B-SegNet : Branched-SegMentor Network For Skin Lesion Segmentation

Shreshth Saini
Indian Institute of Technology
Jodhpur
Rajasthan, India
saini.2@iitj.ac.in

Young Seok Jeon
Saw Swee Hock School of Public
Health, National University Health
System, National University of
Singapore
Singapore
youngseokjeon74@gmail.com

Mengling Feng
Saw Swee Hock School of Public
Health, National University Health
System, National University of
Singapore
Singapore
ephfm@nus.edu.sg

ABSTRACT

Melanoma is the most common form of cancer in the world. Early diagnosis of the disease and an accurate estimation of its size and shape are crucial in preventing its spread to other body parts. Manual segmentation of these lesions by a radiologist however is time consuming and error-prone. It is clinically desirable to have an automatic tool to detect malignant skin lesions from dermoscopic skin images. We propose a novel end-to-end convolution neural network(CNN) for a precise and robust skin lesion localization and segmentation. The proposed network has 3 sub-encoders branching out from the main encoder. The 3 sub-encoders are inspired from Coordinate Convolution, Hourglass and Octave Convolutional blocks: each sub-encoder summarizes different patterns and yet collectively aims to achieve a precise segmentation. We trained our segmentation model just on the ISIC 2018 dataset. To demonstrate the generalizability of our model, we evaluated our model on the ISIC 2018 and unseen datasets including ISIC 2017 and PH². Our approach showed an average 5% improvement in performance over different datasets, while having less than half of the number of parameters when compared to other state-of-the-arts segmentation models.

KEYWORDS

Skin Cancer, Melanoma, Dermoscopy, CNN, Skin Lesion Segmentation, Coordinate Convolution, Octave convolution

ACM Reference Format:

Shreshth Saini, Young Seok Jeon, and Mengling Feng. 2021. B-SegNet : Branched-SegMentor Network For Skin Lesion Segmentation. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '21)*, April 8–10, 2021, Virtual Event, USA. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3450439.3451873>

1 INTRODUCTION

The American Cancer Society estimated that in US alone, nearly 96,480 people will be diagnosed with melanoma and 7,000 of the

Figure 1: Example dermoscopic images from ISIC 2018 [7, 26]. Dataset contains the challenging images with variegation (lighter areas), satellite lesions, confounder like gel or hair, similar skin texture as that of lesion, peripheral or incomplete image.

people will die of melanoma in 2019[1]. Globally, melanoma has caused more than 60,000 deaths out of the 350,000 cases [25]. The annual cost of healthcare for melanoma cancer exceeds \$8 billion [10]. It has also been demonstrated that, with early detection of melanoma, the 5-year survival rate can be increased up to 99%; whereas delaying the diagnosis can significantly reduce the survival rate to 23% [25] once it spreads to other parts of the body. Hence it is crucial to identify the cancerous skin lesion at the earliest to increase the survival rate.

Computer aided methods have been developed to assist physicians to detect early signs of malignant skin lesions. There has been a constant effort by researchers and dermatologists in the worldwide to increase the accuracy and scale of diagnostic method. The International Skin Imaging Collaboration(ISIC) provides an open source dataset[7] which comprises more than 25,000 dermoscopy images. The organisation has been hosting competitions on the dataset for various image analysis tasks every year since 2016. Through the previous ISIC competitions, it is observed that accurate detection of malignant lesions heavily relied on the accurate segmentation of the skin lesions [4]. Therefore, we aim to develop a novel and customized convolution neural network(CNN) for precise segmentation of skin lesions.

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted by ACM, provided that the copies are not made for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM CHIL '21, April 8–10, 2021, Virtual Event, USA

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8359-2/21/04...\$15.00

<https://doi.org/10.1145/3450439.3451873>

2021-03-04 19:11. Page 1 of 1–8.

Segmentation of skin lesion from dermoscopic images is a challenging task. Skin lesions are inherently subjected to a huge variance in its shape, size, and texture as shown in figure 1. While melanoma have very fuzzy lesion boundaries, there are further artifacts introduced due to hair, contrast, light reflection and medical gauze. To address the challenges, we propose a segmentation model that has 3 sub-encoders branching out from the main encoder: each sub-encoder extracts different features from the dermoscopic image: the lesion's location, boundary and a richer representation of the extracted features by performing the high compression-decompression of features. Besides the sub-encoders, we also experimented with using different modules at the bottleneck layer.

2 RELATED

Existing works on skin lesion segmentation can be broadly classified into two categories namely: the conventional non-CNN models and CNN-based models [4].

The conventional non-CNN models generally involve a manual extraction of abstract features of a skin lesions which include their textures, shapes, colours and edges [17]. The process of finding a collection of meaningful features and its association has been a challenging step for researchers as it requires a huge domain knowledge and time investment. Even so, these conventional models were known to generalize poorly over unseen images [12].

CNN-based models have shown many successful cases in lesion segmentation where conventional non-CNN models had struggled [4]. CNN models have a number of strengths over non-CNN models. CNN is a deep parametric model where its parameters can be estimated from data. Through training, the CNN model learns to pick up abstract features by itself where these features are usually more effective in summarising the image. CNN models therefore require very minimal intervention from human which could possibly be biased and ill-represented. Various attempts of using CNN for skin lesion segmentation have been proposed in the past years. A patch wise approach [13] was proposed to use both local features and global features to generate patch-wise lesion segmentation. The model is composed of two separate CNNs. The first CNN looks at small local patches of an image and the second CNN looks at the entire image but down-sampled to match its size to the local patches. Outputs from the two networks are concatenated to give the final prediction. The model however has to exhaust all of the patches to generate the full segmentation prediction.

Fully Convolutional Residual Network(FCRN) [30] and Deep Fully Resolution Convolutional Network(DFRCN) [2] are the cases where they adopted Fully Convolutional Neural Net(FCNN) for lesion segmentation. Unlike the patch-wise approach, FCNN is able to segment the entire lesion image at a single inference. The network does not have pooling layers in order to keep the spatial information intact throughout. This however forces the network to be shallower to compensate with the increasing features size in the hidden layers as the network gets deeper. Shallower network often leads to a less expressive model. To address this limitation, the FCRN uses a deep Resnet model for both segmentation and classification tasks. The model comprises over 50 Res-layers to make the network more expressive and also resistant to gradient vanishing problem. The model has pooling layers to compromise with memory issue

and up-sample layers to transform the encoded features back to the original spatial dimension. Even-though the network is deep, the network will not be able to utilise the embedded features that are in lower spatial dimension.

Focusnet [15] and a work by Vesal et al. [27] are the cases where attention mechanism was adopted to tell the network explicitly where to focus on. Vesal et al.[27] implemented Faster-RCNN [20] for a lesion segmentation. The model first looks for possible regions of interest(ROI). Segmentation is then done only on the image patches that are within the regions of ROIs. The network therefore focuses only on the small patches where it contains skin lesions. Along the similar lines, [22] uses a more robust cascading bottleneck for segmentation network. Focusnet [15] gives a soft attention to embedded features. The attention map is generated by multiplying the encoder features with decoder features. MobileGAN [23] experimented with using adversarial network for a skin lesion segmentation task, where some promising performance was achieved.

It can be observed that the majority of segmentation networks have a huge resemblance to U-Net [21]. These networks commonly have two sub-networks namely: encoder and decoder. These two sub-networks carry out different tasks. Encoder tries to extract abstract features from an image with a series of convolution, pooling and non-linear transformation. Decoder on the other hand tries to combine the pieces of features extracted from the encoder at correct locations. This is done by a series up-sampling, convolution and non-linear transformation. To ease the training, well known classification models like ResNet [11] that is pre-trained on Image-net is chosen as the encoder. Decoder on the other hand is randomly initialized.

Though the vanilla U-Net approach works well for simple segmentation tasks, the model fails when the task gets more demanding [28]. The problem of U-Net lies in its inability to keep the spatial information of an image intact as the network gets deeper. The problem has been raised and addressed by Liu et al. where they introduced Coordinate Convolution [16] which tries to preserve the location information by explicitly adding coordinate index before a convolution. Yunpeng et al. proposed Octave Convolution [6] which processes feature information based on their frequency, hence better locating boundaries of an input image. Another challenge of skin lesion segmentation is the existence of unwanted artifacts in the image, such as hair, gel and etc. To tackle the problem, we took inspiration from PixlSegNet [14] where they introduced the use of hourglass modules to allow their network to see through the obstacles in such non ideal conditions.

3 METHODOLOGY

To address the above issues of 1) loss of spatial information after pooling 2) fuzzy edge segmentation 3) unwanted impurities in an image, our method presents 3 different sub-encoders that can possibly solve each of the issues.

Apart from the main encoder (Ξ), our model contains 3 sub-encoders (Ξ_i) namely: 1) CoordConv (Ξ_1) 2) HourGlass (Ξ_2) and 3) OctConv (Ξ_3) encoders which branches out from the main encoder. Each sub-encoder is designed to serve a different purpose. They are branched out from different layers of the main encoder to ensure

Figure 2: Overall Architecture of proposed Network.

that they begin to extract features starting from the most relevant feature maps and not directly on the input image. This helps us in reducing the overall number of parameters of the model. Instead of using a simple convolutional layer as a bottleneck, we also experimented using dilated convolution to broaden the receptive field, hourglass module to better extract the lesion features even in the presence of obstacles such as hair or gel. To improve the decoder's capacity, we incorporated the deep supervision strategy by recursively introducing sub-decoders from the main decoder. The features after the first block of decoder is branched out to give a smaller sub-decoder and the features from first block of the sub-decoder is likewise branched out to give another sub-decoder. Outputs from the main decoder and sub-decoders are then convoluted after a concatenation to give the final segmentation prediction. An overview of the proposed model is illustrated in figure 2.

$$\hat{Y}_{map} = \sum \mathcal{D}_j(\mathcal{B}(\sum \Xi_i(X_1, X_2, X_3))) \quad (1)$$

In the equation 1, X_1, X_2 , and X_3 are scaled input images as shown in figure 2. $\mathcal{B}$ and $\mathcal{D}_j$ denotes the bottleneck block and decoder block respectively. Effectively the equation represents the model and is differentiable.

3.1 Encoder and Sub-Encoders

Our model starts by copying an input image and scaling them down to 2 different resolutions. Each of the three inputs then passes through a convolutional layer, Attention-block and up-sampling to transform the features back to its original input size. The three feature maps are then added to give a new input feature map to the

main encoder. The down scaling approach described above helps in having a more robust encoder by providing a multi-scale view which helps in extracting both fine and coarse lesion features. Attention-block (figure 3) consists of a channel-wise attention, spatial-wise attention and a small hourglass block to have better represented features and more robust to obstacles in an image. The main encoder is comprised of 3 sub-block. Each sub-block consists of a convolutional layer, Attention-block and 2x2 max-pooling layer. Before doing a max-pooling from each of the sub-block, the features were copied over and fed into 3 different sub-encoders. Eventually, output features from both the main and the 3 sub-encoders are concatenated to generate a feature to the bottleneck layer. Architecture of each of the 3 sub-encoders used and the bottleneck layer are explained in the sections below.

$$\chi_{main} = \Xi_{main}(X_1, X_2, X_3) \quad (2)$$

$$\chi_1 = \Xi_1(\Xi_{main}^1, \Xi_{main}^2, \Xi_{main}^3) \quad (3)$$

$$\chi_2 = \Xi_2(\Xi_1^1, \Xi_1^2) \quad (4)$$

$$\chi_3 = \Xi_3(\Xi_2^1) \quad (5)$$

Equation 2-5 χ_i and Ξ_i^j show output feature map and intermediate features from corresponding i^{th} blocks. Here each X_i consists of either convolution and attention blocks, Coordinate convolution blocks, or octave convolution block. Further each block can be represented as $\Xi_i = C_k(C_{k-1}(\cdot) + \Xi_{i-1}^{k-1})$, where C is any of the above mentioned block.

Figure 3: Combined spatial, channel attention and hourglass modules.

Figure 4: Coordinate convolutional layer. x,y and r represents the coordinate and radial information of input feature map.

3.1.1 Sub-Encoder I: CoordConv-Encoder. CoordConv-encoder is the first sub-encoder that is branched out from the main encoder. CoordConv-encoder (depicted in figure 4 and equation 3) has 3 blocks and each block consists of a coordinate-convolution and a convolutional layer. Output from each block of CoordConv-encoder was also added with features from the main encoder before being sent to the next block of CoordConv-encoder. By introducing the location information explicitly to the features using coordinate-convolutions, the feature computed thereafter will have better representation of location.

3.1.2 Sub-Encoder II: HourGlass-Encoder. HourGlass-Encoder is the second sub-encoder that is branched out from the main encoder. HourGlass-encoder (depicted in figure 5 and equation 4) has 2 blocks and each block consists of an hourglass module and a convolutional layer. Likewise to the CoordConv-encoder, Output from each block of Hourglass-encoder was also added with features from the main encoder before being sent to the next block of Hourglass-encoder. Hourglass module [14] is a bottom down-up approach to better encode and represent the feature maps in

Figure 5: Hourglass module.

a higher dimension. Compressed features allow to extract the features which are part of lesion and give ability to look through small obstacles such as hair or gel. Hourglass module contains small encoder-decoder with convoluted skip connections. The skip connections prevent the issue of information loss within the module.

3.1.3 Sub-Encoder III: OctConv-Encoder. Recent success in the extraction of frequency based information to better understand the details of input image and features was demonstrated in Octave Convolution [6]. We have incorporated this approach to give our model an ability to extract lesion boundary features which could be viewed as features with high frequency. Octave convolution (oct-conv) processes the input feature at different scales and channels to differentiate the high and low frequency regions. Since the lesion boundaries and high color intensity variation in and around the lesion are of high frequency features and identifying them makes our network more robust in the lesion boundary localization and thus achieving a sharper mask prediction.

Oct-conv sub-encoder was extracted from the last (third) block of main encoder with shape $64 \times 64 \times 256$. Output of the sub-encoder was added with the output of third block of main encoder and finally feature map of shape $32 \times 32 \times 256$ was combined with the rest of the feature maps from other sub-encoders. The ratio in which high and low frequency channel were divided was kept same as that in original work [6] i.e. 0.25. Figure 6 shows the overview of octave convolution.

3.2 Bottleneck

Low receptive field at the bottleneck of a network is one reason that leads to a poor performance of a segmentation model [5, 27]. Therefore to broaden the overall receptive field we used a set of dilated convolutions with varying dilation rates of 4, 8, 12 and 16. The dilated convolutions form a D-link [31] structure which has the dilated convolutions in both cascaded and parallel style. Output feature maps from bottleneck modules were then sent to the decoder network to obtain the segmentation mask.

We also experimented with using other possible modules such as hourglass or SE-block as the bottleneck modules. The performance

Figure 6: Octave convolution.

comparison of using different bottleneck modules are provided at the result section.

$$\chi_{bottleneck} = \sum (C_k(\sum (\Xi_i))) \quad (6)$$

Here, input to all the bottleneck block $\mathcal{B}(\cdot)$ are same.

3.3 Decoder

To improve the decoder's capacity, we incorporated the deep supervision strategy by recursively introducing sub-decoders from the main decoder. Figure 2 shows the detailed structure of decoder and of network as whole. Use of cascaded decoder allows to increase the decoder's learning capacity which ultimately give us a more precise and confident mask prediction. Outputs from the main decoder and sub-decoders were added and sent through a convolutional layer with sigmoid activation function to finally generate a probability outcome. Sub-decoders have similar structure as that of sub-encoders other than blocks and additional skip connections.

3.4 Network Training

We trained our model in an end-to-end manner with ISIC 2018 [7, 26] segmentation dataset (see Section 4.1 for details). We augmented our dataset to increase the number of training samples where we used random rotations, shear stretch, contrast, brightness and central cropping. All of the convolutional layer in the network is composed of a 3x3 conv layer, *ReLU* activation and batch-norm layer except for the last convolutional layer which has a sigmoid layer instead.

$$L(y_n, \tilde{y}_n) = 1 - \frac{2 \sum_n y_n \tilde{y}_n}{\sum_n y_n + \sum_n \tilde{y}_n} \quad (7)$$

Dice loss (equation 7) was used to address the label imbalance issue. y_n represents the the true label and $\tilde{y}_n$ denotes the predicted label by the network. We used the Stochastic Gradient Descent (SGD) with learning rate of 0.9, momentum of 0.9 and learning rate of 0.0002. We trained our model for 110 epochs with batch-size of 8. other models for a benchmark were also trained and evaluated with the same training configuration.

2021-03-04 19:11. Page 5 of 1–8.

Figure 7: Qualitative results. The three rows of images show cases where our model is able to perform well under varying conditions. First, second and third images are the inputs, ground-truth masks and predicted masks respectively.

4 EXPERIMENTS

4.1 Dataset

To validate on our model's generalizability, our model's performance was assessed on three different publicly available skin lesion datasets: the ISIC 2018 [7, 26], the ISIC 2017 [8] and the PH² [18]. The ISIC 2018 dataset consists of 2594 training samples with labels and 100 validation samples without labels. ISIC 2017 dataset have 2000 training samples and a separate validation and test of datasets of 150 and 600 respectively with their corresponding ground truth masks. PH² dataset consists of 200 image samples along with its ground-truth masks. The dimensions of ISIC dataset varies from 453x697 to 4499x6748 while PH² samples have a size of 786x560 pixels.

We trained our model only with 1994 training samples from ISIC 2018 dataset and used the remain 600 training samples for internal validation. In the final evaluation, we tested our model on the ISIC 2018 validation dataset through their online submission system, the ISIC 2017 test dataset and the entire PH² dataset. Table 1 summarizes the sample distribution of the three datasets.

4.2 Experimental Results

To evaluation our method we primarily used the *Dice Coefficient (DC)*, *Thresholded Jaccard Index (JI)* and *Accuracy (AC)*.

$$DC = \frac{2 \sum_n y_n \tilde{y}_n}{\sum_n y_n + \sum_n \tilde{y}_n} \quad (8)$$

$$JI = \frac{P(y, \tilde{y})}{P(y, \tilde{y}) + P(\neg y, \tilde{y}) + N(y, \neg \tilde{y})} \quad (9)$$

Where $P(y, \tilde{y})$ is true positive, $N(\neg y, \neg \tilde{y})$ is true negative, $P(\neg y, \tilde{y})$ is false positive and $N(y, \neg \tilde{y})$ is false negative.

Table 1: Description of datasets.

Dataset	Training	Validation	Testing
ISIC 2018 [7, 26]	2594	100	1000
ISIC 2017 [8]	2000	150	600
PH ² [18]	-	-	200

Table 2: Comparison of B-SegNet with state-of-the-art methods on ISIC 2018 [7, 26], ISIC 2017 [8] and PH² [18] datasets.

Dataset	Methods	AC	DC	Thresholded JI	Parameters (million)
ISIC 2018	FocusNet	0.92	0.83	0.76	-
	Shahin et al.	-	0.90	0.84	-
	Our(B-SegNet)	0.98	0.95	0.84	7.53
ISIC 2017	SegNet	0.92	0.82	0.70	11.5
	SLSDeep	0.94	0.88	0.78	46.65
	SegAN	0.94	0.88	0.79	382.17
	NCARG	0.95	0.90	0.83	-
	FrCN	0.96	0.89	0.81	16.3
	Vesal et. al.	0.96	0.93	-	25
	Our(B-SegNet)	0.97	0.96	0.85	7.53
PH ²	FrCN	0.95	0.91	0.84	16.3
	Vesal et al.	0.96	0.94	-	25
	Our(B-SegNet)	0.97	0.96	0.87	7.53

We present our results in Table 2. The table shows the performance comparison of our model against other state-of-the-art models over different datasets. For the ISIC 2017 dataset, our model was benchmarked against the model from SegNet [3], SLSDeep [19], SegAN [29], NCARG [9], FrCN [2] and the model from Vesal et al. [27]. For ISIC 2018 and the PH² datasets, as only a limited number of benchmarking models reported their performance on these datasets, our model was compared against the FocusNet [15] and the model from Shahin et al. [24] for the ISIC 2018 dataset and the FrCN and the model from Vesal et al. for the PH² datasets respectively.

Our model was able to outperform the benchmarking methods over different datasets in term of accuracy, Dice and Thresholded JI. For the ISIC 2018 dataset, measured by Dice, our proposed model outperformed FocusNet by 14.5% and Shahin's model by 5.5%; for the ISIC 2017 dataset, we outperformed all comparing models; for the PH² dataset, measured by Dice, we outperformed FrCN by 5.5% and Vesal by 2%. Moreover, our model contains fewer parameters. For example, the number of parameters in our model is only 2% of the ones for SegAN and 65% of the ones for SegNet (the smallest benchmarking model).

Figure 7 shows the qualitative result to visually confirm the performance of our method. First row of figure 7 shows the case when the ground truth mask was poorly labeled (medical grade) from a dermatologist's perspective while our method gave better results than the ground truth itself. The second row shows the case when the lesion and the background have similar textural features but yet being able to correctly segment the lesion. The third row shows the robustness of our model when an image has obstacles along the lesion. Some failure cases are discussed in figure 10.

Figure 8: Variation in thresholded jaccard index with network configuration.

4.3 Ablation Study

Table 3 and figure 8 provide a trend in the performance of our model when only one of the three sub-encoders were activated as well as different bottleneck modules were used. Baseline model is a model with no sub-encoders and a normal convolutional layers as its bottleneck module. Our model was re-trained when testing for the different configurations. The results show that each of the sub-encoder: CoordConv, HourGlass and OctConv improves the model's performance. We observed the best performance under the baseline condition when all of the three sub-encoders were activated with accuracy, dice and thresholded JI of 0.969, 0.894 and

Table 3: Performance of our method with variation in network structure on ISIC 2017 [8] dataset.

Method		AC	DC	Thresholded JI	Parameters (million)
Base		0.95	0.85	0.69	1.79
Individual sub-encoder	+HG	0.96	0.87	0.72	+1.46
	+OC	0.95	0.87	0.71	+0.95
	+CC	0.96	0.87	0.72	+0.75
Base + all sub-encoders		0.96	0.89	0.78	4.97
Bottleneck module with all sub-encoders	+(SE)	0.96	0.90	0.80	+0.60
	+(HG)	0.96	0.91	0.80	+0.77
	+(dlink)	0.97	0.93	0.82	+1.18
B-SegNet		0.97	0.96	0.85	7.53

Figure 9: Intermediate feature map visualization for last sub decoder output in (a) and spatial and channel attention module output in (b). In (a) and (b), the first-row left image is the feature map for the model with no sub encoder, first-row right image for the model with only coord-conv encoder, second-row left image for the model with all sub-encoders, second-row right image for the model with hourglass-encoder, last-row right image for octconv-encoder, and left one for all sub-encoder with bottleneck modules.

0.787 respectively. We observed the worst performance when all of the sub-encoders were turned off. The performance difference with and without sub-encoders was measured to be 1.4%, 4.3% and 9.48% for accuracy, dice and thresholded JI respectively. The sub-encoders: CoordConv, HourGlass and OctConv required 0.75, 1.46 and 0.95 million parameters respectively.

Different types of bottleneck blocks: D-link, Attention and Hour-Glass were also tested. We observed the highest performance when D-link bottleneck was used. Our model thus achieved the best performance of 0.97, 0.96 and 0.85 for accuracy, dice and Thresholded JI respectively when all three sub-encoders were activated and D-link bottleneck module was used.

2021-03-04 19:11. Page 7 of 1–8.

Figure 10: Failure cases. In figure first, second and third rows show the failure in presence of combination of cases where gel, excessive hair and extremely poor contrast of lesion were present.

We visualize the intermediate feature outputs to understand the differences and impact of different modules in the model's learning capacities. In figure 9, for the same input image, we show feature maps from different sub-models. Feature output from the last sub-decoder is used for comparison purposes as it reflects the overall model output. Feature maps from spatial and channel attention module are visualized to depict the change in attention region over model modifications. Figure 9-(a) shows the difference as when all the sub-encoders are present, and output features show a more precise and distinguishable map for foreground and background regions and fuzzy lesion boundaries compared to when none is present. In the case of hourglass-encoder, feature map outputs present a better classification for foreground pixels. A similar task-specific feature can be observed for other intermediate outputs. Figure (b) shows that output from the attention module shows improved attention to boundary and lesion regions as we move from no sub-encoder to B-SegNet.

5 CONCLUSION AND FUTURE WORK

We proposed a new model, *B-SegNet*, for the segmentation of skin lesion. Our model not only outperformed the state-of-the-art models over all three public skin lesion datasets, and it also contains a much smaller number of parameters. We have also shown that each sub-encoder in our model serves different yet important purpose such as giving confident boundaries and obstacle handling etc. The qualitative review confirmed that our model often produces better segmentation results than the ground-truth mask drawn by a dermatologist. Though we have achieved the best result when all 3 sub-encoders were activated, more thorough research has to be done on the model to validate the true ability of each of the sub-encoders. The roles of each sub-encoders therefore needs to be validated visually with qualitative experiments in future. We also plan to work on creating an end-to-end diagnosis system for skin lesion detection, which will not only localize or segment the lesion but will also classify the lesion type.

ACKNOWLEDGMENTS

REFERENCES

- [1] Melanoma Foundation (AIM). 2019. Melanoma stats, facts and figures. Retrieved August 10, 2020 from <https://www.aimatmelanoma.org/about-melanoma/melanoma-stats-facts-and-figures/>
- [2] Mohammed A. Al-masni, Mugahed A. Al-antari, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. 2018. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Computer Methods and Programs in Biomedicine* 162 (2018), 221–231. <https://doi.org/10.1016/j.cmpb.2018.05.027>
- [3] V. Badrinarayanan, A. Kendall, and R. Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 12 (2017), 2481–2495. <https://doi.org/10.1109/TPAMI.2016.2644615>
- [4] M. Emre Celebi, Quan Wen, Hitoshi Iyatomi, Kouhei Shimizu, Huiyu Zhou, and Gerald Schaefer. 2015. *A State-of-the-Art Survey on Lesion Border Detection in Dermoscopy Images*. 97–129. <https://doi.org/10.1201/b19107-5>
- [5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *CoRR* abs/1606.00915 (2016). <http://arxiv.org/abs/1606.00915>
- [6] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. 2019. Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution. *arXiv:1904.05049* [cs.CV]
- [7] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368* (2019).
- [8] 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. 2018. 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). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 168–172. <https://doi.org/10.1109/ISBI.2018.8363547>
- [9] Yanhui Guo, Amira S. Ashour, and Florentin Smarandache. 2018. A Novel Skin Lesion Detection Approach Using Neutrosophic Clustering and Adaptive Region Growing in Dermoscopy Images. *Symmetry* 10, 4 (2018). <https://doi.org/10.3390/sym10040119>
- [10] Gery P. Guy, Steven R. Machlin, Donatus U. Ekwueme, and K. Robin Yabroff. 2015. Prevalence and Costs of Skin Cancer Treatment in the U.S., 2002/2006 and 2007/2011. *American Journal of Preventive Medicine* 48, 2 (2015), 183–187. <https://doi.org/10.1016/j.amepre.2014.08.036>
- [11] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [12] Janet Weinthal Borko Furth Jack Burdick, Oge Marques. 2018. Rethinking Skin Lesion Segmentation in a Convolutional Classifier. *Journal of Digital Imaging* 31, 435–440. <https://doi.org/10.1007/s10278-017-0026-y>
- [13] M. Hossein Jafari, Ebrahim Nasr-Esfahani, Nader Karimi, S. M. Reza Soroushmehr, Shadrokh Samavi, and Kayvan Najarian. 2017. Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma. *International Journal of Computer Assisted Radiology and Surgery* 12, 6 (01 Jun 2017), 1021–1030. <https://doi.org/10.1007/s11548-017-1567-8>
- [14] Ranjeet Ranjan Jha, Gaurav Jaswal, Divij Gupta, Shreshth Saini, and Aditya Nigam. 2020. PixlSegNet: pixel-level iris segmentation network using convolutional encoder-decoder with stacked hourglass bottleneck. *IET Biometrics* 9, 1 (2020), 11–24. <https://doi.org/10.1049/iet-bmt.2019.0025>
- [15] C. Kaul, S. Manandhar, and N. Pears. 2019. Focusnet: An Attention-Based Fully Convolutional Network for Medical Image Segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 455–458. <https://doi.org/10.1109/ISBI.2019.8759477>
- [16] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. 2018. An intriguing failing of convolutional neural networks and the CoordConv solution. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 9605–9616. <https://papers.nips.cc/paper/8169-an-intriguing-failing-of-convolutional-neural-networks-and-the-coordconv-solution.pdf>
- [17] Ali Madooei, Mark S. Drew, Maryam Sadeghi, and M. Stella Atkins. 2012. Intrinsic Melanin and Hemoglobin Colour Components for Skin Lesion Malignancy Detection. (2012), 315–322. https://doi.org/10.1007/978-3-642-33415-3_39
- [18] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira. 2013. PH2 - A dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 5437–5440. <https://doi.org/10.1109/EMBC.2013.6610779>
- [19] K. S. Mostafa, A. R. Hatem, A. Farhan, F. B. Syeda, S. Adel, K. S. Vivek, U. H. C. Forhad, A. Saddam, R. Santiago, R. Petia, and P. Domenec. 2018. SLSDeep: Skin Lesion Segmentation Based on Dilated Residual and Pyramid Pooling Networks. *ArXiv: abs/1805.10241* (2018).
- [20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc. <https://doi.org/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf>
- [21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
- [22] S. Saini, D. Gupta, and A. K. Tiwari. 2020. Detector-SegMentor Network for Skin Lesion Localization and Segmentation. In *Computer Vision, Pattern Recognition, Image Processing, and Graphics*, R. Venkatesh Babu, Mahadeva Prasanna, and Vinay P. Nambodiri (Eds.). Springer Singapore, Singapore, 589–599. https://doi.org/10.1007/978-981-15-8697-2_55
- [23] K. Sarker, Md. Mostafa, H. Rashwan, M. Abdel-Nasser, V. Singh, S. F. Banu, F. Akram, F. Chowdhury, K. Choudhury, S. Chambon, P. Radeva, and D. Puig. 2019. MobileGAN: Skin Lesion Segmentation Using a Lightweight Generative Adversarial Network. *ArXiv: abs/1907.00856* (07 2019).
- [24] A. H. Shahin, K. Amer, and M. A. Elattar. 2019. Deep Convolutional Encoder-Decoders with Aggregated Multi-Resolution Skip Connections for Skin Lesion Segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 451–454. <https://doi.org/10.1109/ISBI.2019.8759172>
- [25] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. 2018. Cancer statistics, Cancer Facts and Figures 2018. American Cancer Society, 7–30 pages. <https://doi.org/10.3322/caac.21442>
- [26] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* 5 (2018), 180161. <https://doi.org/10.1038/sdata.2018.161>
- [27] Sulaiman Vesal, Shreyas Malakarjun P., Nishant Ravikumar, and Andreas K. Maier. 2018. A Multi-task Framework for Skin Lesion Detection and Segmentation. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer International Publishing, Cham, 285–293. https://doi.org/10.1007/978-3-030-01201-4_31
- [28] Zhao Xiaomei, Yihong Wu, Guidong Song, Li Zhenye, Yong Fan, and Yazhuo Zhang. 2016. Brain Tumor Segmentation Using a Fully Convolutional Neural Network with Conditional Random Fields. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, Cham, 75–87. https://doi.org/10.1007/978-3-319-55524-9_8
- [29] Y. Xue, T. Xu, and X. Huang. 2018. Adversarial learning with multi-scale loss for skin lesion segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 859–863. <https://doi.org/10.1109/ISBI.2018.8363707>
- [30] L. Yu, H. Chen, Q. Dou, J. Qin, and P. Heng. 2017. Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. *IEEE Transactions on Medical Imaging* 36, 4 (April 2017), 994–1004. <https://doi.org/10.1109/TMI.2016.2642839>
- [31] Lichen Zhou, Chuang Zhang, and Ming Wu. 2018. D-LinkNet: LinkNet with Pre-trained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2018), 192–1924. <https://doi.org/10.1109/CVPRW.2018.00034>