourth input channel (the probability map), which meant we could not directly fine-tune the FCN-8s, which expected 3-channel inputs (usually RGB images). As such, we replaced one of the RGB channels to facilitate the fine-tuning process because the alternative would require scratch training a new FCN architecture (with 4-channel input) for which there is insufficient training data. We rotated replacement of the three color channels over stages 2 to 4 to cover all different replacement variations. Our experiments on a small training set found that the order of RGB channel replacement had negligible influence on the final segmentation

results. Each stage took about 12 hours to fine-tune over 200 epochs with a batch size of 20 on a 12GB Titan X GPU, with converged at about the 150th epoch.

III. Experiments and Results

A. Datasets

We used two well-established public benchmark datasets to test the effectiveness of our algorithm.

- The ISBI 2016 Skin Lesion Challenge dataset [51] is a subset of the large International Skin Imaging Collaboration (ISIC) archive, which contains images acquired on a variety of different devices at numerous leading international clinical centers. The challenge dataset provides 900 training images (727 non-melanoma and 173 melanoma) and a separate test dataset of 379 images (304 non-melanoma and 75 melanoma) for evaluation.
- The PH2 public dataset [52] was collaboratively collected by the Universidade do Porto, Técnico Lisboa, and the Dermatology service of Hospital Pedro Hispano in Matosinhos, Portugal. All 200 dermoscopic images (160 non-melanoma and 40 melanoma) were obtained under the same conditions through Tuebinger Mole Analyzer system using a 20-fold magnification. They are 8-bit RGB color images with a resolution of 768×560 pixels.

Both datasets provided ground truth segmentations based on the manual delineations by clinical experts.

B. Experiment Setup

We performed the following experiments on the two datasets: (a) analysis of the performance of each component in our proposed method; (b) comparison of the overall performance of our method with baselines from the state-of-the-art; and (c) analysis of the overall performance of our method on non-melanoma and melanoma dermoscopic images. Our method and all supervised baselines were trained on the ISBI 2016 Skin Lesion Challenge training dataset, and tested on both the ISBI 2016 test dataset and the PH2 dataset.

The baseline skin lesion segmentation methods included: (1) MSCA [25] – multi-scale superpixel based cellular automata; (2) SSLS [24] – saliency based skin lesion segmentation; (3) FCN [32] – fully convolutional networks (FCN-8s).

We also compared our method with dataset specific baselines (i.e., methods optimized to a particular dataset). For the ISBI 2016 dataset, these were the top 5 results (out of 28 teams) from the ISBI 2016 Skin Lesion Challenge [51]: ExB, CUMED (Yu et al [37]), Rahman, SFU and TMU. For the PH2 dataset, these were: (1) SCDRR [26] – sparse coding with dynamic rule-based refinement; (2) DT [27] – skin lesion segmentation using delaunay triangulation; and (3) JCLMM [53] – a recently published method for psoriatic plaque segmentation by joining circular-linear distributions with mixture models, this method was also applied for skin lesion segmentation on PH2 dataset.

C. Evaluation Metrics

The most common skin lesion segmentation evaluation metrics were used for comparison including: dice similarity coefficient (Dic.), Jaccard index (Jac.), sensitivity (Sen.), specificity (Spe.) and accuracy (Acc.). They are defined as:

$$Dic. = \frac{2|GT \cap AP|}{|GT|+|AP|} \tag{7}$$

$$Jac. = \frac{|GT \cap AP|}{|GT \cup AP|} \tag{8}$$

$$Sen. = \frac{|TP|}{|TP|+|FN|} \tag{9}$$

$$Spe. = \frac{|TN|}{|TN|+|FP|} \tag{10}$$

$$Acc. = \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|} \tag{11}$$

Where GT denotes the ground truth, AP is the algorithm predicted segmentation result, TP is the true positive pixels (lesions), TN is the true negative pixels (background), FP is the false positive pixels and FN is the false negative pixels. In addition, we calculated the pixel-level receiver operating characteristic (ROC) curve and the precision-recall (PR) curve for additional comparisons. Both ROC and PR curves have

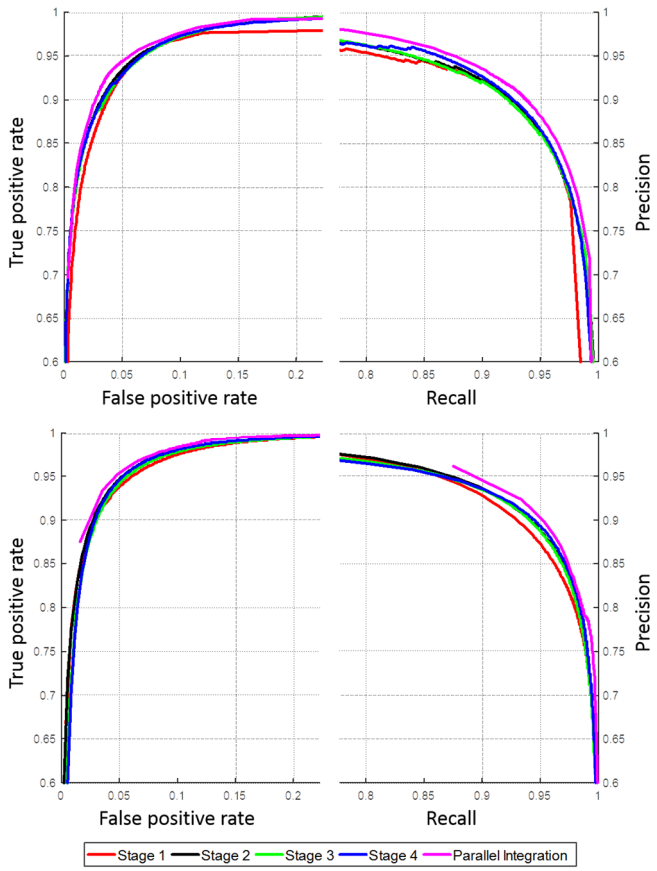

Figure 3. ROC (left) and PR (right) curves of our method at different stages on ISBI 2016 Skin Lesion Challenge dataset (top) and PH2 dataset (bottom).

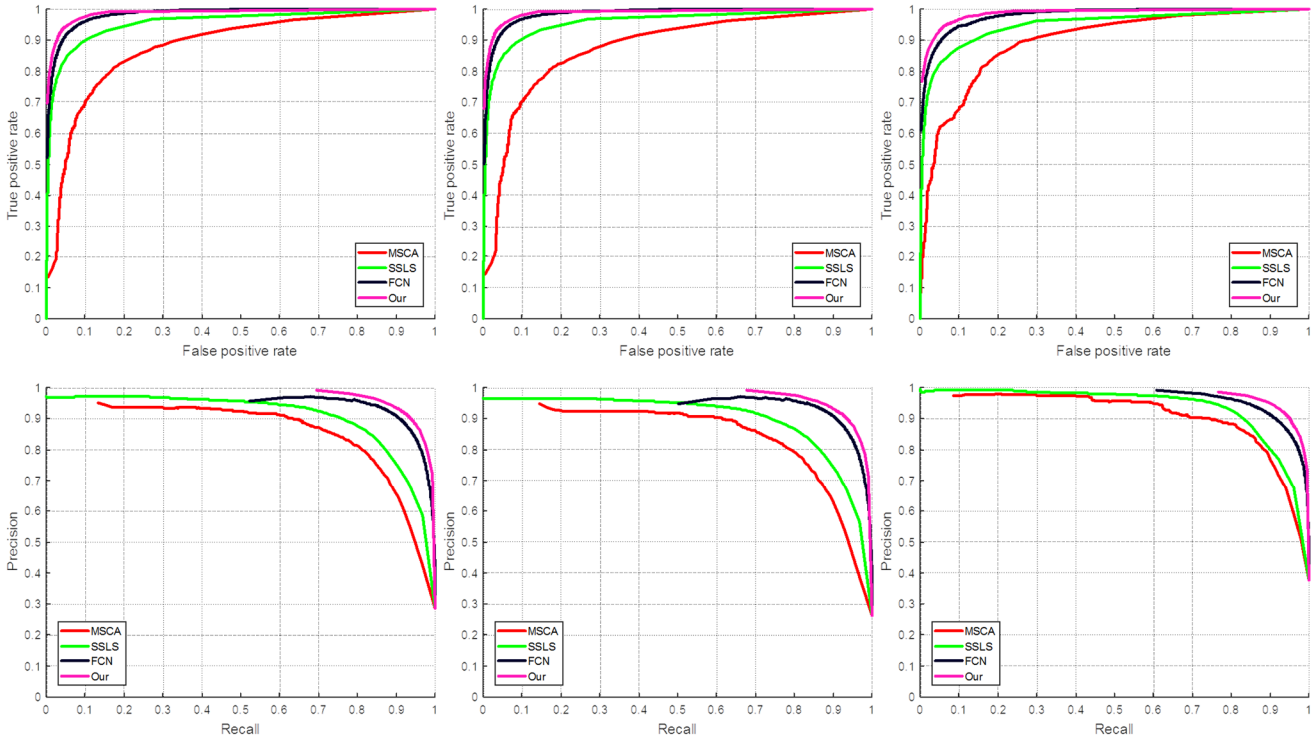

Figure 4. ROC (top) and PR (bottom) curves of different methods on ISBI 2016 Skin Lesion Challenge dataset: overall (left), non-melanoma (middle) and melanoma (right) studies.

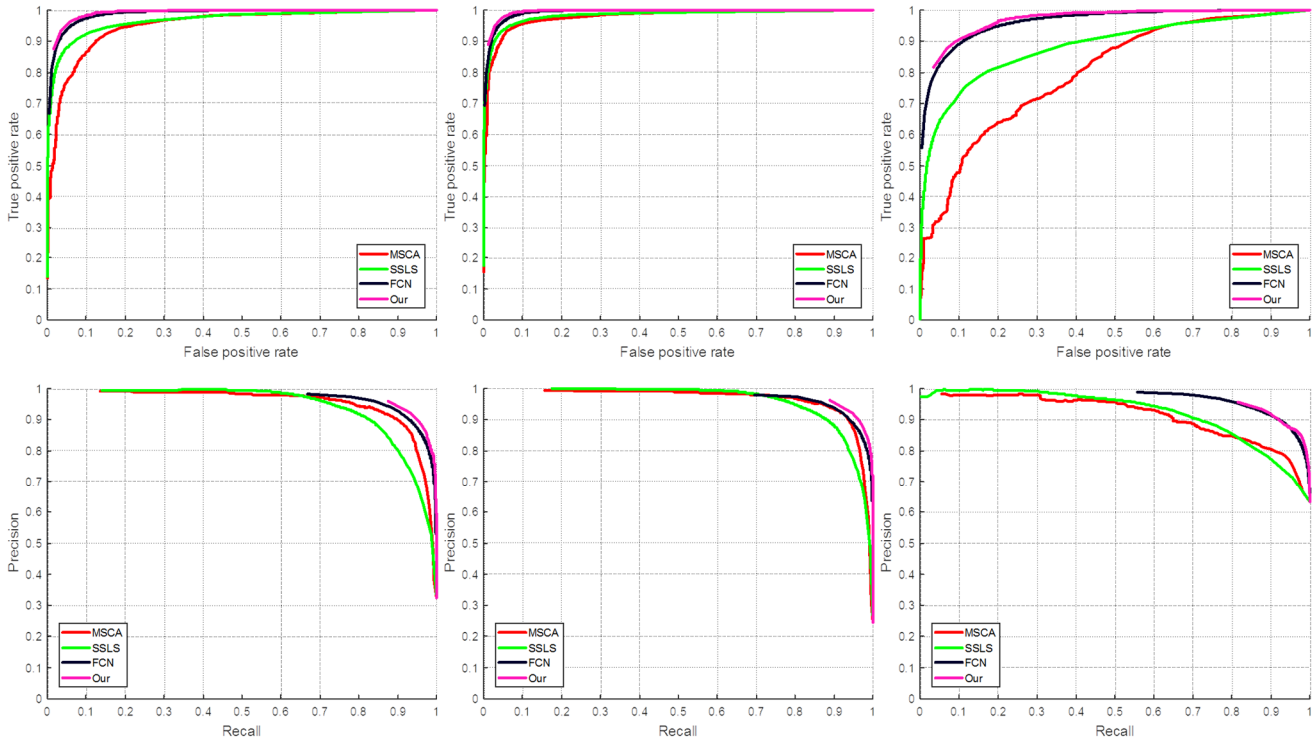

Figure 5. ROC (top) and PR (bottom) curves of different methods on PH2 dataset: overall (left), non-melanoma (middle) and melanoma (right) studies.

been widely used for object detection related problems on general images [54].

D. Component Analysis of our Method

Figure 3 shows the PR and ROC curves of our method at different stages on two datasets. The curves indicate that the results from the later stages of our method can outperform the segmentation at earlier stages. It also illustrates that the parallel integration improved upon the results produced at any

individual stage.

E. Results on ISBI 2016 Skin Lesion Challenge Dataset

Table 1. Segmentation results of our method compared with other methods on the ISBI 2016 Skin Lesion Challenge dataset for all studies, where **Bold** represents the best results.

ISBI 2016 – Overall	Dic.	Jac.	Sen.	Spe.	Acc.
Team - ExB	91.00	84.30	91.00	96.50	95.30
Team - CUMED	89.70	82.90	91.10	95.70	94.90
Team - Rahman	89.50	82.22	88.00	96.90	95.20
Team - SFU	88.50	81.11	91.50	95.50	94.40
Team - TMU	88.80	81.10	83.20	**98.70**	94.60
MSCA	75.88	66.19	78.30	91.31	85.68
SSLS	69.97	57.20	70.04	97.31	84.67
FCN	88.64	81.37	91.70	94.90	94.13
Our	**91.18**	**84.64**	**92.17**	96.54	**95.51**

Table 2. Segmentation results of our method compared with other methods for non-melanoma studies on the ISBI 2016 Skin Lesion Challenge dataset.

ISBI 2016 – Non-melanoma	Dic.	Jac.	Sen.	Spe.	Acc.
Team - ExB	**91.18**	**84.64**	91.12	97.22	**95.78**
Team - CUMED	89.68	82.95	90.82	96.55	95.30
Team - Rahman	89.44	82.04	87.84	97.51	95.70
Team - SFU	88.32	80.88	91.55	95.82	94.93
Team - TMU	88.58	80.73	82.89	**99.04**	94.87
MSCA	75.11	65.57	78.59	91.14	85.84
SSLS	70.81	58.34	72.87	97.15	86.15
FCN	88.66	81.38	91.17	95.87	94.55
Our	90.97	84.34	**91.63**	97.20	95.71

Table 3. Segmentation results of our method compared with other methods for melanoma studies on the ISBI 2016 Skin Lesion Challenge dataset.

ISBI 2016 – Melanoma	Dic.	Jac.	Sen.	Spe.	Acc.
Team - ExB	90.11	82.94	90.57	93.84	93.23
Team - CUMED	89.98	82.90	92.47	92.34	93.21
Team - Rahman	89.93	82.65	88.72	94.44	93.22
Team - SFU	89.44	81.88	91.16	94.13	92.19
Team - TMU	89.68	82.31	84.62	97.48	93.43
MSCA	79.00	68.68	77.12	92.01	85.02
SSLS	66.55	52.59	58.58	**97.94**	78.67
FCN	88.56	81.33	93.83	90.98	92.39
Our	**92.03**	**85.84**	**94.34**	93.89	**94.70**

Tables 1 to 3 and Figure 4 show that our method achieved the overall best performance across all the different measurements. Our method performed better with a large margin for melanoma studies (~3% increase in Jaccard measure) when compared with the best challenge results. Figure 4 also indicates that our proposed method could perform significantly better when compared with the existing fully automatic works with/without supervision.

F. Results on PH2 Dataset

Table 4. Segmentation results on the PH2 dataset for all studies.

PH2 – Overall	Dic.	Jac.	Sen.	Spe.	Acc.
SCDRR	86.00	76.00	-	-	-
DT	-	-	80.24	97.22	89.66
JCLMM	82.85	-	-	-	-
MSCA	81.57	72.33	79.87	95.57	88.75
SSLS	78.38	68.16	75.32	**98.18**	84.85
FCN	89.38	82.15	93.14	93.00	93.48
Our	**90.66**	**83.99**	**94.89**	93.98	**94.24**

Table 5. Segmentation results on the PH2 dataset for non-melanoma studies.

PH2 – Non-melanoma	Dic.	Jac.	Sen.	Spe.	Acc.
DT	-	-	86.79	97.47	93.74
MSCA	85.52	76.88	85.78	96.33	93.86
SSLS	84.72	75.52	83.96	**98.05**	91.77
FCN	89.27	82.01	94.83	94.22	94.79
Our	**90.77**	**84.15**	**95.64**	95.12	**95.61**

Table 6. Segmentation results on the PH2 dataset for melanoma studies.

PH2 – Melanoma	Dic.	Jac.	Sen.	Spe.	Acc.
DT	-	-	54.04	95.97	66.15
MSCA	65.77	54.13	56.25	92.49	68.31
SSLS	53.00	38.73	40.74	**98.67**	57.16
FCN	89.81	82.72	91.39	88.16	88.25
Our	**90.25**	**83.35**	**91.88**	89.42	**88.78**

Tables 4 to 6 and Figure 5 demonstrate that our method significantly improved the existing methods on segmenting skin lesions for PH2 dataset. When compared with the recently published work, it shows an increase of ~8% in Jaccard measure compared with SCDRR, an increase of ~4.5% in Accuracy measure compared with DT, and an increase of ~7.8% in Dice measure compared with JCLMM.

IV. DISCUSSIONS

Our findings show that our method achieved higher accuracy than all other methods on two well-established public datasets. We attribute these benefits to the use of multi-stage FCN with parallel integration and this can be explained as follows: (1) multi-stage FCN enables us to iteratively learn the challenging skin lesion boundaries; (2) the parallel integration approach enables the fusion of the segmentation results to further enhance the detection ability.

We analyzed the main components of our algorithm individually to quantify their contributions to the final segmentation results. Figure 3 shows that the segmentation results at a later stage, e.g., stage 4, can outperform the results at an earlier stage e.g., stage 1. This result underlines the importance of our multi-stage approach to reduce segmentation errors. Figure 3 also shows the advantages from our parallel integration which combines complementary segmentation results produced at individual stages. We attributed the relatively small improvement after stage 2 due to the small training dataset, which limits our method to learn more subtle variations. Nevertheless, our experiments found that the segmentation results after stage 2 still contributed towards a more precise definition of the skin lesion boundary while the early stage results were mainly useful for localizing the skin lesions.

Our experiments indicate that our proposed method has higher overall performance in skin lesion segmentation when compared with the existing methods. In general, while MSCA, SSLS, SCDRR, DT and JCLMM were not able to separate the skin lesions from artifacts e.g., color band (Figure 6a) and performed poorly on challenging skin lesions that have inhomogeneous variations (Figure 6(b, c)) as they are not able to understand image-wide pixel variations and texture differences between the artifacts and the skin lesions. This resulted in ~10% (Table 1) and ~5% (Table 4) lower in

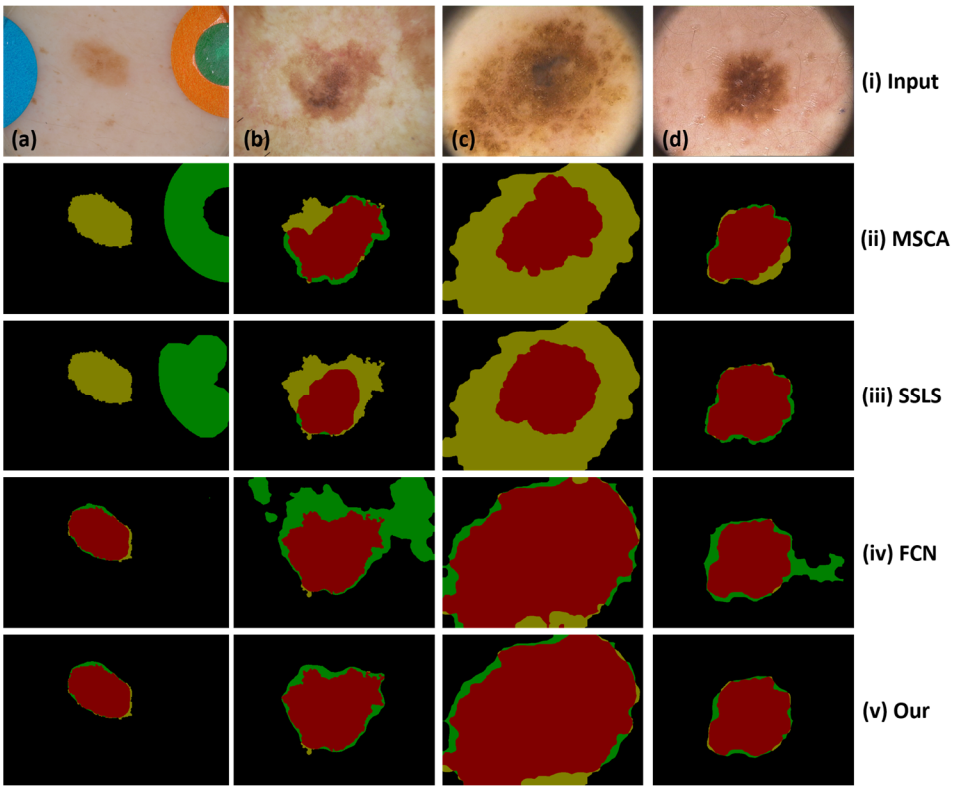

Figure 6. Segmentation results from four example studies. (i) input images, where the first two columns (a, b) are from the ISBI Skin Lesion Challenge dataset and the last two columns (c, d) are from the PH2 dataset; (ii - iv) segmentation results of different comparison methods including MSCA, SSLS and FCN; and (v) segmentation results of our method. The colors represent true positive (red), true negative (black), false positive (green) and false negative (yellow) pixels.

Accuracy measure when compared with the best performing algorithm. These methods had a slightly improved performance on the PH2 dataset, which has less complex variations in lesion locations, illumination and black frame variations. In addition, it would be difficult to adapt these methods for other datasets because their performance is heavily reliant on tuning a large number of parameters and selecting appropriate pre-processing procedures, which will vary widely depending upon the characteristics of the data.

The improvement of FCN over the traditional methods is due to the ability of FCN to combine deep semantic information (upper layers) and shallow appearance information (lower layers) in a hierarchical manner that enables it to encode image-wide location information and semantic characteristics. However, FCN usually generates poor boundary definitions for skin lesions that have fuzzy boundaries and/or low difference in the textures between the foreground and the background (Figure 6b) due to the lack of label refinement and consistency constraints. For these reasons, FCN achieved much lower performance when compared to our method. Evidence for this in Tables 1 and 4, shows that our method had 3.27% (ISBI 2016) and 1.84% (PH2) higher in Jaccard measure when compared with FCN. Figure 6b shows the segmentation results of an example melanoma study, where FCN over-segmented the skin lesion while our proposed method was able to discard the corner regions that had similar visual characteristics to the skin lesions. These results indicate that our method can reduce the segmentation errors by constraining the boundary definitions via integrating different segmentation results.

A comparison with the best ISBI 2016 Skin Lesion Challenge results (Tables 1 to 3) shows that our method had the best overall performance. The methods proposed by ExB and CUMED achieved the most competitive results for overall studies (Table 1) and for non-melanoma studies (Table 2), where in Table 1, ExB was 0.34% lower and CUMED was 1.74% lower while in Table 2, ExB was 0.21% higher and CUMED was 1.39% lower to our method in Jaccard measure. A slightly higher result of ExB method (0.21%) for

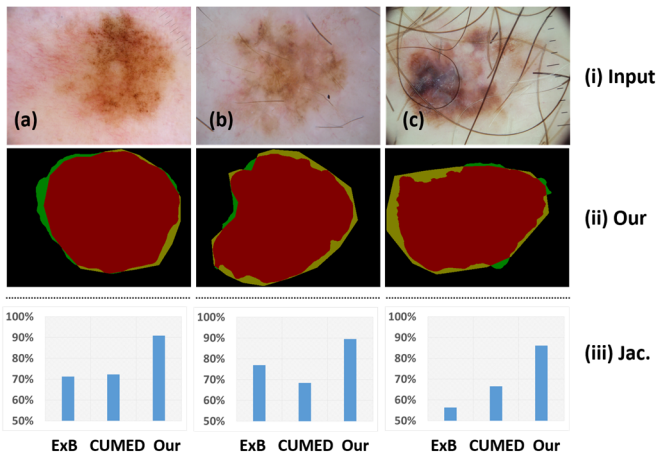

Figure 7. Segmentation results from challenging melanoma studies. (i) input images; (ii) segmentation results of our method; and (iii) comparison of the Jaccard measures of our method with other top methods from ISBI 2016 Skin Lesion Challenge dataset.

non-melanoma studies (Table 2) to our method was likely due to the imbalanced number of melanoma and non-melanoma studies, which causing ExB method overfit to the non-melanoma studies. Furthermore, these methods performed notably poorly for melanoma studies (Table 3), where our method was 2.90% higher than ExB and 2.94% higher than CUMED in Jaccard measure. In general, melanoma studies are more difficult to segment, due to severe inhomogeneity and/or non-uniform boundary patterns. Figure 7 shows the segmentation results of 3 challenging melanoma studies, where the segmentation of the lesions are hindered by hairs, inhomogeneous distributions, and irregular boundaries. In these studies, only our method was able to consistently segment the skin lesions; it had an average of 20% higher in Jaccard measure than ExB and CUMED methods. Furthermore, Tables 1 to 3 also suggest that our method achieved the highest Sensitivity of 92.17%, 91.63% and 94.34%, regardless of skin lesion types (non-melanoma or melanoma studies). Higher Sensitivity means the ability to detect true skin lesions area, which illustrates the robustness of our method on detecting challenging skin lesions. Overall, we attributed the robustness to the use of multi-stage FCN to iteratively learn and infer the visual characteristics of the challenging skin lesions, which ensured that the segmentation errors were always minimized during both training and testing time. In addition, the additional use of parallel integration to integrate the complementary information derived from different stages of mFCN ensured that the challenging skin lesion boundaries were always detected.

V. Conclusions

In this paper, we proposed a new FCN-based method to automatically segment the skin lesions on dermoscopic images. Our method achieved accurate segmentation by combining the important visual characteristics of the skin lesions, which were learned and inferred from multiple embedded FCN stages. Our proposed method leverages the capacity of FCN to segment the skin lesions without using any pre-processing techniques e.g., hair removal or illumination correction. Our experiments on two well-established public datasets demonstrated that our method achieved higher segmentation accuracy compared to the state-of-the-art methods. In the future, we will investigate adaptions of our method to other datasets and as well as potential clinical applications.

References

[1] D. S. Rigel, *et al.*, "The incidence of malignant melanoma in the United States: issues as we approach the 21st century," *Journal of the American Academy of Dermatology,* vol. 34, pp. 839-47, 1996.

[2] M. E. Celebi, *et al.*, "A methodological approach to the classification of dermoscopy images," *Computerized Medical Imaging and Graphics,* vol. 31, pp. 362-73, 2007.

[3] H. Kittler, *et al.*, "Diagnostic accuracy of dermoscopy," *The lancet oncology,* vol. 3, pp. 159-65, 2002.

[4] M. Binder, *et al.*, "Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists," *Archives of Dermatology,* vol. 131, pp. 286-91, 1995.

[5] M. E. Celebi, *et al.*, "Automatic detection of blue-white veil and related structures in dermoscopy images," *Computerized Medical Imaging and Graphics,* vol. 32, pp. 670-77, 2008.

[6] G. R. Day and R. H. Barbour, "Automated melanoma diagnosis: where are we at?," *Skin Research and Technology,* vol. 6, pp. 1-5, 2000.

[7] C. Barata, *et al.*, "Improving dermoscopy image classification using color constancy," *IEEE Journal of Biomedical and Health Informatics,* vol. 19, pp. 1146-52, 2015.

[8] P. Schmid-Saugeona, *et al.*, "Towards a computer-aided diagnosis system for pigmented skin lesions," *Computerized Medical Imaging and Graphics,* vol. 27, pp. 65-78, 2003.

[9] L. K. Ferris, *et al.*, "Computer-aided classification of melanocytic lesions using dermoscopic images," *Journal of the American Academy of Dermatology,* vol. 73, pp. 769-76, 2015.

[10] M. E. Celebi, *et al.*, "Lesion border detection in dermoscopy images," *Computerized medical imaging and graphics,* vol. 33, pp. 148-53, 2009.

[11] M. E. Celebi, *et al.*, "A state-of-the-art survey on lesion border detection in dermoscopy images," *Dermoscopy Image Analysis,* pp. 97-129, 2015.

[12] M. Silveira, *et al.*, "Comparison of segmentation methods for melanoma diagnosis in dermoscopy images," *IEEE Journal of Selected Topics in Signal Processing,* vol. 3, pp. 35-45, 2009.

[13] Z. Ma and J. M. R. Tavares, "A Novel Approach to Segment Skin Lesions in Dermoscopic Images Based on a Deformable Model," *IEEE journal of biomedical and health informatics,* vol. 20, pp. 615-23, 2015.

[14] D. D. Gómez, *et al.*, "Independent histogram pursuit for segmentation of skin lesions," *IEEE Transactions on Biomedical Engineering,* vol. 55, pp. 157-61, 2008.

[15] K. A. Norton, *et al.*, "Three‐phase general border detection method for dermoscopy images using non‐uniform illumination correction," *Skin Research and Technology,* vol. 18, pp. 290-300, 2012.

[16] R. Garnavi, *et al.*, "Border detection in dermoscopy images using hybrid thresholding on optimized color channels," *Computerized Medical Imaging and Graphics,* vol. 35, pp. 105-15, 2011.

[17] F. Peruch, *et al.*, "Simpler, faster, more accurate melanocytic lesion segmentation through meds," *IEEE Transactions on Biomedical Engineering,* vol. 61, pp. 557-65, 2014.

[18] M. Emre Celebi, *et al.*, "Lesion border detection in dermoscopy images using ensembles of thresholding methods," *Skin Research and Technology,* vol. 19, pp. e252-e58, 2013.

[19] H. Zhou, *et al.*, "Mean shift based gradient vector flow for image segmentation," *Computer Vision and Image Understanding,* vol. 117, pp. 1004-16, 2013.

[20] H. Zhou, *et al.*, "Gradient vector flow with mean shift for skin lesion segmentation," *Computerized Medical Imaging and Graphics,* vol. 35, pp. 121-27, 2011.

[21] B. Erkol, *et al.*, "Automatic lesion boundary detection in dermoscopy images using gradient vector flow snakes," *Skin Research and Technology,* vol. 11, pp. 17-26, 2005.

[22] H. Iyatomi, *et al.*, "Quantitative assessment of tumour extraction from dermoscopy images and evaluation of computer-based extraction methods for an automatic melanoma diagnostic system," *Melanoma research,* vol. 16, pp. 183-90, 2006.

[23] M. Emre Celebi, *et al.*, "Border detection in dermoscopy images using statistical region merging," *Skin Research and Technology,* vol. 14, pp. 347-53, 2008.

[24] E. Ahn et al., "Automated Saliency-based Lesion Segmentation in Dermoscopic Images," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015.

[25] L. Bi, *et al.*, "Automated skin lesion segmentation via image-wise supervised learning and multi-scale superpixel based cellular automata," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 1059-62.

[26] B. Bozorgtabar, *et al.*, "Sparse Coding Based Skin Lesion Segmentation Using Dynamic Rule-Based Refinement," in *International Workshop on Machine Learning in Medical Imaging*, 2016, pp. 254-61.

[27] A. Pennisi, *et al.*, "Skin lesion image segmentation using Delaunay Triangulation for melanoma detection," *Computerized Medical Imaging and Graphics,* vol. 52, pp. 89-103, 2016.

[28] P. Wighton, *et al.*, "A fully automatic random walker segmentation for skin lesions in a supervised setting," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2009, pp. 1108-15.

[29] P. Wighton, *et al.*, "Generalizing common tasks in automated skin lesion diagnosis," *IEEE Transactions on Information Technology in Biomedicine,* vol. 15, pp. 622-29, 2011.

[30] A. R. Sadri, *et al.*, "Segmentation of dermoscopy images using wavelet networks," *IEEE Transactions on Biomedical Engineering,* vol. 60, pp. 1134-41, 2013.

[31] Y. He and F. Xie, "Automatic skin lesion segmentation based on texture analysis and supervised learning," in *Asian Conference on Computer Vision*, 2012, pp. 330-41.

[32] J. Long, *et al.*, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-40.

[33] A. Krizhevsky, *et al.*, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-105.

[34] L. Wang, *et al.*, "Saliency detection with recurrent fully convolutional networks," in *European Conference on Computer Vision*, 2016, pp. 825-41.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[36] Y. LeCun, *et al.*, "Deep learning," *Nature,* vol. 521, pp. 436-44, 2015.

[37] Y. Lequan, *et al.*, "Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks," *IEEE Transactions on Medical Imaging,* vol. 36, pp. 994-1004, 2017.

[38] K. He, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-78.

[39] K. Chatfield, *et al.*, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference (BMVC)*, 2014.

[40] H. Chen, *et al.*, "DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation," in *IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 2487-96.

[41] J. Dean, *et al.*, "Large scale distributed deep networks," in *Advances in neural information processing systems*, 2012, pp. 1223-31.

[42] V. Vezhnevets and V. Konouchine, "GrowCut: Interactive multi-label ND image segmentation by cellular automata," *Proc. of Graphicon,* pp. 150-56, 2005.

[43] G. Hernandez and H. J. Herrmann, "Cellular automata for elementary image enhancement," *Graphical Models and Image Processing,* vol. 58, pp. 82-89, 1996.

[44] Y. Qin, *et al.*, "Saliency Detection via Cellular Automata," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 110-19.

[45] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 9, pp. 62-66, 1979.

[46] H.-C. Shin, *et al.*, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE transactions on medical imaging,* vol. 35, pp. 1285-98, 2016.

[47] J. Deng, *et al.*, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-55.

[48] H. Chen, *et al.*, "DCAN: Deep contour-aware networks for object instance segmentation from histology images," *Medical Image Analysis,* vol. 36, pp. 135-46, 2017.

[49] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *ACM international conference on Multimedia*, 2015, pp. 689-92.

[50] A. Kumar, *et al.*, "An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification," *IEEE Journal of Biomedical and Health Informatics,* vol. 21, pp. 31-40, 2016.

[51] 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)," *arXiv preprint arXiv:1605.01397,* 2016.

[52] T. Mendonça, *et al.*, "PH 2-A dermoscopic image database for research and benchmarking," in *IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)* 2013, pp. 5437-40.

[53] A. Roy, *et al.*, "JCLMM: A Finite Mixture Model for Clustering of Circular-Linear data and its application to Psoriatic Plaque Segmentation," *Pattern Recognition,* vol. 66, pp. 160-73, 2017.

[54] X. Li, *et al.*, "Contextual hypergraph modeling for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3328-35.