IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Inferring Skin Lesion Segmentation with Fully Connected CRFs based on Multiple Deep Convolutional Neural Networks

YUMING QIU[1,2,3], JINGYONG CAI[4], XIAOLIN QIN[1,2] and JU ZHANG[3]

[1]Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu, Sichuan, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China
[4]Institute of Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan

Corresponding author: Yuming Qiu (e-mail: ccqiuym@126.com).

This work was supported in part by Sichuan Science and Technology Program under Grants 2018GZDZX0041, 2019ZDZX0005 and 2019ZDZX0006.

ABSTRACT This paper presents a method to infer skin lesion segmentation based on multiple deep convolutional neural network (DCNN) models by employing fully connected conditional random fields (CRFs). This method is on the strength of the synergism between ensemble learning which is responsible for introducing diversity from multiple DCNN models and CRFs inference which is in charge of probabilistic inference based on random fields over dermoscopy images. Contrasting to single DCNN models, the proposed method can gain better segmentation by comprehensively utilizing the advances and performance preferences of multiple different DCNN models. In comparison with simple ensemble schemes, it can effectively and precisely refine the fuzzy lesion boundary by utilizing the information in test images to maximize label agreement between similar pixels. Further, an engineering bonus is the feasibility of parallelization for the heavy operation, predicting on multiple DCNN models. In experiments, we tested the effectiveness and robustness of the proposed method on the mainstream datasets ISIC 2017 and PH2, and the results were competitive with the state-of-art methods. we also confirmed that the proposed method can capture the local information in fuzzy dermoscopy images being able to find more accurate lesion borders with a good boost on Boundary Recall (BR) metric. Moreover, since the hyper-parameters in CRFs are explainable, it is possible to adjust them manually to reach better results case by case, being attractive in practice. This work is of value on integration between the deep learning technologies and probabilistic inference in resolving lesion segmentation, and has great potential to be applied in similar tasks.

INDEX TERMS Pigmented Skin Lesion Segmentation, Fully Connected CRFs, Deep Convolutional Neural Networks, Ensemble Learning

I. INTRODUCTION

Pigmented skin lesion border structure provides valuable information and clinical features such as asymmetry and irregularity for accurate diagnosis, and lesion borders are helpful for extracting other clinical features such as atypical pigment networks, blue-white areas and dots [1]–[5]. Therefore, lesion border detection or lesion segmentation is extremely important in analyzing dermoscopy images that are the major imaging modality in the diagnosis of pigmented skin lesions [1], [5]–[7]. In recent years, many attentions have turned to employment of various deep convolutional neural network (DCNN) methods such as FCNs [8], DeepLab [9], [10] and Mask R-CNN [11], and have made enormous progress [12]–[16] in lesion segmentation. It is known that deep neural network architectures vary in characteristics, strengths and weakness, even different hyper-parameters or initialization parameters in a same DCNN architecture may lead to different segmentation, so their segmentation results from a model may have different emphasis on different aspects, called **performance preferences** here. In practice, single DCNN models usually perform unstably caused by different model architectures and/or hyper-parameters. A promising direction is to integrate all segmentation from different DCNN models to improve the final segmentation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3014787, IEEE Access

Nevertheless a subsequent crucial challenge is how to carry out this idea and make sure it works. Pixel-wise majority voting is an apparent way, but simple voting does not take advantage of the information in test images, resulting in that the voting results usually cut off the obvious lesion continuity and miss some import lesion regions as shown in Figure 2. Moreover, the methods based on some single DCNN-based model resolve lesion segmentation on the global level upon whole training dataset, reckoning without the information in local images.

In this paper, we propose a method which does not only comprehensively take advantage of different DCNN models either homogeneous or heterogeneous but also can address the problem of voting ensemble for lesion segmentation. This new method firstly uses same annotated data to train multiple DCNN models with heterogeneous neural network architectures or homogeneous architectures with different hyper-parameters, and then infers the lesion segmentation of a test dermoscopy image by means of a procedure with three steps, namely obtaining multiple sheets of lesion segmentation from these pretrained DCNN models correspondingly, generating unary potential based on these segmentation, and inferring the final segmentation with fully connected CRFs based on the unary potential and the original test image. Through these steps, we can integrate the lesion segmentation from various DCNN models into a probabilistic inferring procedure. The core idea in this new method is to employ probabilistic graphical models to handle the inconsistency of different models rather than simply using majority voting scheme. In other words, we pose image segmentation and labeling as maximizing a posteriori inference in a Markov Random Fields (MRFs) or its variant Conditional Random Fields (CRFs) defined over pixels or image patches [17]–[20] than simple majority voting. The greatest benefit of this thought is that it is capable of maximizing label agreement between similar pixels by CRFs based on the strong generalization and great diversity from multiple DCNN models, and then able to refine weak and coarse pixel-wise predictions to produce fine-grained segmentation. Contrasting with separate DCNN models, the proposed method can gain better segmentation by comprehensively utilizing the performance preferences of multiple DCNN models. And in comparison with simple voting ensemble scheme, it can effectively segment the lesion boundary by harnessing the information in test images. In addition, the heavy operation – predicting on multiple DCNN models are able to be parallelized in engineering implementation.

In order to validate the effectiveness and robustness of the proposed method, some experiments were carried out on the mainstream dataset ISIC 2017 [21] for skin lesion segmentation. First of all, we trained various models through three remarkable convolutional networks methods FCNs [8], DeepLab [9], [10], and Mask R-CNN [11] with different hyper-parameters, and selected 15 models of them to form one set from which 7 models were selected to form another set containing less models. Then we used the proposed method to infer the lesion segmentation based on the two sets of DCNN models for every test image in ISIC 2017 and PH2. At last, we adopted 5 metrics, mean accuracy (mAC), mean Dice coefficient (mDC), mean Jaccard index (mJI), mean thresholded Jaccard index (mTJI) and mean Boundary Recall (mBR) to evaluate the performance comprehensively, and analyzed the results taking simple ensemble schemes, single DCNN models, single DCNN models plus CRFs as baselines. The experimental results showed that the performance of proposed method on ISIC 2017 and PH2 exceeded baselines on most of all metrics, especially being 5.57% higher than voting ensemble, 4.76% higher than the best score in single DCNN models and 7.59% higher than single DCNN models plus CRFs on ISIC 2017 on mTJI metric, the newest metric used in ISIC 2018 [22]. In comparison with the state-of-the-art methods, the score of our method on the decisive mJI metric is competitive. More remarkably, through introducing the metric Boundary Recall (BR) in skin lesion segmentation, we confirmed that the proposed method can capture the local information in fuzzy dermoscopy images being able to find more accurate lesion borders with a good boost on BR metric. This shows that our proposed method can markedly improve the lesion segmentation in comparison with separate models, separate models plus CRFs and simple ensemble schemes. Moreover, since the hyper-parameters in CRFs have explainable meaning as described in Section III-D, we can manually adjust them to reach better results case by case. It is an attractive character in practice. In addition, under a broader perspective out of the specific skin lesion segmentation task, we believe that our proposed method has great potential to be applied in other fields.

In summary, the contribution of this paper is four-fold:

- This work is one of the first attempts to combine probabilistic inference and the ensemble of multiple DCNN models for skin lesion segmentation.
- We propose a feasible and effective method to infer skin lesion segmentation with fully connect CRFs based on multiple DCNNs.
- The proposed method can effectively and precisely refine the fuzzy lesion boundary by utilizing the information in test images to maximize label agreement between similar pixels. The experimental results are very competitive with the state-of-art methods.
- The proposed method has great potential to be taken as a general framework applicable to other similar tasks.

The rest of this paper is organized as follows. Section II firstly reviews related works. Then the details of proposed method are presented in Section III and experiments on mainstream datasets are introduced in Section IV. Section V discusses some limits and further works. At last conclusions are drawn in Section VI.

II. RELATED WORKS

Lesion segmentation has been a challenging task for many years since the late 1990s. Many traditional methods such

as thresholding [23]–[25], clustering [26], [27], color quantization [28], and region growing [29] which usually exploit pixel values, color, texture and shape, were proposed to settle this challenging task before the deep learning for semantic segmentation became trendy recently. More works can be found in [1], [30] which reviewed numerous efforts comprehensively. With the availability of some public datasets such as PH2 [31] and ISIC series [21], [22], [32] and the rise of deep learning especially the deep convolutional neural networks toward semantic segmentation, attention has turned into DCNN-based methods. In the International Skin Imaging Collaboration (ISIC) 2017 Challenge [21] and 2018 Challenge [33] at the International Symposium on Biomedical Imaging (ISBI), most of top ranked participants took advantage of various dominated DCNN methods such as the FCNs model [8], Mask R-CNN [11] or their variations. In addition, some DCNN-based methods such as FCA-Net [15], SkinNet [34], PA-Net [35], FrCN [36], iMSCGnet [37] and Slsdeep [38] were proposed as well in recent years. These efforts were mainly based on some single model resolving the lesion segmentation problem on the global level upon whole training dataset. As mentioned in Section I, single DCNN models have performance preferences. Hence some researchers turned to harness the ensemble of segmentation on various deep neural networks. The work [39] employed an ensemble of 10 networks, each with different parameter values, was created and the outputs of the 10 were averaged to create a combined segmentation prediction. The work [14] also ensembled the segmentation from Mask R-CNN [11] and DeepLabV3 [10] and gained improvement in ISIC 2017 dataset. Their efforts [14], [39] did not harness the information in original dermoscopy images. In addition, although there are some efforts such as [40]–[42] using CRFs inference to refine the segmentation from DCNNs for general semantic segmentation, they used CRFs inference after the output layer of a DCNN model. Unlike them, we are combining CRFs models and the ensemble from multiple DCNNs rather than a single DCNN model. The key of underlying problems is how to build association between the ensemble and a CRF model, and we will introduce our solution in the next Section.

III. METHODOLOGY

Although deep neural networks have powerful strengths for skin lesion segmentation, the complicated structure and massive parameters make it hard to obtain a stable solution of parameters. On the one hand, different deep neural network methods have different characteristics, strengths and weakness. On the other hand, different training process may get different models leading to unstable lesion segmentation. Therefore, in order to comprehensively take advantage of multiple DCNN models, stabilize the prediction and refine the lesion segmentation, this paper proposes a method to cope with this challenging task. This section firstly presents the overview of this method, and then introduces the three steps of inferring procedure that are the core contents in our proposed method.

A. OVERVIEW

Roughly speaking, the new method can be mainly divided into two phases: training multiple DCNN models and inferring lesion segmentation with CRFs based on these pretrained models. In training phase, various semantic segmentation methods such as FCNs [8], DeepLab [9], [10] and Mask R-CNN [11] are employed to train multiple models. A type of DCNN method can train multiple models by different hyper-parameters or initialization parameters. In inferring phase, the lesion segmentation of a test dermoscopy image is inferred as the following three steps: obtaining multiple sheets of lesion segmentation from different pretrained DCNN models in training phase, generating unary potential based on these segmentation, and inferring the final segmentation with fully connected CRFs. This procedure can be seen in Figure 1 intuitively. The inferring procedure is the essential core of the proposed method. The underlying theoretical thought of this new method is to employ probabilistic graphical models to handle the inconsistency of different models rather than simply using majority voting scheme, based on the observation that most models can get agreement in conspicuous lesion regions but usually are inconsistent in some fuzzy boundaries. Further, CRFs inference for image segmentation is capable of maximizing label agreement between similar pixels, and refine weak and coarse pixel-wise predictions to produce fine-grained segmentation. Therefore, we propose this method to infer the lesion segmentation. In the following three subsections, we will detailedly introduce the three steps in inferring phase.

B. OBTAINING MULTIPLE SHEETS OF SEGMENTATION

For a test dermoscopy image I, we can run each model M_i in all n models pretrained in training phase to predict a corresponding lesion segmentation S_i. It is noticeable that although we have been stating the models are trained by DCNNs, actually a lesion segmentation can be obtained by any methods including artificial assistance and traditional methods mentioned in previous section. This step is relatively simple, and the only discussible question is how to determine the number n of models. Obviously, n should be greater than 1, and it may be unsatisfactory if the number n is too small, because it is generally recognized that the key in ensemble learning is to effectively generate individual learners with strong generalization ability and great diversity. So for lesion segmentation, more individual models are beneficial for inferring in next steps. However, using more individual models will bring about more expensive cost. We empirically suggest that the number n should better be in the range $(5, 20)$.

C. UNARY POTENTIAL GENERATION

After obtaining multiple sheets of lesion segmentation from different DCNN models, it is critical to find a way to build a bridge between these sheets of segmentation and CRFs inference in next step. Here, we firstly blend all sheets of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3014787, IEEE Access

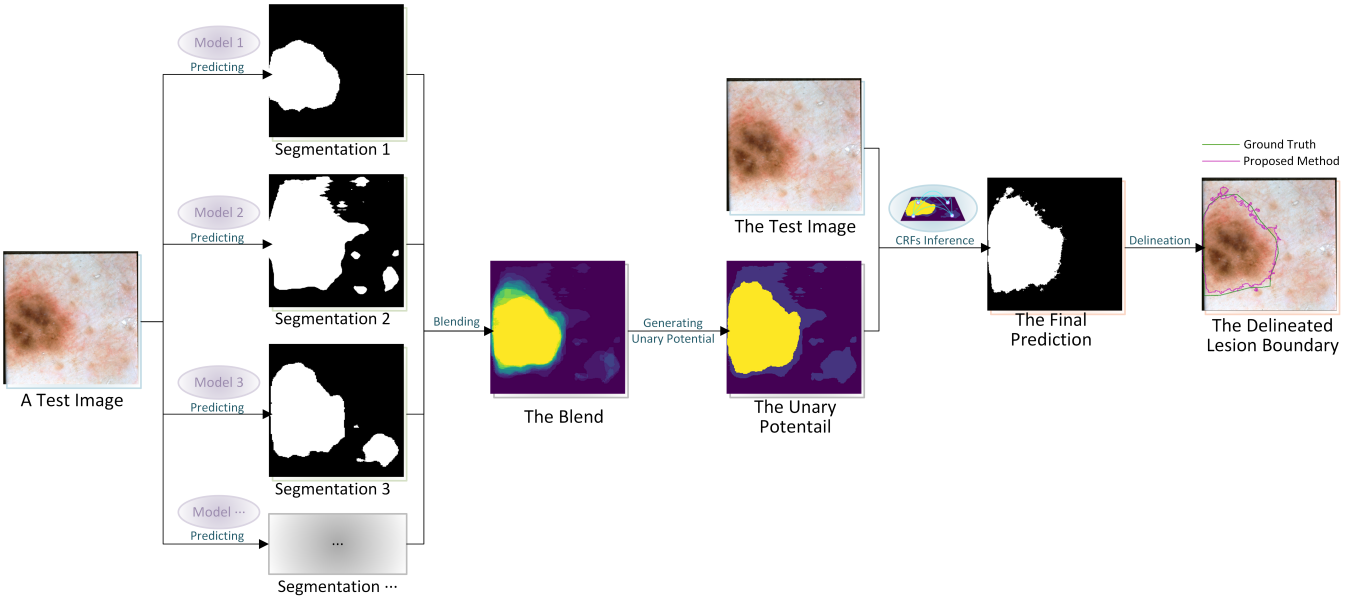


FIGURE 1. The procedure for inferring a test dermoscopy image. For a specified input test image, the blend of multiple sheets of lesion segmentation is firstly obtained from multiple different pretrained DCNNs models. Then the unary potential of fully connected CRFs is generated based on the blend. Lastly, the final prediction is inferred based on the original test image and the unary potential by fully connected CRF.

segmentation and then generate a unary potential based on the blend as an input of CRF models. Formally, for a set of lesion segmentation $\{S_1, S_2, \cdots, S_m | m \in \mathbb{N} > 1\}$, all sheets of segmentation are blended through pixel-wisely adding the labels in each segmentation, namely the blend is $\mathbf{b} = \Sigma_{r=1}^{m} S_r$. In detail, for a pixel i, the value in a same location in the blend is $\mathbf{b}_i = \Sigma_{r=1}^{m} S_{r,i}$ where $S_{r,i}$ means the label value in pixel i in segmentation S_r. If the label is "lesion", the label value is 1, and if the label is "background", the label values is 0. Let u denotes the lesion label, the unary term $\psi_u(x_i)$ measuring the cost of assigning label u to the i-th pixel is

$$\psi_u(x_i) = \begin{cases} -\log(p) & \text{if } \mathbf{b}_i >= \tau \\ -\log(1-p) & \text{if } \mathbf{b}_i <= \kappa \\ -\log(0.5) & \text{otherwise} \end{cases} . \quad (1)$$

Likewise, the unary term for label "background" can be calculated by replace p in Equation 1 with $1-p$. Equation 1 reflects our assumption for the blend of segmentation from multiple DCNNs. Namely, for a pixel, we only consider it as "lesion" with the probability p when at least τ models simultaneously predict it as "lesion"; and we only consider it as "background" with the probability p when at least $(m-\kappa)$ DCNN models simultaneously predict it as "background"; otherwise, we treat it as an unsure label. Obviously, τ, κ and p are hyper-parameters specified empirically or found by some search methods such as grid search and random search. For skin lesion segmentation task, we can further assume κ is 0 according to our experience and experimental observation. In other words, we only consider a pixel as "background" with the probability p when all models simultaneously predict it as "background". Hence, Equation 1 can be simplified as

$$\psi_u(x_i) = \begin{cases} -\log(p) & \text{if } \mathbf{b}_i > \tau \\ -\log(1-p) & \text{if } \mathbf{b}_i = 0 \\ -\log(0.5) & \text{otherwise} \end{cases} . \quad (2)$$

Certainly, it should depend on some specific task and the real situation to ascertained the assignment of κ. For simplification, we will directly use Equation 2 in the rest of this paper. Once the unary potential is calculated, we can take the unary potential and the original test image into CRF models to infer the final segmentation.

D. CRFS INFERENCE

Commonly, a digital dermoscopy image is a raster graphics (or called bitmap image) consisting of a generally rectangular grid of pixels (points of color). But beyond this view, we can assume an undirected graph where each node represents a pixel in an image $\mathbf{I}$ of size N, and be associated with a latent variable $X_i^u \in \mathcal{L}; \mathcal{L} = \{l_1, l_2, \cdots, l_k\}$ indicating a label u of a pixel i where $\mathcal{L}$ is a pre-defined set of labels, and each edge represents relation between pixels. The set of latent variables $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ can be considered as a random field. The pair $(\mathbf{I}, \mathbf{X})$ can be modeled as a conditional random field (CRF) characterized by a Gibbs distribution $P(\mathbf{X} = \mathbf{x}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp(-E(\mathbf{x}|\mathbf{I}))$. $E(\mathbf{x}|\mathbf{I})$ is called the energy of the configuration $\mathbf{x} \in \mathcal{L}^N$ and $Z(\mathbf{I})$ is the partition function [43]. The maximum posteriori labeling of the random field is $x^* = \text{argmax}_{\mathbf{x} \in \mathcal{L}^n} E(\mathbf{x}|\mathbf{I})$. Consider $\mathcal{G} = (\nu, \varepsilon)$ is a graph on $\mathbf{X}$, and each clique c in a set of cliques $\mathcal{C}_\mathcal{G}$ in $\mathcal{G}$, the Gibbs energy of a labeling $\mathbf{x} \in \mathcal{L}^N$ is $E(\mathbf{x}|\mathbf{I}) = \Sigma_{c \in \mathcal{C}_\mathcal{G}} \phi_c(\mathbf{x}_c|\mathbf{I})$ where ϕ_c is the potential of a clique c.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3014787, IEEE Access

According to the assumption of the potentials, there are different CRF models. Basic CRF models [19], [44], [45] are composed of unary potentials on individual pixels or image patches and pair-wise potentials on neighboring pixel or patches. The adjacency CRF structure is limited in its ability to model long-range connection with the image and generally results in excessive smoothing of object boundaries [46]. Some works expanded the basic CRF framework to incorporate hierarchical connectivity and higher-order potentials defined on image regions [17], [18]. The most advanced form is fully connected CRFs, which establishes pairwise potentials on all pair of pixels in the image [46], [47]. Our goal is to obtain the lesion segmentation and lesion boundary based on the blend, the fully connected CRFs is helpful to recover local detailed structure.

In the fully connected pairwise CRF model [46], $\mathcal{G}$ is the complete graph on $\mathbf{X}$, and $\mathcal{C}_{\mathcal{G}}$ is the set of all unary and pairwise cliques. Then, the energy of a label assignment $\mathbf{x}$ is given by

$$E(\mathbf{x}) = \sum_{i} \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j). \tag{3}$$

In Equation 3, $\psi_u(x_i)$ is the unary term measuring the cost of assigning labels to the i-th pixel, and $\psi_p(x_i, x_j)$ is the pairwise term that measures the penalty of assigning labels to pixels i, j. In this work, the unary potentials are generated from the blend in previous step (see Equation 2 in Section III-C). Obviously, the unary term does not consider the smoothness and the consistency of the label assignments. What is gratifying is that the pairwise energies provide an image data-dependent smoothing term that encourages assigning similar labels to pixels with similar properties. Intuitively, assigning same labels to similar pixels is more acceptable than assigning different labels. The unary term and the pairwise term are both fundamental and indispensable for fully connected CRFs. It can be seen that the core problem is how to connect the ensemble of multip DCNN models to fully connected CRFs, and it is obvious that we can not directly utilize the fully connected CRFs to solve this problem. The unary potential generation (described in Section III-C) is the key to bridge the gap between them. As in [46], using contrast-sensitive two-kernel potentials, we can define the pairwise term as

$$\psi(x_i, x_j) = \mu(x_i, x_j)[w^{(1)} k^{(1)}(\mathbf{f}_i, \mathbf{f}_j) + w^{(2)} k^{(2)}(\mathbf{f}_i, \mathbf{f}_j)] \tag{4}$$

where $k^{(1)}$and $k^{(2)}$ are two Gaussian kernels, $\mathbf{f}_i$ and $\mathbf{f}_j$ are feature vectors for pixels i and j, $w^{(1)}$ and $w^{(2)}$ are linear combination weights, and μ is the label compatibility function. Concretely, the first kernel is an appearance kernel as

$$k^{(1)}(\mathbf{f}_i, \mathbf{f}_j) = \exp\left(-\frac{\|p_i - p_j\|^2}{2\theta_\alpha^2} - \frac{\|I_i - I_j\|^2}{2\theta_\beta^2}\right),$$

which depends on both pixel position p and RGB color I. It forces nearby pixels with similar color and position to have similar label, and the degrees of nearness and similarity are controlled by parameters θ_α and θ_β. The second kernel is a smoothness kernel as

$$k^{(2)}(\mathbf{f}_i, \mathbf{f}_j) = \exp\left(-\frac{\|p_i - p_j\|^2}{2\theta_\gamma^2}\right),$$

which only depends on pixel positions. It can remove small isolated regions and the degree is controlled by parameter θ_γ. The label compatibility function μ in Equation 4 introduces a penalty for nearby similar pixels that are assigned different labels. For the lesion segmentation task, the label compatibility function can use the simplest one given by Potts model, namely $\mu(x_i, x_j) = [x_i \neq x_j]$. More details of the pairwise term can be found in [46]. Although some works [48] pointed that the pairwise term definition has drawbacks, neglecting the spatial context between objects and missing high-order interactions between pixels, it is unnecessary to concern because the lesion segmentation is only a binary labeling task.

Minimizing the CRF energy $E(\mathbf{x})$ can yield the most probable label assignment $\mathbf{x}$ for the given dermoscopy image. Obviously, it is intractable to reach exact minimization. In [46], a mean field approximation to the CRF distribution was employed to approximate maximum posterior marginal inference. This approximation yields an iterative message passing algorithm which can be performed using Gaussian filtering in feature space based on the fully connected CRF model. This can reduce the complexity of message passing from quadratic to linear, resulting in an approximate inference algorithm for fully connected CRFs that is linear in the number of variables N and sub-linear in the number of edges in the model. More details of CRFs inference can be found in [46], [49].

IV. EXPERIMENTS

For validating the effectiveness of the proposed method, we conducted some experiments on two mainstream datasets. This section describes the details of these experiments.

A. SETUP

For datasets, we used two mainstream public datasets ISIC 2017 [21] and PH2 [31]. ISIC 2017 is a recent public resource in the study community from the International Skin Imaging Collaboration (ISIC). It contains 2000 dermoscopy images for training, 150 for validation and 600 for testing. We used the validation part to find the hyper-parameters of the CRF model, and used the test part to evaluate the performance. PH2 is a dermoscopy image database for research and benchmarking, which includes the manual segmentation, the clinical diagnosis, and the identification of several dermoscopic structures, performed by expert dermatologists, in a set of 200 dermoscopic images. Due to relatively little data, a prevailing practice [14], [36], [37] is to use all of them as test data to test the models training on ISIC 2017. But the fully connected CRFs need some supervised data to find a set of better hyper-parameters. So we had to divide 200 images into two parts: validation containing 50 once-randomly-selected

images and test part containing the rest 150 images. All experiments for PH2 used an one-off division. For the newest dataset ISIC 2018, it did not share the ground truth of their testing set. More importantly, most of data in ISIC 2018 and 2017, according to our investigation, are completely identical, so it is less indispensable to validate the method on an almost same basis. Some recent works such as [14] were not using ISIC 2018 as well. Therefore, our work was only based on the ISIC 2017. Table 1 provides the datasets' description.

TABLE 1. Description of datasets used in experiments.

Dataset Name	Training	Validation	Test	Memo
ISIC 2017 [21]	2000	150	600	as the official
PH2 [31]	-	50	150	randomly selected

For evaluation, we used 5 metrics, mean Accuracy (mAC), mean Dice Coefficient (mDC), mean Jaccard Index (mJI), mean thresholded Jaccard index (mTJI) and mean Boundary Recall (mBR). AC, DC and JI used in [21], [32] for each test case are defined as:

$$\begin{aligned} \mathrm{AC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \\ \mathrm{DC} &= \frac{2*\mathrm{TP}}{\mathrm{FN}+\mathrm{FP}+2*\mathrm{TP}}, \\ \mathrm{JI} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}, \end{aligned}$$

where TP, TN, FP and FN refer to the number of true positive, true negative, false positive, and false negative pixels respectively. In order to emphasize the higher Jaccard index, thresholded Jaccard Index (mTJI) was introduced in ISIC 2018 [22], which works similarly to standard Jaccard, with one important exception: if the Jaccard value of a particular mask falls below a threshold T, the Jaccard is set to zero. The value of the threshold T defines the point in which a segmentation is considered incorrect. The threshold was set to 0.65 empirically by the official ISIC. In the following text, we will denote it as $\mathrm{mTJI}_{0.65}$. Lastly, in order to assess the boundary-preserving, we introduce a metric Boundary Recall revised from [50] which is the most commonly used metric in evaluation of super-pixels, where it is the fraction of hand-segmented edges which lie within a threshold distance k of any super-pixel edge [51]. For better assessment, we propose to use a self-adapting threshold distance defined as

$$k = k_{\min} + \mathrm{round}((k_{\max} - k_{\min}) * (N_{\mathrm{GT}}/M)), \tag{5}$$

in which N_{GT} denotes the number of pixels in an annotated mask and M refers the number of all pixels. The parameters $k_{\min}$ and $k_{\max}$ control the strictness of boundary-preserving. In our implement, we empirically set $k_{\min}$ and $k_{\max}$ as 2 and 5 respectively. Given an annotation GT and a segmentation S, the boundary recall are defined as

$$\mathrm{BR}(\mathrm{GT}, S) = \frac{\mathrm{TP}(\mathrm{GT}, S)}{\mathrm{TP}(\mathrm{GT}, S) + \mathrm{FN}(\mathrm{GT}, S)}, \tag{6}$$

in which TP and FN are the number of true positive and false negative boundary pixels respectively. Generally, a higher BR score stands for better boundary-preserving.

B. BASELINES

In consideration of the sides related this work, we set baselines from three aspects. The first aspect is about single DCNN models through which we can observe the difference between using DCNNs directly and the proposed method. Concretely, we used the dominated DCNNs for semantic segmentation including FCNs [8], DeepLab [9], [10] and Mask R-CNN [11]. We will use the prefix "S-" to represent the methods in this aspect in the following related tables.

The second aspect is about the methods appending fully connected CRFs after single DCNNs as a post-processing. There were some works [40]–[42] employing fully connected CRFs in deep learning for semantic image segmentation. In our experiments, we used the way described in [9] to append fully connected CRFs after the networks of FCNs [8] and DeeplabV2 [9]. Due to the lack of feasible practices or works about appending CRFs after DeepLabV3 [10] and Mask R-CNN [11], let alone the original DeepLabV3 [10] in which CRFs are deprecated, we did not use fully connected CRFs after them. We used FCNs+CRFs and DeeplabV2+CRFs to show the difference between single DCNN models plus CRFs and the proposed method. We will use the prefix "SC-" to label them in the following related tables.

The last but not least aspect is using other combination schemes on the ensemble of multiple DCNN models. We set three schemes as baselines: Ensemble-Voting, Ensemble-Intersection and Ensemble-Union. As in Section III-C, given a set of lesion segmentation $\{S_1, S_2, \cdots, S_m | m \in \mathbb{N} > 1\}$ and the corresponding blend $\mathbf{b} = \Sigma_{r=1}^{m} S_r$, the Ensemble-Voting can be expressed as

$$\mathrm{EV} = \mathbf{b} > 0.5 * m; \tag{7}$$

the Ensemble-Intersection is

$$\mathrm{EI} = \bigcap_{i=1}^{m} S_i; \tag{8}$$

and the Ensemble-Union is

$$\mathrm{EU} = \bigcup_{i=1}^{m} S_i. \tag{9}$$

Ensemble-Voting is actually the traditional majority voting for ensemble learning, which consider a pixel as "lesion" when at least more than half of DCNN models agree this. Ensemble-Intersection and Ensemble-Union are introduced here only for binary lesion segmentation. Ensemble-Union consider a pixel is "lesion" when it is predict as the label "lesion" in any one of all segmentation; while Ensemble-Intersection requires that all DCNN models must agree it as "lesion". It is worth noting that Ensemble-ADD in [14] is a particular case of the baseline EU. When the m is assigned 2, it is the situation presented in [14]. This aspect baselines

are the main that we want to compare because our method is based on the ensemble of multiple models. We will use the prefix "E-" to refer them in the following related tables.

C. IMPLEMENTATION DETAILS

By using four remarkable convolutional networks FCNs [8], DeepLab [9], [10] Mask R-CNN [11] and HRnet [52], we trained dozens of different models with different hyper-parameters upon the training part of ISIC 2017 dataset. Among these models, we finally selected 15 models to implement the proposed method, and we additionally selected 7 ones in 15 selected models to observe the performance on less basic models in order to ease concerns about too much time consumption of the proposed method. The differences between selected models with same network architecture were the learning rate, epoch and augmentations used in training phase. In detail, FCNs models were trained by two learning rates 1e-4, 2e-4 and max epoch 50, and the top 4 models were selected from all checkpoints at every epochs according to the mJI scores rank on validation data part; DeepLabV2 models were trained by learning rate 2.5e-4 and max epoch 60, and the top 4 models were selected from all checkpoints; DeepLabV3 models were trained by learning rates 0.005, 0.01 and max epoch 30, and the top 4 models were selected; Mask R-CNN models were trained by learning rate 0.005 and max epoch 50, and the 3 best models were selected. HRnet were trained with learning rate 4e-3 and max epoch 120, and no model was selected because the best one is not good enough. In addition, we used geometric augmentation in training all models according the cue from [53]. All experiments were run on a tower server equipped two NVidia TITAN Xp GPUs. Let T_F, T_{D2}, T_{D3}, and T_M denote the time consumption of once training for FCNs, DeeplabV2, DeeplabV3 and MaskRCNN respectively, all training time consumption of 15 models is $T_E = 2T_F + T_M + T_{D2} + 2T_{D3}$. From the 15 selected models, we selected 7 ones to evaluate the performance on less basic models. Concretely, 7 models are as follows: three models through two training phases in selected FCNs models, two models through once training in selected DeeplabV2 models, and two models through once training in selected DeeplabV3 models. Likewise, all training time consumption of 7 models is $T'_E = 2T_F + T_{D2} + T_{D3}$. In our implementation, T_F, T_M, T_{D2} and T_{D3} were roughly equivalent to 0.7, 21.5, 15 and 25.5 hours respectively, so, T_E and T'_E roughly equal 88.9 and 41.9 hours respectively.

Then we used the proposed method to infer the lesion segmentation based on the two groups of selected DCNN models for every test image in ISIC 2017 and PH2. Every test image was resized with unchaged length-width-ratio to keep the long side is 512. It is not difficult to prove that metrics mAC, mDC, mJI and mTJI_t in Section IV-A are not sensitive to the scale of image. The reason why we scaled all the test images in a uniform size is to make better sense of the metric mBR because of big size difference between dermoscopy images in ISIC 2017. Further, we combined the grid search and random search to find the optimization of the hyper parameters τ and p in Equation 2 and the hyper parameters in the fully connected CRF on the validation parts in ISIC 2017 and PH2. Generally, τ assigned to about $0.3 * m$ is proper in ISIC 2017 and about $0.8 * m$ in PH2 where m is the number of DCNN models, and p usually was taken as about 0.95 in our experiments.

At last, we used the five metrics mentioned in Section IV-A to measure the performance, and then analyzed the results in comparison with the baseline methods.

D. RESULTS ON 15 DCNN MODELS

Based on 15 DCNN models, the evaluation results on ISIC 2017 and PH2 are shown in Table 2 and Table 3 respectively where every numerical value represents the best score on a metric listed in row head by using some method listed in column head. All things considered, the proposed method performs better than all baseline methods on most metrics introduced in Section IV-A. Especially upon the newest $\mathrm{mTJI}_{0.65}$ metric, the score on ISIC 2017 of the proposed method achieves 74.22% which is 5.57% higher than voting ensemble, 4.76% higher than the best score in single DCNN models and 7.59% higher than single DCNN models plus CRFs. On PH2 dataset, the scores on the main three metrics mDC, mJI and $\mathrm{mTJI}_{0.65}$ are also superior to the compared methods. In other words, in comparison with single models with or without post-processing CRF, the proposed method enhanced the performance of lesion segmentation, and this indicates that the propose method can effectively generate strong generalization and great diversity from multiple DCNN models. In comparison with three ensemble baselines, the proposed method is more inference-backed and superior than them to infer the lesion segmentation, and this indicates that the proposed method can capture the local information in fuzzy dermoscopy images and be able to find more accurate lesion border.

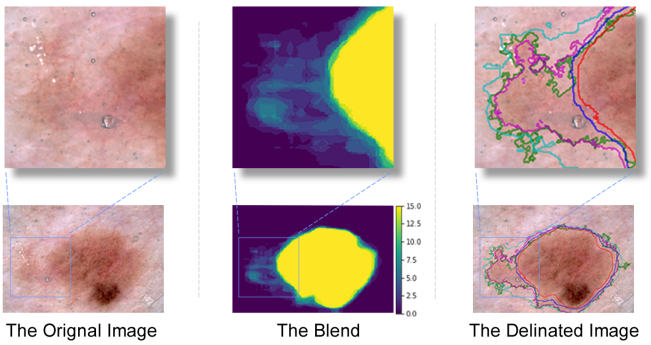


FIGURE 2. An example of the effectiveness of the proposed method. The middle column shows the diversity introduced from multiple DCNN models in the left-lesion-region. The right column presents the lesion segmentation in comparison with three baseline methods, in which green and magenta represent Ground Truth and the proposed method respectively, and blue, red and cyan represent Ensemble-Voting, Ensemble-Intersection, Ensemble-Union respectively.

Let's take a dermoscopy image containing a melanoma lesion as an example as shown in Figure 2, most DCNN models

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3014787, IEEE Access

failed to predict the left-lesion-region which is very fuzzy but extremely vital to diagnose whether it is a melanoma lesion or to decide whether these missed regions should be excised. If we use Ensemble-Voting, the segmentation delineated as blue curve crossing the lesion region which resembles right lesion region and is distinct from the background fails to contain the left-lesion-region. Likewise, using Ensemble-Intersection and Ensemble-Union can not predict the accurate lesion border. However, if we use the propose method, it is successful to predict most pixels in the left-lesion-region by taking advantage of the diversity from multiple DCNN models in the left-lesion-region. The key for this is the integration of the ensemble learning which is responsible for introducing diversity from multiple DCNN models and the CRFs inference which is in charge of probabilistic inference based on the random field over the whole dermoscopy image. More vivid examples can be found in Figure 3 for ISIC 2017 and Figure 4 for PH2. It is worth mentioning that we only show the mask images of baselines related ensemble combination schemes because they are the main baselines deserved to compare with the proposed method and other baselines are involved too many mask images to place into the two figures.

TABLE 2. Performance comparison on 15 DCNN models on ISIC 2017 [21].

Method	mAC	mDC	mJI	$\mathrm{mTJI}_{0.65}$	mBR	TTC
S-FCN8s	80.54	84.04	75.45	66.54	43.52	T_F
S-MaskRCNN	78.50	83.16	75.04	67.72	43.03	T_M
S-DeeplabV2	80.78	84.08	75.27	66.14	38.87	T_{D2}
S-DeeplabV3	85.55	85.70	77.51	69.46	44.85	T_{D3}
SC-FCN8s	80.50	84.06	75.53	66.63	43.40	T_F
SC-DeeplabV2	79.30	83.80	75.39	66.23	41.93	T_{D2}
E-Voting	80.80	84.98	76.75	68.65	44.77	T_E
E-Intersection	66.83	75.37	65.50	51.05	29.34	T_E
E-Union	**92.30**	85.32	76.90	69.04	38.02	T_E
Our Method	90.04	**87.49**	**80.02**	**74.22**	**47.32**	T_E

Fields "mAC", "mDC" and "mJI" refer mean of Accuracy, Dice coefficient and Jaccard Index on all test images respectively; "$\mathrm{mTJI}_{0.65}$" refers the thresholded Jaccard Index with threshold 0.65 and "mBR" denotes mean of Boundary Recall on all test images. Detailed description of the five metrics can be found in Section IV-A. The last column TTC lists the time consumption of training the least models for various methods; namely the time consumption of one "S-" and "SC-" baseline refers how many hours we spent to train a model for some method in our experiments, and the time consumption of ensemble-based methods and our method refers the total hours consumed by all basic models. T_F, T_M, T_{D2} and T_{D3} are roughly equivalent to 0.7, 21.5, 15 and 25.5 hours respectively in our implementation, and $T_E = 2T_F + T_M + T_{D2} + 2T_{D3} \approx 88.9h$. For rows, one of "S-" series of baselines refers using single models of some DCNN method; one of "SC-" series of baselines refers utilizing CRFs as post-processing after some DCNN method; one of "E-" series of baselines refers some combination scheme on ensemble of multiple models.

When we compare the records between Table 2 and Table 3, an obvious phenomenon is that applying CRFs after single DCNNs ("SC-" series) works better on PH2 than ISIC 2017. In our observation, this is caused by the difference of data characteristics between two datasets. On the one hand, ISIC 2017 dataset has more types of dermoscopy images than PH2, leading more difficulty and complication. On the other hand, much fuzzier lesion borders in ISIC 2017 result in

TABLE 3. Performance comparison on 15 DCNN models on PH2 [31].

Method	mAC	mDC	mJI	$\mathrm{mTJI}_{0.65}$	mBR	TTC
S-FCN8s	95.84	91.10	84.35	81.08	55.02	T_F
S-MaskRCNN	96.19	91.22	84.55	80.79	55.41	T_M
S-DeeplabV2	97.45	91.38	84.70	82.18	52.61	T_{D2}
S-DeeplabV3	98.82	91.00	84.06	81.06	49.11	T_{D3}
SC-FCN8s	95.41	92.39	86.54	83.63	64.25	T_F
SC-DeeplabV2	97.39	92.98	87.43	85.29	64.65	T_{D2}
E-Voting	97.43	91.64	85.13	82.51	54.55	T_E
E-Intersection	91.40	91.61	85.92	82.66	69.65	T_E
E-Union	**99.73**	84.47	74.47	63.25	20.66	T_E
Our Method	96.20	**94.14**	**89.20**	**89.20**	**68.10**	T_E

All notes are same as in Table 2.

that the Softmax outputs for ISIC 2017 are poorer than PH2. Although using CRFs in DCNNs is effective for PH2 dataset, its performance is still lower than the proposed method. In our opinion, the most important reason is that the proposed method can employ the performance preferences of various DCNN models and generate more diversity.

In the tables, we also present the training time consumption. Using multiple DCNN models can give us more diversity, but it inevitably needs more time consumption in both training phase and predicting phase. For time consumption in predicting phase, the heavy operation – predicting on multiple DCNN models are able to parallelize in engineering implementation. For training time, it is indeed a problem worthy of attention and discussion in the methods trying to use ensemble learning especially for our method which used many basic learners needing more time to train. Fortunately, this issue is not very critical. On the one hand, we can obtain multiple models in different epochs in an once-training. On the other hand, the proposed method is not very sensitive to the DCNN hyper-parameters, so it is not necessary to try numerous hyper-parameters in training basic DCNN learners. However, trying numerous hyper-parameters is a common practice in training a best single DCNN model, probably resulting in more time consumption. Therefore, it is hard to say which one consumes more training time. In the best situation, all single DCNN models only need once training, and the ensemble-based methods need multifold training. So, to give a reference in the best situation, we added the actual training time consumption in our implementation only for the selected models.

E. RESULTS ON 7 DCNN MODELS

In order to ease concerns about too much time consumption of the proposed method, we additionally selected 7 models in 15 selected models to observe the performance on less basic models. The evaluation results on ISIC 2017 and PH2 are shown in Table 4 and Table 5 respectively.

From these experimental results, we can conclude three points as follows. First, the performance of the proposed method is still superior to the compared methods even based on 7 DCNN models. Second, the difference between 15 models and 7 models is not significant, thus it is feasible to obtain a relatively good performance on less basic models,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3014787, IEEE Access

TABLE 4. Performance comparison on 7 DCNN models on ISIC 2017 [21].

Method	mAC	mDC	mJI	$\mathrm{mTJI}_{0.65}$	mBR	TTC
S-FCN8s	80.54	84.04	75.45	66.54	43.52	T_F
S-MaskRCNN	78.50	83.16	75.04	67.72	43.03	T_M
S-DeeplabV2	80.78	84.08	75.27	66.14	38.87	T_{D2}
S-DeeplabV3	85.55	85.70	77.51	69.46	44.85	T_{D3}
SC-FCN8s	80.50	84.06	75.53	66.63	43.40	T_F
SC-DeeplabV2	79.30	83.80	75.39	66.23	41.93	T_{D2}
E-Voting	78.53	83.70	74.98	65.32	43.14	T'_E
E-Intersection	71.58	79.15	69.81	57.50	38.22	T'_E
E-Union	**88.31**	86.62	78.68	72.35	39.67	T'_E
Our Method	85.97	**87.33**	**79.81**	**73.65**	**47.43**	T'_E

The training time consumption is $T'_E = T_F + T_{D2} + T_{D3} \approx 41.9h$.
Other notes are same as in Table 2.

TABLE 5. Performance comparison on 7 DCNN models on PH2 [31].

Method	mAC	mDC	mJI	$\mathrm{mTJI}_{0.65}$	mBR	TTC
S-FCN8s	95.84	91.10	84.35	81.08	55.02	T_F
S-MaskRCNN	96.19	91.22	84.55	80.79	55.41	T_M
S-DeeplabV2	97.45	91.38	84.70	82.18	52.61	T_{D2}
S-DeeplabV3	98.82	91.00	84.06	81.06	49.11	T_{D3}
SC-FCN8s	95.41	92.39	86.54	83.63	64.25	T_F
SC-DeeplabV2	97.39	92.98	87.43	85.29	64.65	T_{D2}
E-Voting	96.73	91.95	85.67	82.66	58.28	T'_E
E-Intersection	93.71	91.88	86.07	82.44	65.92	T'_E
E-Union	**99.15**	88.74	80.45	74.72	34.09	T'_E
Our Method	94.98	**93.98**	**89.07**	**88.43**	**70.76**	T'_E

All notes are same as in Table 4.

being able to reduce the time consumption simultaneously on both training and test phase. Third, we can generally obtain improvement through using more good models if we have enough resources, because generally more good models can bring about more diversity. In addition, these experimental results bear out the discussion about the number of DCNN models in Section III-B

F. COMPARISON WITH STATE OF THE ARTS

Studying on skin lesion segmentation is very active in recent years, and many works have been done. Here, we compare the performance between our method and some important state-of-the-art methods. According to the results on ISIC 2017 as shown in Table 6, it is clearly demonstrated that the proposed method is considerably competitive especially on the primary metric mJI. In addition, we do not list the training time consumption because the training time consumption of all compared methods has not been made public.

TABLE 6. Comparison with the state-of-the-arts on ISIC 2017 [21] dataset.

Method	mJI	mDC	mAC
Goyal, et.al [14]	79.34	87.10	94.10
FCA-Net [15]	78.65	**87.80**	**94.95**
FocusNet [54]	75.62	83.15	92.14
SkinNet [34]	76.70	85.50	93.20
DCL-PST [55]	77.70	85.60	94.00
PA-Net [35]	77.60	85.80	93.60
FrCN [36]	77.11	87.08	94.03
iMSCGnet [37]	77.75	85.83	93.58
Ours (7 models)	79.81	87.33	85.97
Ours (15 models)	**80.02**	87.49	90.04

For PH2 dataset, it is hard to compare the performance directly between previous works and the proposed method. As mentioned in Section IV-A, our method needs some samples to guide the selection of hyper-parameters, so the results may be a little bit unfair for related references if we use all images in PH2 as test like [14], [36], [37]. In our experiments, we used the validation part described in Section IV-A to guide selecting hyper-parameters and the test part to evaluate. Hence, it is not viable to compare to the previous works in a completely fair way. However, in order to reflect the research status on PH2, we still list them into Table 7 for referencing but not comparing, and do not mark the best scores in Table 7.

TABLE 7. Performance lists of the state-of-the-arts and our method on PH2.

Method	mJI	mDC	mAC	Training	Test	Validation
Goyal et al. [14]	83.90	90.70	93.80	ISIC2017	PH2	-
FrCN [36]	84.79	91.77	95.08	ISIC2017	PH2	-
iMSCGnet [37]	88.21	93.36	95.71	ISIC2017	PH2	-
Ours (7 models)	89.07	93.98	94.98	ISIC2017	$\frac{3}{4}$PH2	$\frac{1}{4}$PH2
Ours (15 models)	89.20	94.14	96.20	ISIC2017	$\frac{3}{4}$PH2	$\frac{1}{4}$PH2

Due to requiring some samples to guide the hyper-parameter selecting, only three quarters in PH2 were used to carry out the evaluation. So we list them here just for referencing but not comparing.

In addition, there is a noteworthy phenomenon in Table 2 to 6 that a method on mJI is usually better than other methods while its mAC score is relatively poor than them, although better mAC and better mJI both indicate a better performance. This phenomenon does not merely occur in our experimental results, extensively existing in many previous works such as FCA-Net [15], SkinNet [34] and iMSCGnet [37], and there has been no reasonable explanation for it in community. It is known that the metrics mJI and mAC were both introduced in the official ISIC 2017 [21] and mJI was the finally decisive metric adopted, and the official had not explained the specific reasons. It is known that different metrics have different purposes to measure different things, and it is natural to have different rank by different metrics. But it is very difficult to make clear why they have different ranks. For mAC and mJI here, it remains difficult to give a clear explanation in a mathematical manner, because the mean of ACs and the mean of JIs have different denominators. Intuitively, mJI measures the intersection over union for only lesion pixels, while mAC measures the accuracy of both lesion and background pixels. What we want to fight for lesion segmentation should focus on the lesion pixels rather than all pixels. We assume this is the reason for why mJI was specified as the final metric. And a possible reason why mAC is relatively poor is that the proportion of background pixels is greater than lesion pixels on the whole, resulting in that the change speed of mAC is slower than mJI. It is indeed a problem should be studied seriously, but discussing the rank mechanism of mAC and mJI is out of our current concentration and the scope of this paper, and may be taken as a further research topic.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3014787, IEEE Access

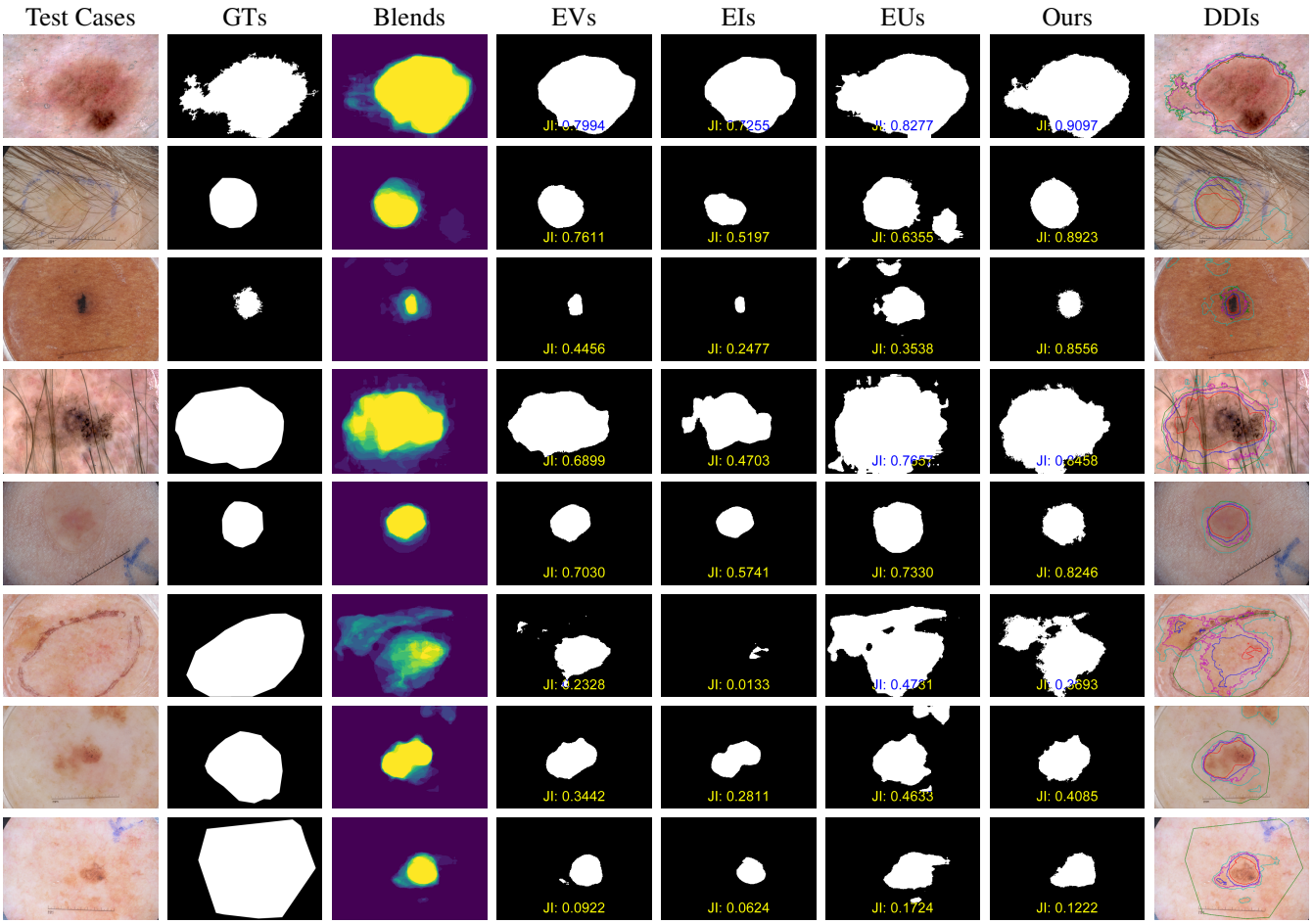


FIGURE 3. Examples of lesion segmentation on 15 DCNN models on several dermoscopy images in ISIC 2017. For columns, the first two columns list all test dermoscopy images and their corresponding ground truths (GTs); the third column shows the blended result from multiple DCNN models; the next four columns present the mask images of lesion segmentation obtained by three baseline methods namely Ensemble-Voting (EV), Ensemble-Intersection (EI) and Ensemble-Union (EU) and our proposed method respectively, and the last column shows the delineated dermoscopy images. For rows excluding headers, the first five rows are good cases and the last three rows are bad cases. The text in the mask image of lesion segmentation is the score of Jaccard index metric. For the colors of lesion boundary in last column, green and magenta represent Ground Truth and the proposed method respectively, and blue, red and cyan represent Ensemble-Voting, Ensemble-Intersection, Ensemble-Union respectively.

V. DISCUSSION

The proposed method has a drawback that its performance is limited by the lesion segmentation from DCNN models. If all original segmentation from DCNN models are very bad, the final inferring segmentation is intractable to achieve ideality. Several bad cases are shown in Figure 3 and 4. There are two possible ways to alleviate this problem. One is to introduce the artificial assistance to add a rough area as a source of the blend. The other one is to use the test image self to improve the blend, which may be a promising direction for the further works. Then another further direction of efforts is to find more method to generate the unary potential for CRF models. Next, owing to the difference of experimental implementation and dependency on multiple DCNN models affected by hyper parameters, the performance of the proposed method for skin lesion segmentation may have slight fluctuations. So we believe that there must be higher performance using the proposed method in lesion segmentation task because our hardware condition and computing capacity limited our experimental scale. Another problem worth discussing is that we need a portion of samples to guide us to find better hyper-parameters. What we want to point out is that the hyper-parameters in the proposed method are explainable and we can adjust them manually and empirically case by case, rather than find a global optimum to try to get a best score on an overall test dataset. This character is very attractive in clinical practice. In addition, blending and generating the unary potential is the key to connect the ensemble of multiple DCNN models and the fully connected CRFs, and it is valuable and promising to try any other schemes for them as a further research problem.

VI. CONCLUSION

Lesion segmentation is fundamental in analyzing the dermoscopy images. This work presents a new method to employ fully connected conditional random fields (CRFs) to infer lesion segmentation based on segmentation from multiple deep convolutional neural networks (DCNN) models.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3014787, IEEE Access

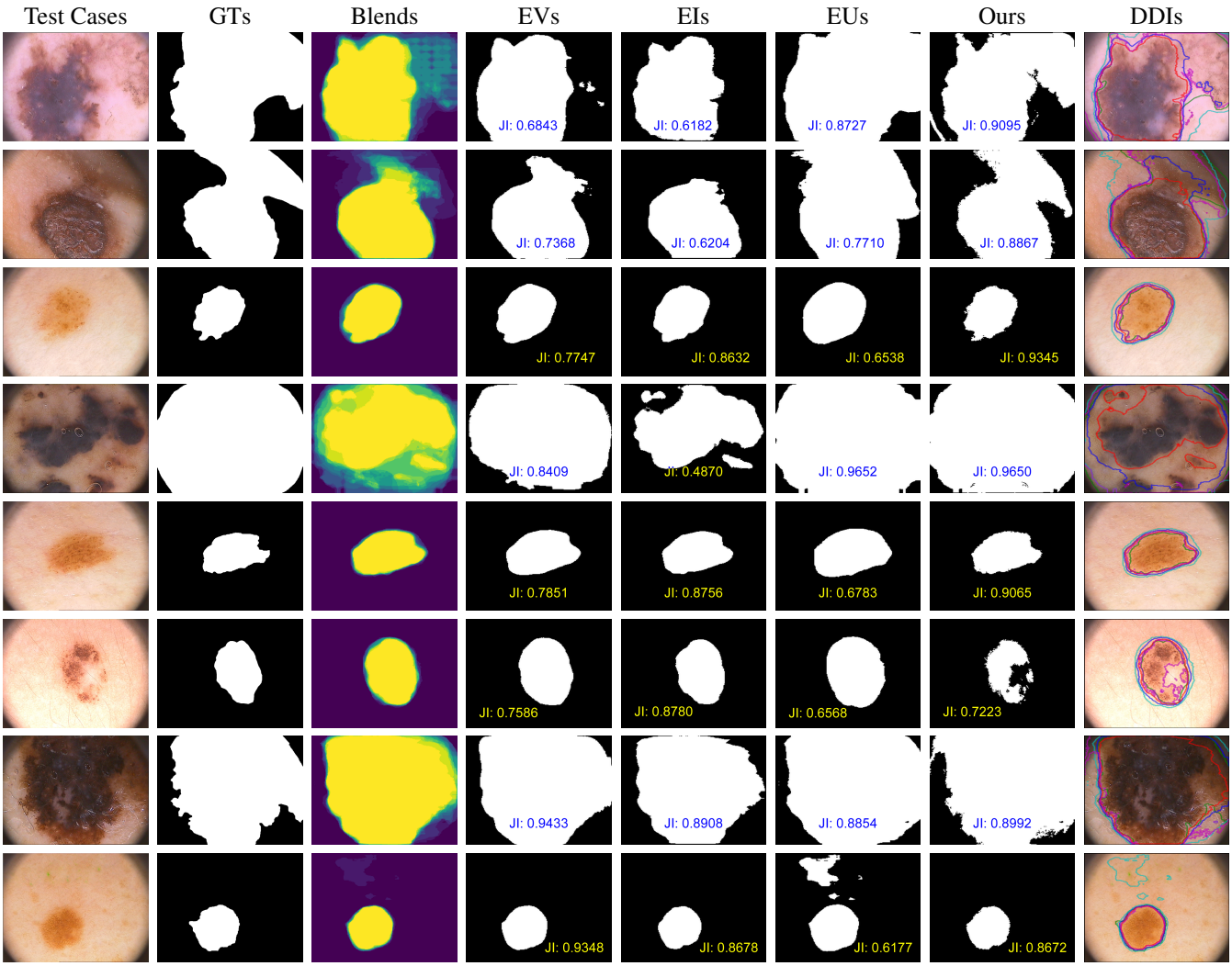


FIGURE 4. Examples of lesion segmentation on 15 DCNN models on several dermoscopy images in PH2. All notes are same as in Figure 3.

It does not only make use of the performance preferences of different DCNN models to improve the performance, but also can utilize the local information in original test images to refine the lesion boundaries. We tested the effectiveness and robustness of proposed method on the mainstream dataset ISIC 2017 and PH2, and the experimental results showed that the proposed method gained better performance than baselines. This shows that the proposed method is more inference-backed and superior than single DCNN models and simple ensemble methods, and further indicates that it can effectively generate strong generalization and great diversity from multiple DCNN models, and is able to capture the local information in fuzzy dermoscopy images and find more accurate lesion borders. Under a broader perspective out of the specific skin lesion segmentation task, we believe that our proposed method has great potential to act as a general framework to be applied in other fields.

REFERENCES

[1] M. E. Celebi, T. Mendonca, and J. Marques, "A state-of-the-art survey on lesion border detection in dermoscopy images," in *Dermoscopy Image Analysis*, ser. Digital Imaging and Computer Vision. CRC Press, 2015, pp. 97–129. [Online]. Available: http://www.crcnetbase.com/doi/book/10.1201/b19107

[2] B. Jonathan, *Diagnostic Dermoscopy : The Illustrated Guide*. John Wiley & Sons, 2012.

[3] G. Argenziano, H. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, G. Ferrara, R. Hofmann-Wellenhof, M. Landthaler, S. W. Menzies, H. Pehamberger, D. Piccolo, H. S. Rabinovitz, R. Schiffner, S. Staibano, W. Stolz, I. Bartenjev, A. Blum, R. Braun, H. Cabo, P. Carli, V. De Giorgi, M. G. Fleming, J. M. Grichnik, C. M. Grin, A. C. Halpern, R. Johr, B. Katz, R. O. Kenet, H. Kittler, J. Kreusch, J. Malvehy, G. Mazzocchetti, M. Oliviero, F. zdemir, K. Peris, R. Perotti, A. Perusquia, M. A. Pizzichetta, S. Puig, B. Rao, P. Rubegni, T. Saida, M. Scalvenzi, S. Seidenari, I. Stanganelli, M. Tanaka, K. Westerhoff, I. H. Wolf, O. Braun-Falco, H. Kerl, T. Nishikawa, K. Wolff, and A. W. Kopf, "Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet," *Journal of the American Academy of Dermatology*, vol. 48, no. 5, pp. 679–693, 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0190962203003530

[4] S. Duma, "Dermoscopy of pigmented skin lesions," in *European Handbook of Dermatological Treatments*, A. D. Katsambas, T. M. Lotti, C. Dessinioti, and A. M. D'Erme, Eds. Springer Berlin Heidelberg, 2015, pp. 1167–1177. [Online]. Available: http://link.springer.com/10.

1007/978-3-662-45139-7_116

[5] J. Malvehy, S. Puig, R. P. Braun, A. A. Marghoob, and A. W. Kopf, *Handbook of Dermosocopy*. Taylor & Francis, 2006.

[6] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 31, no. 6, pp. 362–373, 2007. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0895611107000146

[7] S. Pathan, P. C. Siddalingaswamy, and K. G. Prabhu, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—A review," *Biomedical Signal Processing & Control*, vol. 39, no. 2018, pp. 237–262, 2017.

[8] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *http://arxiv.org/abs/1606.00915*, 2016.

[10] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *http://arxiv.org/abs/1706.05587*, 2017.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *http://arxiv.org/abs/1703.06870*, 2017.

[12] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2065–2074, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7942129/

[13] Y. Yuan and Y.-C. Lo, "Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 519–526, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8239798/

[14] M. Goyal, A. Oakley, P. Bansal, D. Dancey, and M. H. Yap, "Skin lesion segmentation in dermoscopic images with ensemble deep learning methods," *IEEE Access*, vol. 8, pp. 4171–4181, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8936444/

[15] V. K. Singh, M. Abdel-Nasser, H. A. Rashwan, F. Akram, N. Pandey, A. Lalande, B. Presles, S. Romani, and D. Puig, "FCA-Net: Adversarial learning for skin lesion segmentation based on multi-scale features and factorized channel attention," *IEEE Access*, vol. 7, pp. 130 552–130 565, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8832175/

[16] R. Mishra and O. Daescu, "AlgoDerm: An end-to-end mobile application for skin lesion analysis and tracking," in *Proceedings of the 2019 International Conference on Health Informatics and Medical Systems*, 2019, pp. 3–9.

[17] Xuming He, R. Zemel, and M. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2004, pp. 695–702. [Online]. Available: http://ieeexplore.ieee.org/document/1315232/

[18] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *International Conference on Computer Vision*. IEEE, 2005, pp. 1284–1291 Vol. 2. [Online]. Available: http://ieeexplore.ieee.org/document/1544868/

[19] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., vol. 3951. Springer Berlin Heidelberg, 2006, pp. 1–15. [Online]. Available: http://link.springer.com/10.1007/11744023_1

[20] P. Kohli, L. Ladicky, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," in *computer vision and pattern recognition*. IEEE, 2008, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/4587417/

[21] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)," *http://arxiv.org/abs/1710.05006*, 2017.

[22] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv:1902.03368 [cs]*, 2019. [Online]. Available: http://arxiv.org/abs/1902.03368

[23] C. Grana, G. Pellacani, R. Cucchiara, and S. Seidenari, "A new algorithm for border description of polarized light surface microscopic images of pigmented skin lesions," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 959–964, 2003.

[24] R. Garnavi, M. Aldeen, M. E. Celebi, G. Varigos, and S. Finch, "Border detection in dermoscopy images using hybrid thresholding on optimized color channels," *Computerized Medical Imaging and Graphics*, vol. 35, no. 2, pp. 105–115, 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0895611110000819

[25] M. E. Celebi, Q. Wen, S. Hwang, H. Iyatomi, and G. Schaefer, "Lesion border detection in dermoscopy images using ensembles of thresholding methods," *Skin Research and Technology*, vol. 19, no. 1, pp. e252–e258, 2013. [Online]. Available: http://doi.wiley.com/10.1111/j.1600-0846.2012.00636.x

[26] R. Melli, C. Grana, and R. Cucchiara, "Comparison of color clustering algorithms for segmentation of dermatological images," in *Proceedings of SPIE - The International Society for Optical Engineering*, J. M. Reinhardt and J. P. W. Pluim, Eds., vol. 61443S, 2006, pp. 1–9. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.652061

[27] D. Alvarez and M. Iglesias, "k-Means clustering and ensemble of regressions: An algorithm for the ISIC 2017 skin lesion segmentation challenge," *arXiv:1702.07333 [cs]*, 2017. [Online]. Available: http://arxiv.org/abs/1702.07333

[28] M. E. Celebi, Y. A. Aslandogan, W. V. Stoecker, H. Iyatomi, H. Oka, and X. Chen, "Unsupervised border detection in dermoscopy images," *Skin Research and Technology*, vol. 13, no. 4, pp. 454–462, 2007. [Online]. Available: http://doi.wiley.com/10.1111/j.1600-0846.2007.00251.x

[29] M. Emre Celebi, H. A. Kingravi, H. Iyatomi, Y. Alp Aslandogan, W. V. Stoecker, R. H. Moss, J. M. Malters, J. M. Grichnik, A. A. Marghoob, H. S. Rabinovitz, and S. W. Menzies, "Border detection in dermoscopy images using statistical region merging," *Skin Research and Technology*, vol. 14, no. 3, pp. 347–353, 2008. [Online]. Available: http://doi.wiley.com/10.1111/j.1600-0846.2008.00301.x

[30] R. B. Oliveira, M. E. Filho, Z. Ma, J. P. Papa, A. S. Pereira, and J. M. R. Tavares, "Computational methods for the image segmentation of pigmented skin lesions: A review," *Computer Methods and Programs in Biomedicine*, vol. 131, pp. 127–141, 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169260716303418

[31] T. Mendonca, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH^2 - A dermoscopic image database for research and benchmarking," in *international conference of the ieee engineering in medicine and biology society*. IEEE, 2013, pp. 5437–5440. [Online]. Available: http://ieeexplore.ieee.org/document/6610779/

[32] D. Gutman, N. C. F. Codella, M. E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "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)," *http://arxiv.org/abs/1605.01397*, 2016.

[33] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 Dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *http://arxiv.org/abs/1803.10417*, 2018.

[34] S. Vesal, N. Ravikumar, and A. Maier, "SkinNet: A deep learning framework for skin lesion segmentation," in *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*. IEEE, 2018, pp. 1–3. [Online]. Available: https://ieeexplore.ieee.org/document/8824732/

[35] H. Wang, G. Wang, Z. Sheng, and S. Zhang, "Automated segmentation of skin lesion based on pyramid attention network," in *Machine Learning in Medical Imaging*, H.-I. Suk, M. Liu, P. Yan, and C. Lian, Eds., vol. 11861. Springer International Publishing, 2019, pp. 435–443, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-030-32692-0_50

[36] M. A. Al-masni, M. A. Al-antari, M.-T. Choi, S.-M. Han, and T.-S. Kim, "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks," *Computer Methods and Programs in Biomedicine*, vol. 162, pp. 221–231, 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169260718304267

[37] Y. Tang, Z. Fang, S. Yuan, C. Zhan, Y. Xing, J. T. Zhou, and F. Yang, "iMSCGnet: Iterative multi-scale context-guided segmentation of skin lesion in dermoscopic images," *IEEE Access*, vol. 8, pp. 39 700–39 712, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9007375/

[38] M. M. K. Sarker, H. A. Rashwan, F. Akram, S. F. Banu, A. Saleh, V. K. Singh, F. U. H. Chowdhury, S. Abdulwahab, S. Romani, P. Radeva, and D. Puig, "SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Springer International Publishing, 2018, vol. 11071, pp. 21–29, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00934-2_3

[39] N. C. F. Codella, Q.-B. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, and J. R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM Journal of Research and Development*, vol. 61, no. 4, pp. 5:1–5:15, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8030303/

[40] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *International Conference on Computer Vision*, 2015, pp. 1529–1537. [Online]. Available: http://arxiv.org/abs/1502.03240

[41] A. Arnab, S. Jayasumana, S. Zheng, and P. Torr, "Higher order conditional random fields in deep neural networks," *IEEE Access*, 2015. [Online]. Available: http://arxiv.org/abs/1511.08119

[42] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *http://arxiv.org/abs/1412.7062*, 2014.

[43] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning*, 2001, pp. 282–289.

[44] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *International Conference on Computer Vision*. IEEE, 2009, pp. 670–677. [Online]. Available: http://ieeexplore.ieee.org/document/5459175/

[45] J. Verbeek and B. Triggs, "Scene segmentation with crfs learned from partially labeled images," *neural information processing systems*, pp. 1553–1560, 2007.

[46] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *http://arxiv.org/abs/1210.5644*, 2012.

[47] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *computer vision and pattern recognition*. IEEE, 2008, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/4587799/

[48] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *International Conference on Computer Vision*. IEEE, 2015, pp. 1377–1385. [Online]. Available: http://ieeexplore.ieee.org/document/7410519/

[49] P. Krähenbühl and V. Koltun, "Parameter learning and convergent inference for dense random fields," in *International Conference on Machine Learning*, vol. 28, 2013, pp. III–513—III–521.

[50] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 530–549, 2004. [Online]. Available: http://ieeexplore.ieee.org/document/1273918/

[51] W. Benesova and M. Kottman, "Fast superpixel segmentation using morphological processing," in *Proceedinks of the International Conference on Machine Vision and Machine Learning - MVML 2014*, 2014, pp. 67–1–67–9.

[52] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv:1904.04514 [cs]*, 2019. [Online]. Available: http://arxiv.org/abs/1904.04514

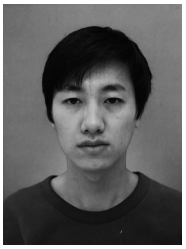
[53] Y. Qiu, X. Qin, and J. Zhang, "Low effectiveness of non-geometric-operation data augmentations for lesion segmentation with fully convolution networks," in *Proceeding of 2018 3rd IEEE International Conference of Image, Vision and Computing*. IEEE, 2018, pp. 299–305.

[54] C. Kaul, S. Manandhar, and N. Pears, "Focusnet: An attention-based fully convolutional network for medical image segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 455–458. [Online]. Available: https://ieeexplore.ieee.org/document/8759477/

[55] L. Bi, J. Kim, E. Ahn, A. Kumar, D. Feng, and M. Fulham, "Step-wise integration of deep class-specific learning for dermoscopic image segmentation," *Pattern Recognition*, vol. 85, pp. 78–89, 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0031320318302772

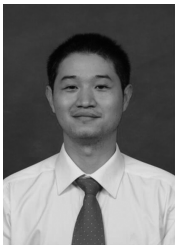
YUMING QIU received the M.S. degree from Nanjing Normal University, Nanjing, China, in 2010. He is currently pursuing the Ph.D. degree in Computer Software and Theory at Chengdu Institute of Computer Applications, University of Chinese Academy of Sciences, Chendu, China.

He was a senior engineer in Alibaba Group from 2010 to 2013, and then has been working at Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences. His research interest includes machine learning, medical knowledge computing and computer vision especially methods on medical image segmentation.



JINGYONG CAI received the B.S. degree in electronic information science and technology from Anhui Polytechnic University, Wuhu, China, in 2016 and the M.S. degree in computer and information sciences from Tokyo University of Agriculture and Technology, Tokyo, Japan, in 2019. He is currently pursuing the Ph.D. degree in computer and information sciences at Tokyo University of Agriculture and Technology, Tokyo, Japan.

His research interest includes training algorithms of machine learning, image processing and hardware optimization for artificial neural networks.



XIAOLIN QIN is currently a professor at Chinese Academy of Sciences (CAS) as well as University of CAS, China. He also serves as director of Laboratory for Automated Reasoning and Programming of CAS and Academician Office. He received his Ph.D degree in computer software and theory from Graduate University of CAS, in 2011. From May 2014 to June 2015, he was a postdoctoral fellow at Department of Computer and Information Science, Linköping University, Sweden.

He is a member of Youth Innovation Association and Western youth scholar for CAS. He is a member of TianfuTen-thousand Talents Program of Sichuan. He serves as a vice chairman for Chengdu Youth Association of Science and Technology and the technical committee of biology computation and biology information processing, Chinese Institute of Electronics.

JU ZHANG received the B.S. degree in mathematics from Sichuan University in 1983, and the M.S. degree in computer sciences from Beijing University in 1983. He received the Ph.D. degree in computer sciences from University of Rochester in 1994.

Through 2012, he had worked mostly in financial industry. Since 2012, he has been a research professor with the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences. His research interests include computer algorithms and application of artificial intelligence in medical sciences.