IJE TRANSACTIONS B: Applications Vol. 34, No. 02, (February 2021) 458-469

International Journal of Engineering

Journal Homepage: www.ije.ir

Semantic Segmentation of Lesions from Dermoscopic Images using Yolo-DeepLab Networks

F. Bagheri[a], M. J. Tarokh*[a], M. Ziaratban[b]

[1] *Department of Information Technology Engineering, Faculty of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran*
[2] *Department of Electrical Engineering, Faculty of Engineering, Golestan University, Gorgan, Iran*

PAPER INFO

Paper history:
Received 23 July 2020
Received in revised form 07 November 2020
Accepted 08 December 2020

Keywords:
Deep Learning
Deeplab3+
Semantic Segmentation
Skin Lesion
Yolov3

ABSTRACT

Accurate segmentation of lesions from dermoscopic images is very important for timely diagnosis and treatment of skin cancers. Due to the variety of shape, size, color, and location of lesions in dermoscopic images, automatic segmentation of skin lesions remains a challenge. In this study, a two-stage method is presented for the segmentation of skin lesions using Deep Learning. In the first stage, convolutional neural networks (CNNs) estimate the approximate size and location of the lesion. A sub-image around the estimated bounding box is cropped from the original image. The sub-image is resized to an image of a predefined size. In order to segment the exact area of the lesion from the normal image, other CNNs are used in the DeepLab structure. The accuracy of the normalization stage has a significant impact on the final performance. In order to increase the normalization accuracy, a combination of four networks in the structure of Yolov3 is used. Two approaches are proposed to combine the Yolov3 structures. The segmentation results of the two networks in the DeepLab v3+ structure are also combined to improve the performance of the second stage. Another challenge is the small number of training images. To overcome this problem, the data augmentation is used along with different modes of an image in each stage. In order to evaluate the proposed method, experiments are performed on the well-known ISBI 2017 dataset. Experimental results show that the proposed lesion segmentation method outperforms the state-of-the-art methods.

***doi**: 10.5829/ije.2021.34.02b.18*

1. INTRODUCTION

Nowadays, cancer is one of the most common reasons of mortality in humans worldwide. One of the most prevalent cancers is melanoma skin cancer. This disease is initiated when a specific type of skin cell named melanocyte starts to over-grow out of control [1]. Therefore, attempts to diagnose this disease in early stages are very important for more rapid treatment and increasing the chance of survival [2]. Visual inspection during laboratory assessments and medical examination of skin lesions might cause misdiagnosis due to similarities of skin lesions and normal skin tissues [3]. In the recent decade, dermatologists have begun to use an invasive imaging tool called dermoscopy which provides an enlarged image of skin lesion through polarized light [4]. It shows more details of the skin structure and improves the correctness of the diagnosis in comparison to visual observation. However, observation and verification of dermoscopy images by dermatologists is subjective, difficult, and time-consuming [4]. Thus, an automatic accurate skin lesion recognition system is very critical to support dermatologists in decision-making. One essential primary stage in any computer-based diagnostic system for detecting melanoma is automated segmentation of skin lesions [5–7]. The lesion segmentation remains a challenge due to the large variety of skin lesions in color, shape, texture, location, and size in dermoscopy images. In addition, low contrast borders between lesions and surrounding tissues, existence of ruler sign, blood vessels, hair, air bubbles, and changes of brightness are amongst the barriers to accurate segmentation [8].

Generally, there are various methods for image segmentation, such as methods based on edge detection, thresholding, region detection, feature clustering [9-10],

*Corresponding Author Email: *mjtarokh@email.kntu.ac.ir* (M. J. Tarokh)

Please cite this article as: F. Bagheri, M. J. Tarokh, M. Ziaratban, Semantic Segmentation of Lesions from Dermoscopic Images using Yolo-DeepLab Networks, International Journal of Engineering, B: Applications Vol. 34, No. 02, (2021) 458-469

as well as supervised methods such as those based on deep neural networks [11]. Various supervised methods have been used for segmentation of skin lesions such as decision tree, support vector machine, and neural networks. Indeed, these methods use low-level features [8]. Recently, image segmentation based on deep learning has become one of the main image segmentation methods. Deep learning based on Convolutional Neural Networks (CNNs) is a powerful method. CNN is capable to exclude more appropriate features in comparison to the features extracted by conventional methods [12–14]. The first application of CNN-based method in semantic segmentation was presented by Ciregan et al. [15, 16].

In 2017, Burdick et al. investigated the effect of segmentation boundary expansion involving pixels around the target lesion. They used ISBI 2016 Challenge dataset [17] to evaluate the experiments. They found the preprocessing techniques that create bounds larger than the actual lesion can potentially improve the performance of the classifier [18]. You et al., in 2017 prersented a two-stage method to segment and classify skin lesions using fully convolutional residual network (FCRN). They examined their method on the ISBI 2016 dataset, and achieved the accuracy of 94.9% [19]. Yuan et al. in 2017 presented a method for segmentation of skin lesions using deep fully convolutional networks (FCN). They employed the Jaccard distance as a loss function of the FCN. They evaluated their method on ISBI 2016 and PH2 datasets, and reached the accuracy of 95.5%, and 93.8%, respectively [20].

In 2017, Lin et al. comprised two skin lesion segmentation approaches, C-means clustering and U-Net-based histogram equalization. Their work was evaluated on the ISBI 2017 dataset. The clustering technique achieved a dice index of 61% and the U-Net method results in the accuracy of 77% [21]. Li et al., proposed another approach based on a lesion index calculation unit (LICU) and multi-scale fully-convolutional residual networks. They evaluated their approach on the ISBI 2017 dataset and achieved 71.8% in Jaccard index [22]. Bi et al. followed the FCN architecture to add convolutional and deconvolutional layers, which upsample the feature maps derived from Resnet to output the score mask. They achieved Jaccard index of 76.1% on ISBI 2017 dataset [23]. Yuan and Lo proposed a method for segmentation of skin lesions based on convolutional-deconvolutional neural networks (CDNN). They trained their model through various color spaces. Their method was ranked first in the ISBI 2017 lesion segmentation challenge with a Jaccard index of 76.5% [24]. Al-Masni et al. (2018) conducted a study on segmentation of skin lesions and designed a full resolution convolutional network. They performed the examinations on ISBI 2017 and PH2 datasets, and achieved 77.11% and 84.79% by Jaccard criteria, respectively [8].

Baghersalimi et al. presented a full convolutional neural network, DermoNet, to segment skin lesions. In DermoNet, subsequent layers could reuse the information extracted from previous layers. The Jaccard values of DermoNet on ISBI 2017 dataset was 78.3% [25].

Hasan et al., proposed a Dermoscopic Skin Network (DSNet) to segment skin lesions. To reduce the number of network parameters, depth-wise separable convolution layers were used in their network. They achieved the Jaccard value of 77.5% on the ISBI 2017 dataset [26].

Tang et al. developed a skin lesion segmentation method based on separable U-Net and took advantage of the separable convolutional block and the U-Net architectures, simultaneously. The Jaccard index of their method on ISBI 2017 dataset was 79.26% [27].

For many applications, both local and global information on lesions and normal tissues are required to increase the segmentation accuracy. Many researchers have used multi-flow architectures to combine local and global information [28]. Chen et al. used three CNNs which receive information on lesion from different aspects as input. The features extracted from each CNN were concatenated as output, constituting the final feature vector [29]. Similarly, Kawaraha and Hamarneh introduced a method for classifying skin lesions using multi-flow CNN. In this method, the flows worked on various versions of resolution of the image [30].

These studies indicated that the combination of several CNNs with various details can improve the final performance. In this study, combinations of CNNs have been used to improve the accuracy of each stage of the proposed method.

Our contributions in this work are as follows: Using normalization stage before the segmentation stage.

- Using state-of-the-art CNNs in both normalization and segmentation stages.
- Combination of Yolo networks to improve the accuracy of the normalization stage.
- Proposing a novel combined structure for Yolov3 to combine the results of Yolo networks.
- Combination of DeepLab v3+ networks to improve the segmentation accuracy.
- Using various modes of images to overcome large variety of lesions and low number of training images.

2. MATERIALS AND METHODS

2. 1. Dataset The proposed segmentation method was evaluated on a well-known and open ISBI 2017 challenge dataset. This dataset was prepared by the International Skin Imaging Collaboration (ISIC) archive [31], and was presented online [32]. This dataset consists of 8-bit RGB dermoscopy images of sizes from 540×722 to 4499×6748 pixels. Out of 2750 images, 2000, 150, and

600 images have been categorized for training, validation, and test, respectively.

2. 2. Proposed Method The purpose of lesion segmentation is extraction of skin lesions from dermoscopy images in order to help disease diagnosis. In recent years, various methods such as U-Net and FCN have been used for medical image segmentation. FCN and U-Net, as well as other single-stage methods are sensitive to the lesion size. Very large and very small lesions decrease the accuracy of single-stage segmentation methods. In addition, various locations of lesions in images increase the complexity of networks and reduce the performance. In our experiments by using single-stage methods, we observed that a significant number of inaccurate segmentation occurred in two categories of skin images: the images in which the lesion was very big or very small, and the images in which the lesion was not in the center. Therefore, it is better to add a stage before the segmentation stage to normalize the size and location of lesions in images. This will reduce complexity of the network training in the segmentation stage. The proposed method consists of two stages of normalization and segmentation. The normalization stage estimates the approximate size and location of lesions. This stage yields normal images in which the lesions have similar sizes and are placed in the center. In the following stage, lesions will be more accurately segmented from the normalized images compared to the original input images. The overall framework of the proposed lesion segmentation method is illustrated in Figure 1.

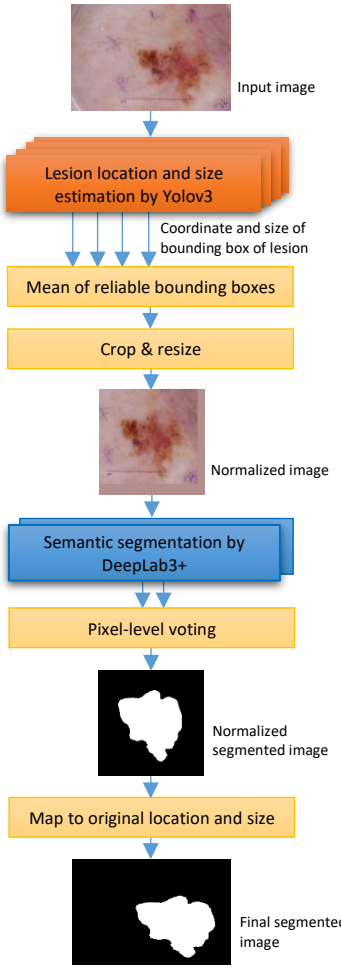

Figure 1. The overall framework of the proposed method

2. 3. Size and Location Normalization of Lesions Any error in the normalization stage leads to high costs in performance of the segmentation stage. Hence, the accuracy of the normalization stage is very important. One of the possible errors in the normalization stage occurs when the cropped image does not include any part of the lesion. It means that some pixels of the skin lesion do not exist in the output image of the normalization stage. For these images, before entering the segmentation stage, a part of the lesion is missed. Therefore, the high accuracy in the normalization stage is very important. If the accuracy of the normalization is not large enough, it might cause reduction in the final accuracy compared to single-stage segmentation methods (without normalization stage). In the proposed method, convolutional neural networks with definite structures presented for object detection will be used as the normalization stage. CNNs are very competent and practical in applications of object detection and classification. Various common deep networks based on CNN were being proposed and used for the above applications [33].

Object detection networks such as R-CNN [34], Fast R-CNN [35], and Faster R-CNN [36] combine convolutional networks with region proposal networks. Methods of Single Shot multi-box Detector (SSD) [37] and You Only Look Once (Yolo) [38] detect objects only in one convolutional process without region proposals [33]. Amongst various methods for object detection, Faster R-CNN usually shows appropriate accuracy, but its computational cost is very high as compared to Yolo [39]. The accuracy of Faster R-CNN might be higher than that of Yolo in many applications of detection, but the speed of Yolo is far greater than that of Faster R-CNN [40]. On the other hand, in implementations, the score value for Faster R-CNN is usually very close to 1, even in cases of misdetection. However, the score value in Yolo is usually proportional to the correctness of detection. In other words, Yolo presents lower scores for the samples which cannot be detectted definitely. As pointed above, one of the approaches that can be implemented to enhance the accuracy of the detection is the combination of the results of several detectors. This ability of Yolo that presents the scores proportional to the detection accuracy is very important and applicable in combining detectors. In this paper, due to the ability of

Yolo for combination as well as its reasonable accuracy and speed appropriateness, a combination of several Yolo detectors is used in the normalization stage.

Yolo is a CNN-based object detection algorithm which divides the image into several sub-regions. Then, it predicts bounding boxes and class probabilities for each of the sub-regions. Yolo algorithm predicts two values for any anchor box: one of them is the class probabilities, and the other one is the bounding box characteristics [38].

To improve the performance of Yolov1, a second version, Yolov2, was developed. Yolov2 uses an identity mapping and concatenating feature maps from a previous layer to capture low-level features [41].

In Yolov3, the feature map are taken from one of the last layers of a pre-trained network. The feature map is upsampled by 2. Another feature map from earlier in the network is concatenated with the upsampled features. This allows Yolov3 to get both meaningful semantic and finer-grained information from the feature maps. Some convolutional layers process these combined feature maps [42]. Yolov3 has two outputs in scales of 1 and 2 that are used in the training phase. We use the latest version of Yolo, Yolov3, in our experiments.

2. 4. Combining Networks in The First Stage

In order to improve the performance of the normalization stage, a combination of several networks is used with the overall structure of Yolov3 containing different pre-trained networks. Several pre-trained networks exist with each possessing specific characteristics. Indeed, if an inappropriate pre-trained network is used for an application, a suitable efficiency will not be achieved. The difference of each pre-trained network is due to the number of layers, the number of convolutional filters, and their complexities [43]. Using transfer learning concept, the weights of a network trained based on a specific dataset such as ImageNet, can be used and trained again by a different dataset to be used in another application. Utilizing learned weights in pre-trained networks, the model can be trained at a higher speed based on the new dataset. The first layers of pre-trained networks were trained to detect primary and main features of an image such as borders, corners, round formats, basic geometric shapes, and colors [33]. In this study, various pre-trained networks such as VGG [44], AlexNet [13], Resnet [45], GoogleNet [46], and Inception [47] are used as the basis networks of the Yolov3 structures. The constructed Yolo networks are further investigated, and those with a higher performance in the validation set are selected for combining the results in the normalization stage. As shown in Figure 1, the outputs of Yolo networks include the coordinates and size of the estimated bounding boxes of detected lesions. In the proposed method, two approaches to combine the Yolo networks are introduced.

In the first approach, the outputs of Yolo networks are combined by averaging the coordinates and size of the bounding boxes obtained by each Yolo network. Meanwhile, the outputs of some Yolo detectors might have a low score. Therefore, for an input image, amongst all N outputs of detectors, the M ($M \leq N$) outputs with the largest score are used to combine and determine the final bounding box.

To improve the performance of each Yolo network in the normalization stage, for an input image, totally four modes are considered and applied to the input of each Yolo network, as follows:

1. Input image
2. Horizontal flip of input image
3. Vertical flip of input image
4. Input image with 180 degrees rotation

Four corresponding outputs will be calculated by applying their inverse transforms. Thus, for each input image, each detector makes four bounding boxes with corresponding scores in the output. By combining N detectors, totally $4N$ bounding boxes will be achieved. $3N$ out of $4N$ bounding boxes with the largest score will be considered for averaging and determining the final bounding box as follows:

$$x = \frac{\sum_{i=1}^{N}\sum_{j=1}^{4}\alpha_{ij}x_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{4}\alpha_{ij}} \quad (1)$$

$$\alpha_{ij} = \begin{cases} 1 & Score_{ij} \text{ is in the set of } 3N \text{ largest scores} \\ 0 & \text{elsewhere} \end{cases} \quad (2)$$

where $Score_{ij}$ and x_{ij} are the estimation score and the x coordinate of upper left corner of the bounding box of the j-th mode of the input image estimated by the i-th network, respectively. y, w, and h of the final bounding box are calculated in a similar way.

In the second approach, the trained Yolo networks and an additional convolutional network are combined to construct a novel combined Yolo structure as shown in Figure 2. In this figure, the yellow boxes are the Yolov3 networks which are trained separately. The weights of layers of these networks are freezed during the training of the combined Yolo structure. To have better results, the outputs of trained Yolo networks should be combined with respect to the content of the input image. Hence, an additional convolutional network (green boxes) are employed to extract useful features from the input image to be used in the combination procedure. Briefly, in the second approach, the combined Yolov3 structure learns how to combine the outputs of frozen Yolo networks according to features of input images. Four parameters of convolution layers in Figure 2 are respectively the filter size, number of filters, stride, and the zero padding size.

To avoid missing any parts of lesion in the normal image, it is better to consider a margin around the estimated bounding box before the image cropping. To do this, the estimated bounding box is extended on both

sides in both vertical and horizontal directions. The extended box is cropped from the input image and is sent to the segmentation stage. The extent of margin around the estimated bounding box is considered as 30% of the size of bounding box in each direction. The bounding boxes estimated by Yolov3 structures based on four pre-trained networks, final estimated bounding box, extended box, and the normalized image for an instance image are displayed in Figure 3.

The bounding boxes estimated by Yolov3 structures based on four pre-trained networks, final estimated bounding box, extended box, and the normalized image for an instance image are displayed in Figure 3. The red, green, blue, and yellow colours in Figure 3(a-d) are related to the first, second, third, and forth modes of the input image. Green rectangle in Figure 3(e) is the correct bounding box of the lesion. The red and blue rectangles are the bounding boxes estimated by the first and the second Yolo results combination approaches, respectively. The red rectangles with dashed lines are the extended box around the lesion estimated by the first combination approach. The extended box is cropped and resized to construct the normal image (Figure 3(f)).

2. 5. Segmentation Stage Various methods and networks are used for semantic segmentation of images in different applications. One of the novel structures is the DeepLab structure [48].

DeepLab is a model of deep learning for segmentation of images. In general, the DeepLab architecture is based on a combination of two common Spatial Pyramid Pooling and Encoder-decoder networks architectures [49].

Different DeepLab structures have been proposed over time. DeepLab v1 [48], DeepLab v2 [50], DeepLab v3 [51], and DeepLab v3+ [49] are the various structures of DeepLab. DeepLab v1 uses atrous convolution to control the resolution at which feature maps are

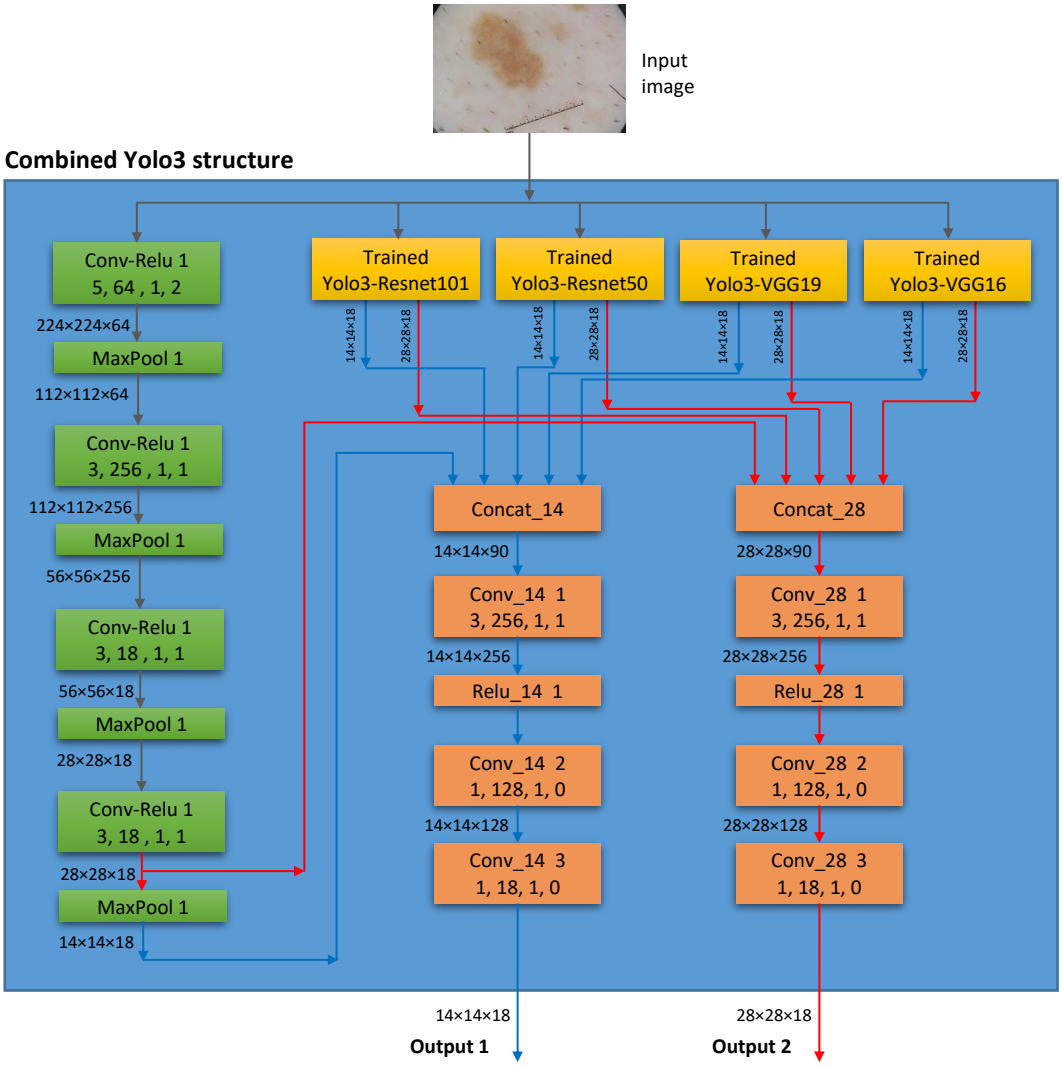

Figure 2. The proposed combined Yolov3 structure

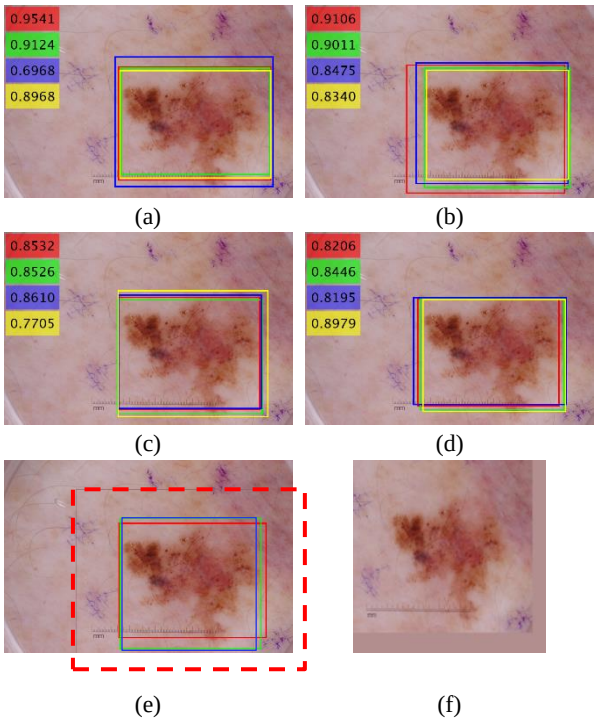

Figure 3. Result of normalization stage for a sample image. (a), (b), (c), and (d) are the bounding boxes estimated by Yolov3 structures based on Resnet50, Resnet101, VGG16, and VGG19, respectively. (e) The correct bounding box is shown with a green box. The red and blue rectangles are the bounding boxes estimated by the first and the second Yolo results combination approaches, respectively. The red dashed rectangle is the extended box around the lesion estimated by the first combination approach. (f) Normalized image.

computed [48]. In DeepLab v2, by using atrous spatial pyramid pooling (ASPP), objects are segmented on multiple scales with effective fields-of-view and filters at multiple sampling rates [50]. To capture more information, DeepLab v3 augments the ASPP module via image-level feature.

It also includes batch normalization parameters. DeepLab v3+ includes an effective decoder module to improve the segmentation results [49].

We use DeepLab3+ structure [49] in the segmentation stage of our proposed method. To improve the performance of our segmentation stage, totally eight different modes of input image are considered as follows:

- Input image, Horizontal and Vertical flips of the image, the image rotated by -45, 45, 90, 180, and 270 degrees.

The output of each input mode is rotated or is flipped back to the original mode. The final result is obtained by combining the outputs. The final output of the combination is a binary image in which, a pixel is considered as lesion if the corresponding pixel in at least n out of m output images are recognized as lesion.

2. 6. Combining Networks in the Second Stage Similar to the normalization stage, in order to increase the accuracy, combinational results of some networks can be used in the segmentation stage. In our experiments, combination of segmentation results of VGG19 and Resnet50 networks in DeepLab3+ structure has been used to improve the overall lesion segmentation performance.

2. 7. Evaluation Metrics A commonly used metric to evaluate object detection methods is the mean average precision (mAP). In our experiments, to evaluate performance of the normalization stage in more details, a metric named $BoxIOU$ is defined as the intersection over union (IOU) of the estimated bounding box with the correct bounding box of lesions in the ground truth:

$$BoxIOU = \frac{TP_{Box}}{TP_{Box}+FN_{Box}+FP_{Box}} \quad (3)$$

For evaluating semantic segmentation methods, the following metrics have been used in the literature. Sensitivity (SEN) represents the rate of pixels of skin lesion correctly detected. On the other hand, specificity (SPE) is the rate of pixels of non-skin lesions classified correctly [52]. The Jaccard index (JAC) is an intersection over union (IOU) of the result mask with the ground truth mask [53]. Index of Dice (DIC) measures the similarity of classified skin lesions through ground truth [54]. Accuracy (ACC) shows the overall performance of segmentation [53]. The Matthew correlation coefficient (MCC) measures the correlation between the segmented and annotated pixels. MCC returns values in a range of [−1 +1] [53]. All these criteria are computed from the confusion matrix elements as follows [53]:

$$SEN = \frac{TP}{TP+FN} \quad (4)$$

$$SPE = \frac{TN}{TN+FP} \quad (5)$$

$$JAC = \frac{TP}{TP+FN+FP} \quad (6)$$

$$SEN = \frac{2.TP}{(2.TP)+FP+FN} \quad (7)$$

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

$$MCC = \frac{TP.TN-FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}} \quad (9)$$

2. 8. Results In our experiments, cropped images in the normalization stage were resized to 448×448 pixels. Due to the hardware limitation, the mini-batch size was set to 8 samples. Experiments were performed by using 6GB NVIDIA GeForce RTX2060 GPU.

Table 1 reports the values of the mean average precision (mAP) and $BoxIOU$ for each of the pre-trained networks using the Yolov3 structure in the normalization stage. The Jaccard index of final semantic segmentation is the main factor for comparison of different networks [32]. The values of this metric are reported in the last column of Table 1. The DeepLab3+ structure based on Resnet50 was used for segmentation in this table.

In the segmentation stage, two pre-trained networks, VGG19 and Resent50 were used in the DeepLab3+ structure. For an input image, by considering eight input modes and two segmentation structures, totally 16 output images were achieved. In the final output binary image, I_{BW}, a pixel was considered as lesion if at least it was recognized as lesion in 6 out of 16 output images as follows:

$$p_r\,(x\,,y) = \frac{1}{16}\left[\sum_{i=1}^{2}\sum_{j=1}^{7} o_{ij}\,(x,y)\right] \quad (10)$$

$$I_{BW}\,(x\,,y) = \begin{cases} 1 & p_r(x,y) \geq P_0 \\ 0 & else \end{cases} \quad (11)$$

where o_{ij} is the binary output image of i-th network corresponding to the j-th mode of the input image and $P_0 = \frac{6}{16}$. Table 2 provides the results of various combinations of networks in the normalization and segmentation stages. A comparison among various methods based on 7 evaluation metrics are given in Table 3. In the proposed method I, Yolov3 structure based on VGG19, and DeepLab3+ structure based on Resnet50 were used in the normalization and segmentation stages, respectively. The first and second combination approaches were employed in the normalization stages of the proposed method II and the proposed method III, respectively. The segmentation stages of the proposed method II and proposed method III consisted of the combination of two DeepLab3+ structures based on VGG19, and Resnet50.

TABLE 1. Performance of different Yolov3 structures based on various pre-trained networks

Backbone network of the Yolov3	*mAP* (%)	Mean BoxIOU (%)	Overall segmentation Jaccard (%)
Vgg 19	91.36	79.62	**79.12**
Resnet 101	90.55	77.97	78.92
Vgg 16	91.85	79.03	78.79
Resnet 50	90.68	78.93	78.77
Resnet 18	90.87	78.04	78.43
Densenet 201	90.68	78.86	78.40
Mobilenet v2	88.99	78.00	78.04
Shufflenet	89.49	77.52	78.04
Alexnet	88.87	75.81	77.75
Googlenet	84.69	74.06	76.31
Squeezenet	59.48	54.98	59.26
Xception	51.57	48.46	54.21
Inception v3	45.29	44.46	50.15

TABLE 2. Performance of different combinations of Yolov3 structures in normalization stage and different combinations of DeepLab3+ structures in segmentation stage

Normalization stage										Segmentation stage		
Vgg 19	Vgg 16	Resnet 50	Resnet 101	Resnet 18	Densenet 201	1st combination approach	2nd combination approach	*mAP* (%)	Mean BoxIOU (%)	Resnet 50	Vgg 19	Overall Jaccard (%)
*	*					*		92.25	79.84	*		79.38
*		*				*		91.68	79.61	*		79.39
*	*	*				*		91.67	79.74	*		79.48
*	*	*	*			*		92.20	80.09	*		79.53
*	*	*	*	*		*		93.09	80.24	*		79.52
*	*	*	*	*	*	*		92.92	80.40	*		79.48
*	*	*	*			*		92.20	80.09		*	79.49
*	*	*	*			*		92.20	80.09	*	*	79.77
*	*	*	*				*	92.29	80.31	*	*	**79.96**

3. DISCUSSION

In this paper, a method based on deep learning was proposed to segment lesions from dermoscopic images. Deep neural networks require many training images due to a large number of trainable parameters.

In applications for which enough training images are not available, two general techniques are used to compensate the lack of enough training data: data augmentation and transform learning. In this paper, for the data augmentation, rotation, horizontal, and vertical flips, image resizing with the ratio between 0.8 and 1.2, and brightness alteration were randomly applied to the training images and the augmented training set consisted of 8000 images.

In the proposed method, by adding the normalization stage prior to the segmentation stage, the inputs of the segmentation stage contained normalized lesions with far fewer varieties in size and location. This caused reduction in complexity of the training procedure in the segmentation stage and improved the segmentation performance. As can be observed in Table 1, the use of

TABLE 3. Comparison among various methods based on different metrics

	SEN	*SPE*	*ACC*	*MCC*	*AUC*	*DIC*	*JAC*
Yuan et al. [24]	82.50	97.50	93.40	-	-	84.90	76.50
Li et al. [22]	82.00	97.80	93.20	-	-	84.70	76.20
Bi et al. [23]	82.20	**98.50**	93.40	-	-	84.40	76.00
Lin et al. [21]	-	-	-	-	-	77.00	62.00
Al-masni et al. [8]	85.40	96.69	94.03	83.22	91.04	87.08	77.11
Baghersalimi et al. [25]	-	-	-	-	-	-	78.30
Tang et al. [27]	**89.53**	96.32	**94.31**	-	-	86.93	79.26
Hasan et al. [26]	87.5	95.5	-	-	-	-	77.5
Proposed method I	88.90	95.47	93.94	83.53	92.17	87.01	79.12
Proposed method II	89.07	96.01	94.26	84.19	92.54	87.46	79.77
Proposed method III	89.21	96.08	94.29	**84.35**	**92.61**	**87.57**	**79.96**

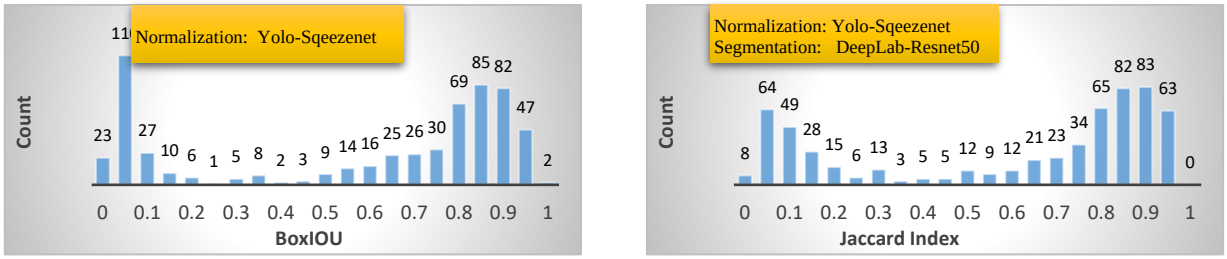

Figure 4. Distributions of BoxIOU (Detection Jaccard index) and overall Jaccard index obtained by Yolo-Squeeznet and DeepLab-Resnet50 respectively in the normalization and segmentation stages

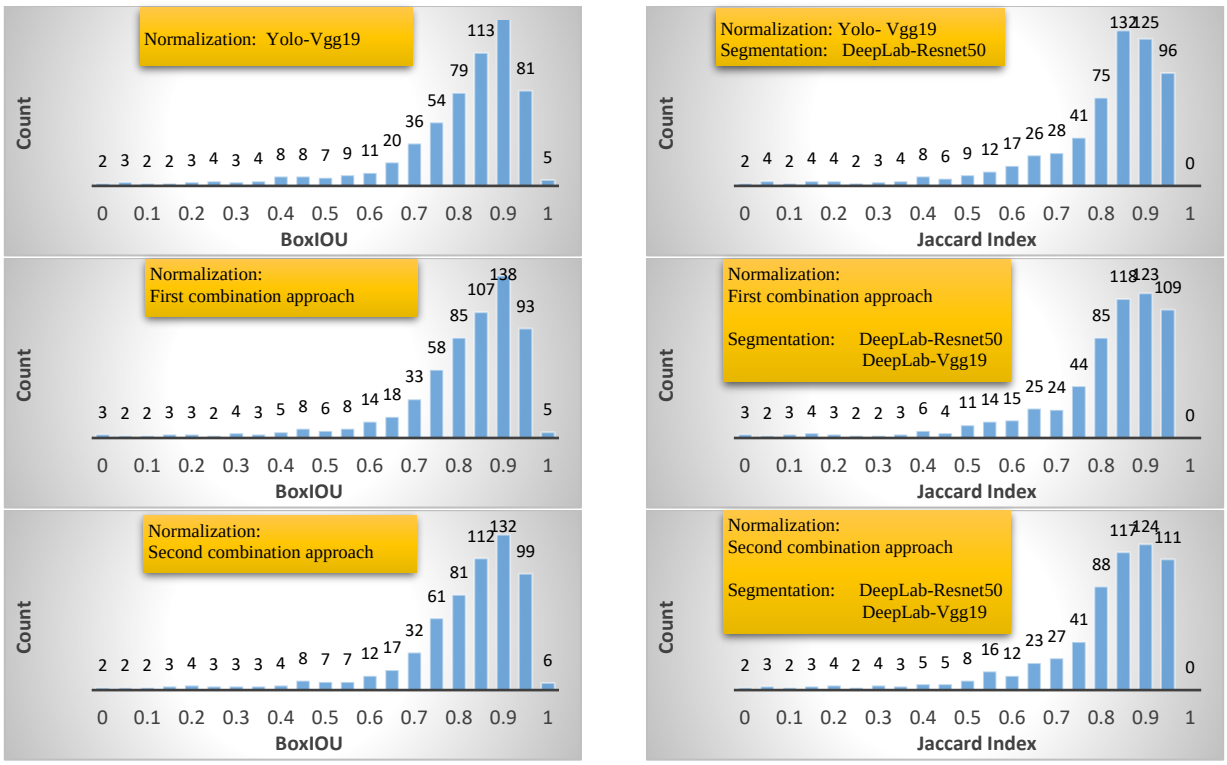

Figure 5. Distributions of *BoxIOU* (Detection Jaccard index) and overall Jaccard index obtained by different combinations of networks in the normalization and segmentation stages

inappropriate pre-trained networks in the normalization stage considerably reduced the overall performance.

The mAP values of the detection stage associated with some pre-trained networks such as Sqeezenet, Xception, and Inception v3 were obtained lower than 60%. The reason is that these networks had been particularly trained and optimized for mobile applications [39].

Further, Table 1 indicates that the final segmentation performance is proportional to the performance of the normalization stage. In other words, in cases where the value of mAP was achieved considerably larger than other cases, the value of final Jaccard was definitely greater.

To illustrate the effect of the normalization stage on the performance of our overall lesion segmentation method, distributions of $BoxIOU$ and overall Jaccard index method have been shown in Figures 4 and 5. In each row of these figures, similarity between distributions of $BoxIOU$ and overall Jaccard index demonstrates that the performance of the overall segmentation is highly affected by the performance of the detection in the normalization stage.

The use of the Yolo structure in the normalization stage made it possible to apply valid score values of the detection for combining the outputs of several networks. The results in Table 2 indicate that by combining four networks of Yolov3 based on the Resent and VGG networks, the value of mAP in the normalization stage as well as the final Jaccard index is increased.

The left and the right images in Figure 6 are respectively the results of the normalization stage and the final segmentation results of four difficult sample images. The proposed method could not correctly segment the lesions in these images and their Jaccard values have been obtained lower than 20%. As can be observed in the left column of Figure 6, the main reason of low segmentation accuracy of these images is that the proposed method could not accurately detect the lesion area in the normalization stage.

In Figures 6 and 7, rectangles with solid red and green lines are the estimated and the correct bounding boxes, respectively. The red rectangles with dashed lines are the extended estimated bounding boxes, which have been cropped and resized to enter the segmentation stage.

From the images in the right side, the green, red, yellow, and black areas respectively represent TP, FP, FN, and TN of the confusion matrix. In the first and second rows of Figure 6, the detected lesions were much wider than the correct lesions. In the third and fourth rows, the lesions have been detected smaller than the correct ones. As can be observed, segmentation of lesions in these images are very difficult even for experts.

On the other hand, four difficult images for which the Jaccard values have been obtained greater than 85% are illustrated in Figure 7.

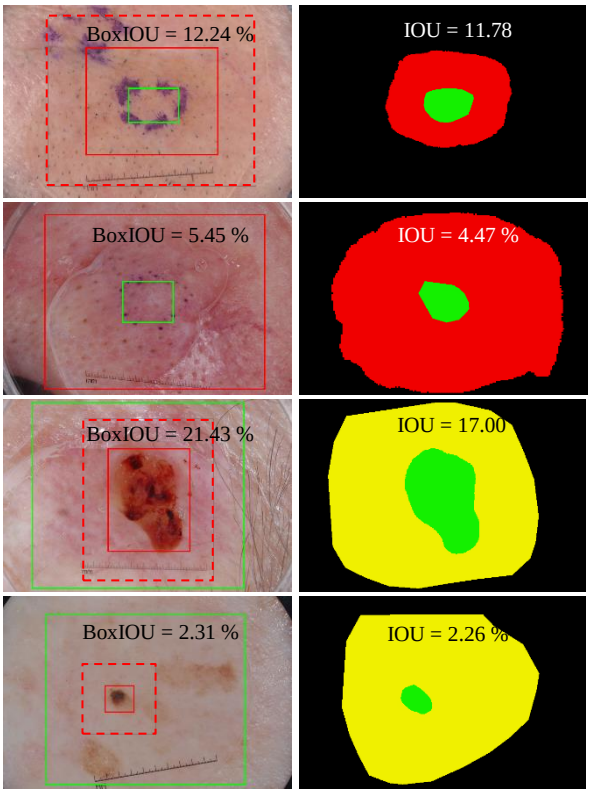

Figure 6. Four difficult sample images that have not been accurately normalized and segmented

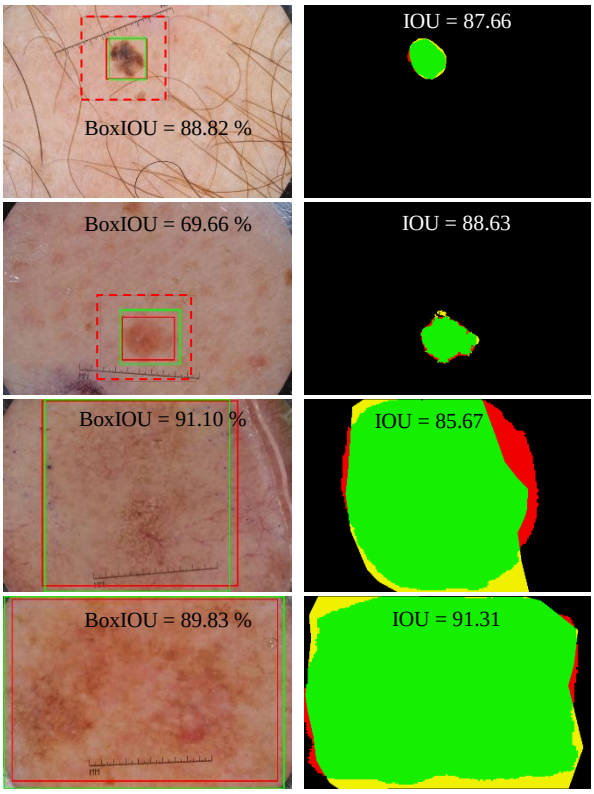

Figure 7. Four difficult sample images that have been accurately normalized and segmented

In these sample images, very small and very big lesions, lesions with low-contrast boundaries, various lesion locations, and existence of hairs and regions similar to lesions are the main challenges. However, the proposed method accurately detected and segmented the lesions.

The best results were obtained by using our combined Yolov3 structure in the normalization stage. The second combination approach in the normalization stage performed better compared to the first approach. The reason is that in the second approach, the combined structure was trained to combine the results of the Yolo networks. While in the first approach, the combination was performed without any learnable parameters.

4. CONCLUSION

Developing a highly accurate lesion segmentation system considerably helps dermatologists to diagnose skin cancer in a timely and correct manner. In this paper, a two-stage model was presented to improve the performance of skin lesion segmentation. In the proposed method, the images entered the normalization stage, in which the variety of sizes and locations of the lesions in the input images were reduced. A novel combined Yolov3 structure was proposed to combine results of four Yolov3 networks. The output of the normalization stage was an image, in which the lesion was approximately located in the centre and had a predefined size. The normalized images in the first stage were imported into the second stage. The segmentation stage consisted of a combination of two CNNs in the DeepLab3+ structure.

The main reason of applying the normalization stage before the segmentation part was that the segmentation methods are generally sensitive to the size and location of objects. The varieties of sizes and locations of objects in images complicate the training of the model. Normalization of the images greatly improved the performance of the proposed lesion segmentation method.

5. REFERENCES

1. National Cancer Institute, "SEER Cancer Stat Facts: Melanoma of the Skin", (2017). Retrieved from Bethesda, MD: https://seer. cancer. gov/statfacts/html/melan. html.
2. Balch, C. M., Gershenwald, J. E., Soong, S.-J., Thompson, J. F., Atkins, M. B., Byrd, D. R., Buzaid, A. C., and Al., E., "Final version of 2009 AJCC melanoma staging and classification", ***Journal of Clinical Oncology***, Vol. 27, No. 36, (2009), 6199–6206. doi:10.1200/JCO.2009.23.4799
3. Vestergaard, M. E., Macaskill, P. H. P. M., Holt, P. E., and Menzies, S. W., "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting", ***British Journal of Dermatology***, Vol. 159 No. (3), (2008), 669–676. doi 10.1111/j.1365-2133.2008.08713.x
4. Binder, M., Schwarz, M., Winkler, A., Steiner, A., Kaider, A., and Wolff, K. H., "a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists", ***Arch Dermatol***, Vol. 131, No. 3, (1995), 286–291. doi: 10.1001/archderm.1995.01690150050011
5. Celebi, M. E., Iyatomi, H., Schaefer, G., and Stoecker, W. V., "Lesion border detection in dermoscopy images", ***Computerized Medical Imaging and Graphics***, Vol. 33, No. 2, (2009), 148–153. doi:10.1016/j.compmedimag.2008.11.002
6. Ganster, H., Pinz, P., Rohrer, R., Wildling, E., Binder, M., and Kittler, H., "Automated melanoma recognition", ***IEEE Transactions on Medical Imaging***, Vol. 20, No. 3, (2001), 233–239. doi:10.1109/42.918473
7. Celebi, M. E., Wen, Q., Iyatomi, H., Shimizu, K., Zhou, H., and Schaefer, G., "A state-of-the-art survey on lesion border detection in dermoscopy images", ***Dermoscopy Image Analysis***, Vol. 10, (2015), 97–129. doi:10.1201/B19107-5
8. Al-masni, M. A., Al-antari, M. A., Choi, M., Han, S., and Kim, T., "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks", ***Computer Methods and Programs in Biomedicine***, Vol. 162, (2018), 221–231. doi:10.1016/j.cmpb.2018.05.027
9. Hassanpour, H., and Yousefian, H., "An improved pixon-based approach for image segmentation", ***International Journal of Engineering, Transactions A: Basics***, Vol. 24, No. 1, (2011), 25–35. Retrieved from http://www.ije.ir/article_71880_c9bb4bcef741659a548a0979992905c5.pdf
10. Nikbakhsh, N., Baleghi Damavandi, Y., and Agahi, H., "Plant classification in images of natural scenes using segmentations fusion", ***International Journal of Engineering Transactions C: Aspects***, Vol. 33, No. 9, (2020), 1743–1750. doi:10.5829/ije.2020.33.09c.07
11. Liu, X., Deng, Z., and Yang, Y., "Recent progress in semantic image segmentation", ***Artificial Intelligence Review***, Vol. 52, No. 2, (2019), 1089–1106. doi:10.1007/s10462-018-9641-3
12. LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning", ***Nature***, Vol. 521, (2015), 436–444. doi: 10.1038/nature14539
13. Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks", ***Advances in Neural Information Processing Systems***, Vol. 25, No. 2, (2012), 1097–1105. doi:10.1145/3065386
14. Al-Masni, M. A., Al-Antari, M. A., Park, J. M., Gi, G., Kim, T. Y., Rivera, P., Valarezo, E., Han, S.-M., and Kim, T.-S., "Detection and classification of the breast abnormalities in digital mammograms via regional Convolutional Neural Network", Proceedings of 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), (2017), 1230–1233. doi:10.1109/EMBC.2017.8037053
15. Ciregan, D., Meier, U., and Schmidhuber, J., "Multi-column deep neural networks for image classification", 2012 IEEE Conference on Computer Vision and Pattern Recognition, (2012), 3642–3649. doi:10.1109/CVPR.2012.6248110
16. Cernazanu-Glavan, C., and Holban, S., "Segmentation of bone structure in X-ray images using convolutional neural network", ***Advances in Electrical and Computer Engineering***, Vol. 13, No. 1, (2013), 87–94. doi:10.4316/AECE.2013.01015
17. ISIC: ISBI, Skin lesion analysis towards melanoma detec- tion. Vol. 3, (2016). [Online]. Retrieved from Https://Goo.Gl/2A1913, Accessed 2016.
18. Burdick, J., Marques, O., Weinthal, J., and Furht, B., "Rethinking Skin Lesion Segmentation in a Convolutional Classifier", ***Journal of Digital Imaging***, Vol. 31, No. 4, (2017), 435–440. doi:10.1007/s10278-017-0026-y
19. Yu, L., Chen, H., Dou, Q., Qin, J., and Heng, P.-A., "Automated melanoma recognition in dermoscopy images via very deep

residual networks", ***IEEE Transactions on Medical Imaging***, Vol. 36, No. 4, (2017), 994–1004. doi:10.1109/TMI.2016.2642839

20. Yuan, Y., Chao, M., and Lo, Y.-C., "Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance", ***IEEE Transactions on Medical Imaging***, Vol. 36, No. 9, (2017), 1876–1886. doi:10.1109/tmi.2017.2695227
21. Lin, B. S., Michael, K., Kalra, S., and Tizhoosh, H. R., "Skin lesion segmentation: UNets versus clustering", 2017 IEEE Symposium Series on Computational Intelligence (SSCI), (2017). doi:10.1109/SSCI.2017.8280804
22. Li, Y., and Shen, L., "Skin lesion analysis towards melanoma detection using deep learning network", ***Sensors***, Vol. 18, No. 2 (556), (2018), 1–16. doi:10.3390/s18020556
23. Bi, L., Kim, J., Ahn, E., and Feng, D., "Automatic Skin Lesion Analysis using Large-scale Dermoscopy Images and Deep Residual Networks", (2017), 6–9. Http://Arxiv.Org/Abs/1703.04197
24. Yuan, Y., and Lo, Y.-C., "Improving Dermoscopic Image Segmentation With Enhanced Convolutional-Deconvolutional Networks", ***IEEE Journal of Biomedical and Health Informatics***, Vol. 23, No. 2, (2019), 519–526. doi:10.1109/JBHI.2017.2787487
25. Baghersalimi, S., Bozorgtabar, B., Schmid-saugeon, P., Ekenel, H. K., and Thiran, J., "DermoNet : densely linked convolutional neural network for efficient skin lesion segmentation", ***EURASIP Journal on Image and Video Processing***, Vol. 71, (2019), 1–10. doi:https://doi.org/10.1186/s13640-019-0467-y
26. Hasan, M. K., Dahal, L., Samarakoon, P. N., Tushar, F. I., and Martí, R., "DSNet: Automatic dermoscopic skin lesion segmentation", ***Computers in Biology and Medicine***, Vol. 120, No. April, (2020), 103738. doi:10.1016/j.compbiomed.2020.103738
27. Tang, P., Liang, Q., Yan, X., Xiang, S., Sun, W., Zhang, D., and Coppola, G., "Efficient skin lesion segmentation using separable-Unet with stochastic weight averaging", ***Computer Methods and Programs in Biomedicine***, Vol. 178, (2019), 289–301. doi:10.1016/j.cmpb.2019.07.005
28. Litjens, G., Kooi, T., Bejnordi, B. E., Arindra, A., Setio, A., Ciompi, F., Ghafoorian, M., Laak, J. A. W. M. Van Der, Ginneken, B. Van, and Sánchez, C. I., "A survey on deep learning in medical image analysis", ***Medical Image Analysis***, Vol. 42, No. December 2012, (2017), 60–88. doi:10.1016/j.media.2017.07.005
29. Shen, W., Yang, F., Mu, W., Yang, C., Yang, X., and Tian, J., "Automatic localization of vertebrae based on convolutional neural networks", ***Proceedings of the SPIE on Medical Imaging***, Vol. 9413, (2015), 94132E. doi:10.1117/12.2081941
30. Kawahara, J., and Hamarneh, G., "Multi-resolution-tract CNN with hybrid pretrained and skin-lesion trained layers", Machine Learning in Medical Imaging. Cham: Springer International Publishing, Vol. 10019, (2016), 164–171. doi: 10.1007/978-3-319-47157-0_20
31. Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., and Halpern, A., "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)", Retrieved from ArXiv:1710.05006v3, (2017).
32. ISIC, Skin lesion analysis towards melanoma detection. (2017). Available: [Accessed 08 02 2020], Retrieved from Https://Challenge.Isic-Archive.Com/Landing/2017
33. Vaidya, B., and Paunwala, C., "Deep Learning Architectures for Object Detection and Classification (Chapter 4)", Springer International Publishing, (2019), 53–79. doi:10.1007/978-3-030-03131-2
34. Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2014), 580–587.
35. Girshick, R., "Fast R-CNN", 2015 IEEE International Conference on Computer Vision (ICCV), (2015). doi:10.1109/ICCV.2015.169
36. Ren, S., He, K., Girshick, R., and Sun, J., "Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks", NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems, Vol. 1, (2015), 91–99.
37. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., and Berg, A. C., "SSD : Single Shot MultiBox Detector", Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol 9905. Springer, Cham, (2016). doi:10.1007/978-3-319-46448-0_2
38. Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., "You Only Look Once: Unified, Real-Time Object Detection", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016). doi:10.1109/CVPR.2016.91
39. Zou, Z., Shi, Z., Guo, Y., and Ye, J., "Object Detection in 20 Years: A Survey", (2019), 1–39. Retrieved from Http://Arxiv.Org/Abs/1905.05055
40. He, Y., Zeng, H., Fan, Y., Ji, S., and Wu, J., "Application of Deep Learning in Integrated Pest Management: A Real-Time System for Detection and Diagnosis of Oilseed Rape Pests", ***Mobile Information Systems***, Vol. 2019, (2019), 1–14. doi:10.1155/2019/4570808
41. Redmon, J., and Farhadi, A., "YOLO9000: Better, faster, stronger", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017). doi:10.1109/CVPR.2017.690
42. Redmon, J., and Farhadi, A., "YOLOv3: An Incremental Improvement", (2018). Retrieved from ArXiv:1804.02767v1
43. Haridas, R., and R L, J., "Convolutional neural networks: A comprehensive survey", ***International Journal of Applied Engineering Research***, Vol. 14, No. 3, (2019), 780–789. doi:10.37622/IJAER/14.3.2019.780-789.
44. Simonyan, K., and Zisserman, A., "Very Deep Convolutional Networks For Large-Scale Image Recognition", International Conference on Learning Representations, (2015), 1–14.
45. He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016). doi:10.1109/CVPR.2016.90
46. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going Deeper with Convolutions", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2015). doi:10.1109/CVPR.2015.7298594
47. Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A., "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning", AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, (2017), 4278–4284.
48. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L., "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs", International Conference on Learning Representations, (2015). Retrieved from https://arxiv.org/abs/1412.7062
49. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H., "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", Proceedings of the European Conference on Computer Vision (ECCV), (2018), 801–818.

50. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L., "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 40, No. 4, (2017), 834–848. doi:10.1109/TPAMI.2017.2699184

51. Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H., "Rethinking Atrous Convolution for Semantic Image Segmentation", (2017). ArXiv Preprint Http://Arxiv.Org/Abs/1706.05587

52. Al-antari, M. A., Al-masni, M. A., Park, S. U., Park, J. H., Metwally, M. K., Kadah, Y. M., Han, S. M., and Kim, T.-S., "An automatic computer-aided diagnosis system for breast cancer in digital mammograms via deep belief network", ***Journal of Medical and Biological Engineering***, Vol. 38, (2018), 443–456. doi:10.1007/S40846-017-0321-6

53. Powers, D. M., "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", ***Journal of Machine Learning Technologies***, Vol. 2, No. 1, (2011), 37–63. Retrieved from https://arxiv.org/abs/2010.16061

54. Pereira, S., Pinto, A., Alves, V., and Silva, C. A., "Brain tumor segmentation using convolutional neural networks in MRI images", ***IEEE Transactions on Medical Imaging***, Vol. 35, No. 5, (2016), 1240–1251. doi:10.1109/TMI.2016.2538465

Persian Abstract

چکیده

جداسازی دقیق ضایعات از تصاویر پوستی در تشخیص و درمان به موقع سرطان پوست و جلوگیری از مرگ بیماران بسیار مهم است. به دلیل تنوع شکل، اندازه، رنگ و محل ضایعات در تصاویر درموسکوپی، جداسازی خودکار ضایعات پوستی همچنان یک چالش محسوب می‌شود. در این مطالعه، یک روش دو مرحله ای برای جداسازی ضایعات پوستی مبتنی بر یادگیری عمیق ارائه می‌شود. در مرحله اول، اندازه و موقعیت مکانی تقریبی ضایعه توسط شبکه های عصبی پیچشی تخمین زده می‌شود. یک زیرتصویر پیرامون مستطیلِ تخمین زده شده‌ی محیط بر ضایعه، از تصویر اصلی جدا شده و به یک تصویر با ابعادِ از پیش تعیین شده، نرمال می‌شود. به منظور جداسازی ناحیه دقیق ضایعه از تصویر نرمال شده، شبکه های عصبی پیچشی در ساختار DeepLab مورد استفاده می‌گیرند. دقت مرحله نرمال سازی بر عملکرد نهایی تاثیر بسزایی دارد. به منظور افزایش دقت مرحله نرمال سازی، از ترکیب چهار شبکه در ساختار Yolov3 استفاده می‌شود. دو روش به منظور ترکیب ساختارهای Yolov3 پیشنهاد می شود. نتایج جداسازی توسط دو شبکه در ساختار DeepLab3+ نیز با هم ترکیب می‌شوند تا دقت مرحله دوم نیز بهبود یابد. یکی دیگر از چالشها در این زمینه وجود تعداد کم تصاویر آموزشی است. برای غلبه بر این مساله، از راهکارهای اضافه کردن تعداد تصاویر آموزشی با ایجاد تغییراتی در تصاویر موجود و همچنین به کارگیری مودهای مختلفِ یک تصویر در هر مرحله از روش پیشنهادی استفاده می‌شود. به منظور ارزیابی روش پیشنهادی، آزمایشها بر روی مجموعه داده شناخته شده ISBI 2017 انجام می‌شوند. نتایج آزمایشها نشان می‌دهد که روش پیشنهادی عملکرد بهتری نسبت به تمامی روشهای موجود ارائه می‌کند.