SESV: Accurate Medical Image Segmentation by Predicting and Correcting Errors

Yutong Xie, *Graduate Student Member, IEEE*, Jianpeng Zhang, Hao Lu, Chunhua Shen, *Senior Member, IEEE*, and Yong Xia, *Member, IEEE*

***Abstract*— Medical image segmentation is an essential task in computer-aided diagnosis. Despite their prevalence and success, deep convolutional neural networks (DCNNs) still need to be improved to produce accurate and robust enough segmentation results for clinical use. In this paper, we propose a novel and generic framework called Segmentation-Emendation-reSegmentation-Verification (SESV) to improve the accuracy of existing DCNNs in medical image segmentation, instead of designing a more accurate segmentation model. Our idea is to predict the segmentation errors produced by an existing model and then correct them. Since predicting segmentation errors is challenging, we design two ways to tolerate the mistakes in the error prediction. First, rather than using a predicted segmentation error map to correct the segmentation mask directly, we only treat the error map as the prior that indicates the locations where segmentation errors are prone to occur, and then concatenate the error map with the image and segmentation mask as the input of a re-segmentation network. Second, we introduce a verification network to determine whether to accept or reject the refined mask produced by the re-segmentation network on a region-by-region basis. The experimental results on the CRAG, ISIC, and IDRiD datasets suggest that using our SESV framework can improve the accuracy of DeepLabv3+ substantially and achieve advanced performance in the segmentation of gland cells, skin lesions, and retinal microaneurysms. Consistent conclusions can also be drawn when using PSPNet, U-Net, and FPN as the segmentation network, respectively. Therefore, our SESV framework is capable of improving the accuracy of different DCNNs on different medical image segmentation tasks.**

***Index Terms*— Medical image segmentation, deep convolutional neural network, correction learning.**

I. Introduction

MEDICAL image segmentation plays a pivotal role in clinical practices and research settings [1]. Since manual segmentation is time-consuming and expensive, automated solutions have been extensively studied [1], [2]. Traditional approaches to medical image segmentation are mainly based on hand-crafted features [1]. Recently, deep convolutional neural networks (DCNNs) have achieved remarkable success in many vision tasks including image segmentation. Such success has prompted investigators to apply DCNNs to medical image segmentation [3]–[29].

Although more accurate than traditional approaches, DCNNs still suffer from limited segmentation accuracy, mainly due to the small amount of training data, complex anatomical variations among subjects, low soft tissue contrast, and various imaging artifacts. For clinical use, the inaccuracy in segmentation results may mislead medical professionals and thus carry serious repercussions for the diagnosis and treatment [30]. There are several ways to improve the segmentation accuracy of DCNNs, such as designing more effective network architectures, acquiring and annotating sufficient training data, and enhancing the quality of medical images. Each, however, has its limitations. Alternatively, it is intuitive to obtain more accurate segmentation results via identifying where a DCNN makes errors and then correcting the errors accordingly. Wang *et al*. [31] employed user interactions to generate the error map of each segmentation result, and then fed the error map and other image information to a deep model to improve segmentation accuracy. However, it is expensive to record user interactions in clinical practice, since such interactions require a high degree of concentration and expertise.

Manuscript received August 13, 2020; accepted September 10, 2020. Date of publication September 21, 2020; date of current version December 29, 2020. The work of Yutong Xie, Jianpeng Zhang, and Yong Xia was supported in part by the National Natural Science Foundation of China under Grant 61771397, in part by the Science, Technology and Innovation Commission of Shenzhen Municipality under Grant JCYJ20180306171334997, in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX202010, and in part by the China Scholarship Council (CSC). *(Yutong Xie and Jianpeng Zhang contributed equally to this work.) (Corresponding author: Yong Xia.)*

Yutong Xie and Jianpeng Zhang were with the School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia. They are now with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: xuyongxie@mail.nwpu.edu.cn; james.zhang@mail.nwpu.edu.cn).

Hao Lu was with the School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia. He is now with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: poppinace@foxmail.com).

Chunhua Shen is with the School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: chunhua.shen@adelaide.edu.au).

Yong Xia is with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Research and Development Institute, Northwestern Polytechnical University, Shenzhen 518057, China (e-mail: yxia@nwpu.edu.cn).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.3025308

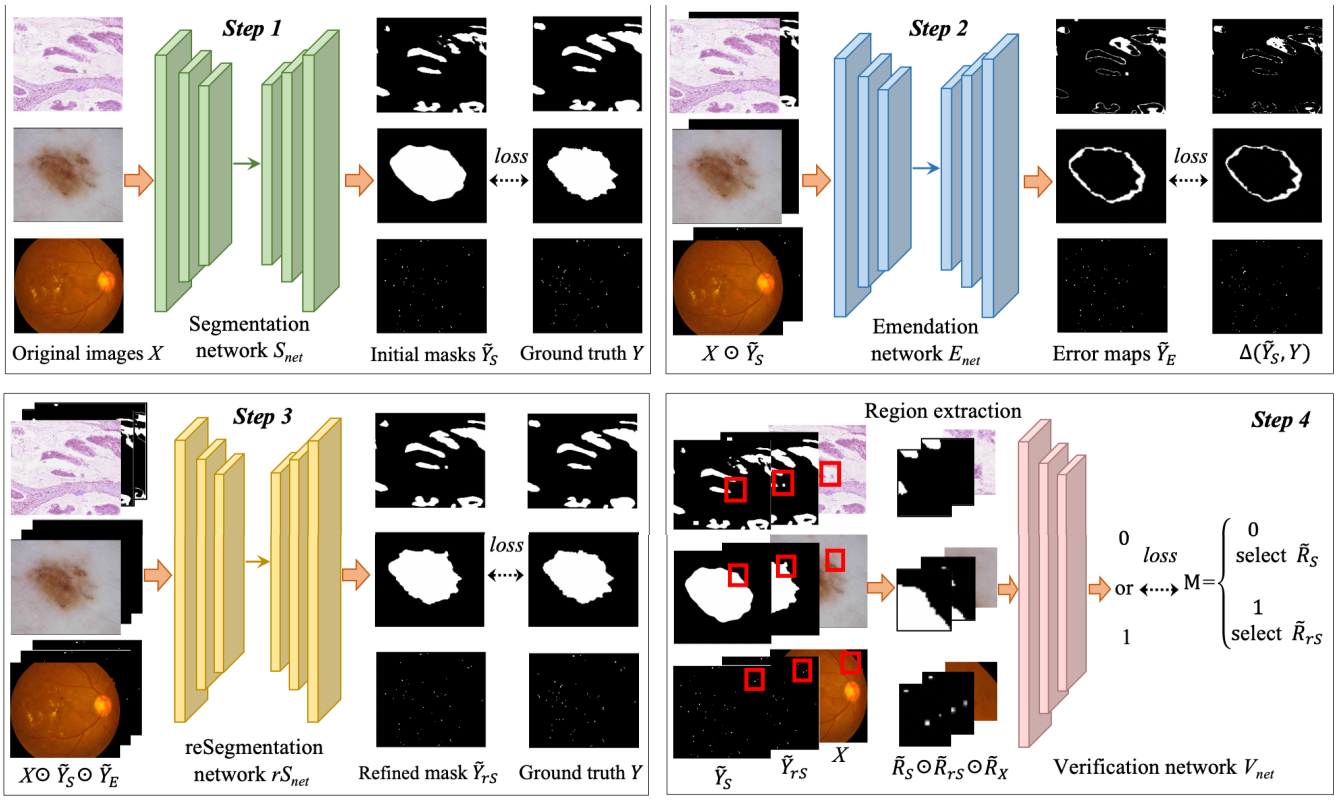

Fig. 1. Technical pipeline of the proposed SESV framework, which consists of four steps: (1) applying a segmentation network S_{net} to a medical image X to generate an initial segmentation mask $\tilde{Y}_S$, (2) feeding the concatenation of X and $\tilde{Y}_S$ into an emendation network E_{net} to predict a segmentation error map $\tilde{Y}_E$, and (3) feeding the concatenation of X, $\tilde{Y}_S$, and $\tilde{Y}_E$ into a re-segmentation network rS_{net} to generate a refined segmentation mask $\tilde{Y}_{rS}$, and (4) using a verification network V_{net} to determine whether to accept or reject $\tilde{Y}_{rS}$ on a region-by-region basis. The symbol $\odot$ represents the concatenation operation.

To construct a fully automated segmentation approach, in our pilot study [14], we trained an emendation network to predict segmentation errors, and then employed the predicted errors to revise the segmentation result directly. Unfortunately, predicting segmentation errors is also extremely difficult and often poorly done. If more than half of the predicted erroneous pixels are not truly mis-segmented, using predicted segmentation errors to revise the result may deteriorate, instead of improving, the segmentation accuracy. Nevertheless, although less-accurate, the predicted segmentation errors in most cases can indicate the locations where errors are prone to occur, since it is much easier to predict the rough location of a segmentation error than to predict the erroneous region. To make the segmentation-emendation strategy more practical and tractable, we therefore advocate to treat the predicted segmentation errors as the prior that indicates the locations of segmentation errors and feed them to a re-segmentation network to refine segmentation results.

In this work, we propose the Segmentation-Emendation-reSegmentation-Verification (SESV) framework (see Figure 1) to improve the accuracy of existing medical image segmentation models. We first apply a base segmentation network to a medical image, aiming to generate an initial segmentation mask. Then, we concatenate the obtained initial mask with the input image and feed them into an emendation network to predict a segmentation error map. Next, we further concatenate the predicted error map, which indicates the locations of possible segmentation errors, with the initial mask and input image, and feed the concatenation into a re-segmentation network to produce a refined segmentation mask. Finally, since even the predicted error locations could be wrong (see Figure 2), we employ a verification network to determine whether accept or reject the segmentation refinement on a region-by-region basis. We choose DeepLabv3+ [32] as the base segmentation network and thus build the SESV-DLab model. Experimental results show that this model achieves advanced performance in the segmentation of gland cells, skin lesions, and retinal microaneurysms. We also evaluate the proposed SESV framework that uses other well-established segmentation networks, such as PSPNet [33], U-Net [15], and FPN [34], and consistent performance improvements are observed.

The pilot data of this research was presented in MICCAI 2019 [14]. In this paper, we reported a completely new solution with much improved performance. The major differences are two-fold. First, we take the inaccuracy of predicted segmentation error maps into consideration, and hence abandon the idea of using these maps to correct the segmentation results directly. In contrast, we treat these maps as the prior that indicates the locations of segmentation errors and employ them to perform re-segmentation for refinement. Second, we also pay extra attention to the inaccuracy of segmentation refinement, and

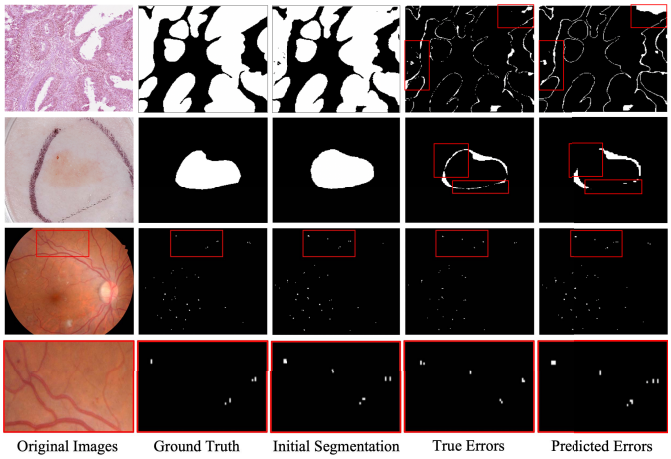

Fig. 2. Three examples and their predicted segmentation error maps, which could be incorrect in some regions (highlighted by red rectangles).

accordingly add a verification process to determine whether accept or reject the 'refined' results. Meanwhile, the proposed SESV framework has been evaluated comprehensively by using four base segmentation networks and on three medical image segmentation tasks. Our contributions include:

- We propose the SESV framework to improve the accuracy of existing medical image segmentation models via segmentation error prediction, error-guided re-segmentation, and refinement verification.
- To make our SESV framework tolerate less-accurate prediction of segmentation errors, we treat each predicted error map as the prior that indicates the locations of segmentation errors, use the error map as part of the input to perform re-segmentation, and then construct a verification network to reject incorrect 'refinements'.

II. Related Works

A. Medical Image Segmentation

A huge number of DCNN-based methods have been proposed for medical image segmentation [3]–[29], [35], [36]. In this section, we mainly review the approaches to the segmentation of glands, skin lesions, and retinal microaneurysms, which are closely related to our study.

1) Gland Segmentation: Accurate segmentation of glands on histology microscopy images is an effective means to assist pathologists in diagnosing the malignancy of adenocarcinoma [37]. Currently, many advanced gland segmentation methods are based on DCNNs [9]–[14]. Most of them attempt to improve the segmentation performance by exploiting the contour information from different perspectives or by designing advanced network architectures and loss functions to preserve the spatial information. Chen *et al*. [10] developed a deep contour-aware network (DCAN) to learn the regions and contours of glands simultaneously for gland segmentation. Graham *et al*. [11] proposed the minimal information loss dilated network (MILD-Net), which counters the loss of information caused by max-pooling and uses atrous spatial pyramid pooling for resolution maintenance and multi-level aggregation of features.

2) Skin Lesion Segmentation: Accurate detection of the boundaries of skin lesions on dermoscopic images can help remove distractions from those images, and thus is highly beneficial for improving the accuracy of skin lesion diagnosis [5], [38]. Many skin lesion segmentation methods have been proposed in the literature [3]–[7], [39]–[43]. Among them, those based on DCNNs have achieved tremendous success [3]–[7], [38]–[40]. Lei *et al*. [40] proposed a novel generative adversarial network for skin lesion segmentation, which consists of a skip connection and dense convolution U-Net based segmentation module and a dual discrimination module. One discriminator focuses on the segmentation errors at lesion boundaries, and the other examines the contextual environment of skin lesions. Mirikharaji and Hamarneh [39] proposed a new loss term that encodes the star shape prior into the loss function of an end-to-end trainable fully convolutional network (FCN) framework, and thus guaranteed a global structure in segmentation results.

3) Retinal Microaneurysms Segmentation: Microaneurysms are the earliest clinically visible changes of diabetic retinopathy (DR) [44]. Accurate microaneurysms segmentation in retinal images is conducive to the early detection of DR and hence can reduce the chances of severe vision loss [45]. Currently, most microaneurysms segmentation methods are based on DCNNs [46]–[50]. Kou *et al*. [48] reported the deep recurrent U-Net (DRU-Net) that incoporates the deep residual model and recurrent convolutional operations into U-Net for microaneurysms segmentation. Sarhan *et al*. [49] introduced a two-stage deep learning approach to microaneurysm segmentation. In this approach, microaneurysms are first segmented at two scales and the results are then refined by a classification network, into which the triplet embedding loss with a selective sampling routine is incorported. Zhou *et al*. [46] proposed a collaborative learning model to jointly improve the performance of DR grading and microaneurysm segmentation, in which the lesion attention module can refine the lesion maps using class-specific information to fine-tune the segmentation module in a semi-supervised manner.

Although these DCNN-based methods have advanced, more or less, the performance of medical image segmentation, which, however, has not been considered as satisfactory in the scenario of clinical practices [30].

B. Segmentation Failure Estimation and Correction

To relieve the adverse effects of segmentation errors, many approaches to segmentation failure estimation and correction have been proposed [14], [30], [31]. Wang *et al*. [31] proposed a deep learning-based interactive segmentation method to improve segmentation results. They first used user interactions to indicate mis-segmentations and then fed the user interactions and initial segmentation to another DCNN for segmentation refinement. To avoid the expensive user interactions and design a fully automated method, we proposed the deep segmentation-emendation (DSE) model in our pilot study [14]. This model consists of a segmentation network and an emendation network. The former generates segmentation results, and the latter learns to predict over- and under-segmented

regions in each result, which are then used to revise the result for better accuracy. Predicting segmentation errors, however, remains challenging. Hence, directly using the predicted over- and under-segmented regions to revise a segmentation result may inherit the mistakes mad in the final output. In this study, we take the inaccuracy of predicted errors into consideration and treat predicted segmentation errors as the prior that indicates the locations where potential segmentation errors in a re-segmentation process [31]. Even the prior is wrong, we have a verification network to reject incorrect segmentation 'refinements'. Thus, the novel solution can tolerate some mistakes in predicted errors.

III. Method

A. Overview of the SESV Framework

The proposed SESV framework consists of a segmentation network S_{net}, an emendation network E_{net}, a re-segmentation network rS_{net} and a verification network V_{net}. Under this framework, the pipeline of medical image segmentation is composed of four major steps (see Figure 1). First, we apply S_{net} to a medical image X to generate an initial segmentation mask $\tilde{Y}_S$. Then, we feed the concatenation of X and $\tilde{Y}_S$ to E_{net} to predict a segmentation error map $\tilde{Y}_E$. Next, we use $\tilde{Y}_E$ as the prior of the locations of segmentation errors and feed the concatenation of X, $\tilde{Y}_S$, and $\tilde{Y}_E$ to rS_{net} to generate a refined segmentation mask $\tilde{Y}_{rS}$. Finally, we use V_{net} to determine whether accept or reject $\tilde{Y}_{rS}$ on a region-by-region basis.

B. Segmentation Network S_{net}

The proposed SESV framework is generic and can be applied to any segmentation models. In this study, we choose the DeepLabv3+ [32] pre-trained on the MS-COCO [51] and PASCAL VOC datasets [52] as the segmentation network S_{net} and fine-tune it to generate the initial segmentation mask $\tilde{Y}_S$.

C. Emendation Network E_{net}

The true segmentation error $\Delta(\tilde{Y}_S, Y)$ is the discrepancy between an initial segmentation mask $\tilde{Y}_S$ and the corresponding ground truth Y, defined by

$$\Delta_i(\tilde{Y}_{s_i}, Y_i) = \begin{cases} 0 & \text{if } \tilde{Y}_{s_i} = Y_i \\ 1 & \text{if } \tilde{Y}_{s_i} \neq Y_i \end{cases}, \tag{1}$$

where $\tilde{Y}_{s_i}$ and Y_i are the class label of the i-th pixel in $\tilde{Y}_S$ and Y, respectively. The $\Delta_i(\tilde{Y}_{s_i}, Y_i) = 0$ means that the initial pixel label $\tilde{Y}_{s_i}$ is correct, whereas $\Delta_i(\tilde{Y}_{s_i}, Y_i) = 1$ means incorrect. We focus only on predicting whether a pixel is correctly or incorrectly segmented in $\tilde{Y}_S$, which is simpler than predicting the over- or under-segmented pixels [14] and can alleviate the burden of training E_{net}.

The emendation network E_{net} takes the concatenation of the image X and initial segmentation mask $\tilde{Y}_S$ as its input, and hence has an extra input channel for $\tilde{Y}_S$. We use the fine-tuned S_{net} as the backbone of E_{net} and average the parameters for the RGB channels of S_{net} to initialize the $\tilde{Y}_S$ channel of E_{net}, as done in [38], [53]. The network E_{net} aims to learn and predict the segmentation error, which is formulated as

$$\tilde{Y}_E = E_{net}(X \odot \tilde{Y}_S; \Theta_E), \tag{2}$$

where $\odot$ denotes the concatenation operation, and Θ_E is the parameters of E_{net}. This network can be optimized via minimizing the difference between the predicted segmentation error $\tilde{Y}_E$ and true error $\Delta(\tilde{Y}_S, Y)$, shown as follows

$$\min_{\Theta_E} \mathcal{L}_{E_{net}}(\tilde{Y}_E, \Delta(\tilde{Y}_S, Y)), \tag{3}$$

Considering the class-imbalance issue in each $\Delta(\tilde{Y}_s, Y)$ (see Figure 3), we use the following combined class-weighted Dice loss and cross-entropy loss

$$\mathcal{L}_{E_{net}} = \frac{1}{N}\sum_{n=1}^{N} \omega_n \left\{ \left[1 - \frac{2\sum \tilde{Y}_E^n \Delta(\tilde{Y}_s^n, Y^n)}{\sum(\tilde{Y}_E^n + \Delta(\tilde{Y}_s^n, Y^n)) + \varepsilon} \right] - \mathbb{E}\left[\Delta(\tilde{Y}_s^n, Y^n) log \tilde{Y}_E^n \right] \right\}, \tag{4}$$

where N is the number of classes, $\mathbb{E}$ is the expectation operator, ε is a smoothing factor, and ω_n is the weight of class n. The weight ω_n is adaptively updated during the iterative process according to the following rule

$$\omega_n = log \frac{\sum_{n=1}^{N} V^n}{V^n}, \tag{5}$$

where V^n is the number of pixels in class n.

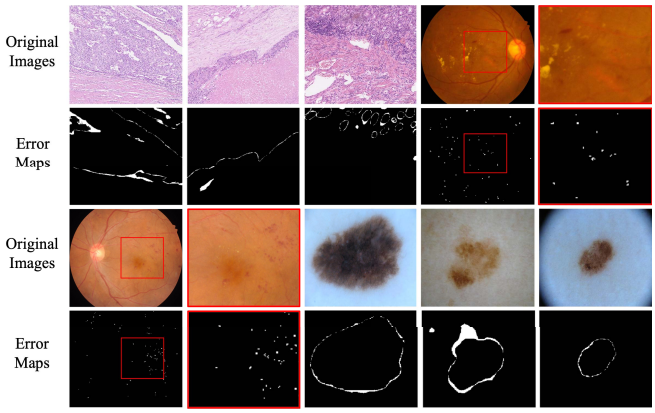

Fig. 3. Visualization of segmentation error maps $_{1\triangleright}\tilde{Y}_S, Y_{\triangleleft}$ chosen from three different datasets. It shows that incorrectly segmented pixels (with a value of 1) are significantly less than correctly segmented pixels (with a value of 0) in each $_{1\triangleright}\tilde{Y}_S, Y_{\triangleleft}$.

D. Re-segmentation Network rS_{net}

Due to the difficulties of training E_{net}, the predicted segmentation error map $\tilde{Y}_E$ may not be accurate enough for emendating the initial segmentation mask $\tilde{Y}_S$. However, although less-accurate, the $\tilde{Y}_E$ can provide the prior locations of possible segmentation errors. Incorporating such a prior into the segmentation process is definitely beneficial for generating accurate segmentation masks. Therefore, we construct a re-segmentation network rS_{net}, which takes the concatenation

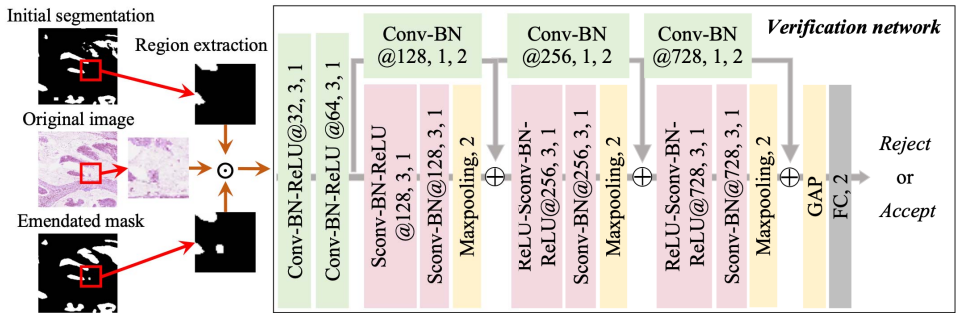

Fig. 4. Architecture of V_{net}. 'Conv': convolutional layers; 'Sconv': separable convolutional layers; @*a*, *b*, *c*: the number of filters *a*, kernel size *b* and stride *c*; Yellow: 2×2 max-pooling layers or global average pooling (GAP) layers; 'BN-ReLU': batch normalization (BN) and ReLU activation; 'FC': fully connected layers.

of the original image X, initial segmentation mask $\tilde{Y}_S$, predicted segmentation error map $\tilde{Y}_E$ as its input, to generate a refined segmentation mask $\tilde{Y}_{rS}$, shown as follows

$$\tilde{Y}_{rS} = rS_{net}(X \odot \tilde{Y}_S \odot \tilde{Y}_E; \Theta_{rS}), \tag{6}$$

where Θ_{rS} denotes the parameters of rS_{net}.

We use the trained S_{net} as the backbone of rS_{net}. Similarly, we use the average parameters in the RGB channels of S_{net} to initialize the $\tilde{Y}_S$ channel and $\tilde{Y}_E$ channels of rS_{net} [38], [53]. The re-segmentation network rS_{net} is optimized via minimizing the the discrepancy between the refined segmentation mask $\tilde{Y}_{rS}$ and the ground truth Y, shown as follows

$$\min_{\Theta_{rS}} \mathcal{L}_{rS_{net}}(\tilde{Y}_{rS}, Y). \tag{7}$$

Since the expected outputs of S_{net} and rS_{net} are the same, *i.e.*, the ground truth Y, we employ the same loss function for both networks. The loss function is designed according to the specific medical image segmentation task.

E. Verification Network V_{net}

If the predicted erroneous regions are completely wrong, using such predictions as part of the input may result in segmentation errors, instead of corrections. To avoid possible deterioration of the initial segmentation mask, we construct a verification network V_{net} to determine whether accept the refined segmentation mask as the final result or not. Since there might be both correct and incorrect predictions on each error map $\tilde{Y}_E$, such verification must be performed on a region-by-region basis.

The verification network V_{net} contains two normal convolutional layers, three max-pooling layers, six separable convolutional layers, three shortcuts of 1×1 convolution, and a fully-connected classification layer with two output units followed by *softmax* (see Figure 4). Since we have only a limited number of training images, the depthwise separable convolutions [54] are used to reduce the number of model parameters, hence limiting overfitting.

In the training stage, we first randomly extract small regions $\tilde{R}_S$, $\tilde{R}_{rS}$ and $\tilde{R}_X$ from the initial segmentation mask $\tilde{Y}_S$, refined segmentation mask $\tilde{Y}_{rS}$ and original image X, respectively, and then concatenate those regions as the input of V_{net}. The expected output of V_{net} indicates which region is more consistent with the ground truth, formally expressed as follows

$$\mathbb{M} = \begin{cases} 0 & \text{if } Acc(\tilde{R}_{rS}, R_Y) \geq Acc(\tilde{R}_S, R_Y) \\ 1 & \text{if } Acc(\tilde{R}_{rS}, R_Y) < Acc(\tilde{R}_S, R_Y) \end{cases}, \tag{8}$$

where R_Y is the region extracted from the ground truth Y, $Acc(R_1, R_2)$ represents the pixel-level accuracy between R_1 and R_2. Thus, $\mathbb{M} = 1$ means the region $\tilde{R}_S$ is more accurate than $\tilde{R}_{rS}$, whereas $\mathbb{M} = 0$ means the opposite. We optimize V_{net} via minimizing the following cross-entropy loss

$$\min_{\Theta_V} \mathcal{L}_{V_{net}}(V_{net}(\tilde{R}_S \odot \tilde{R}_{rS} \odot \tilde{R}_X; \Theta_V), \mathbb{M}), \tag{9}$$

where $\tilde{Y}_V = V_{net}(\tilde{R}_S \odot \tilde{R}_{rS} \odot \tilde{R}_X; \Theta_V)$ is the prediction made by V_{net}.

In the inference stage, a test image X of size $H \times W$ and its predicted binary segmentation masks $\tilde{Y}_S$ and $\tilde{Y}_{rS}$ are first divided into partly-overlapped regions of size $L_R \times L_R$, denoted by $\tilde{R}_X$, $\tilde{R}_S$ and $\tilde{R}_{rS}$, respectively. Then, the $\tilde{R}_S$, $\tilde{R}_{rS}$ and $\tilde{R}_X$ are concatenated as input of the trained V_{net}. Let the output of V_{net} be denoted by $\tilde{Y}_V$, the optimal regions O_R can be obtained as follows

$$O_R = \tilde{Y}_V \times \tilde{R}_S + (1 - \tilde{Y}_V) \times \tilde{R}_{rS}. \tag{10}$$

The final segmentation result of X is generated by recomposing and averaging the optimal regions.

IV. Experiments

A. Materials and Evaluation Metrics

We applied the proposed SESV-DLab model to three medical image segmentation tasks.

1) Gland Segmentation: We used the public ColoRectal Adenocarcinoma Gland (CRAG) dataset [11] that contains 173 training and 40 testing images. To assess the segmentation performance, we calculated three metrics suggested in [11], including the object-level F1 score (O-F1) that evaluates the accuracy of detecting individual glands, the object-level Dice (O-Dice) that represents the accuracy of segmenting individual glands, and the object-level Hausdorff distance (O-Hausdorff) that measures the shape similarity of each individual gland.

2) Skin Lesion Segmentation: The datasets used for this task were provided by the International Skin Imaging Collaboration (ISIC) skin lesion segmentation challenges held in 2017 [55] and 2018 [56], [57]. The ISIC-2017 dataset contains 2000 training, 150 validation, and 600 test dermoscopic images. The ISIC-2018 dataset consists of 2594 training and 1000 test images. To evaluate segmentation results, we adopted the metrics suggested by the challenge organizers [55], including the Jaccard index (Jac), Dice coefficient (Dice), pixel-wise accuracy (Acc), pixel-wise sensitivity (Sen), and pixel-wise specificity (Spe). Note that the ISIC-2018 challenge uses a new metric, *i.e.*, the threshold of Jac (Jac_th), to rank the segmentation performance of each method on the test dataset. If Jac is less than 0.65, Jac_th equals 0, otherwise, Jac_th equals the Jac.

3) Retinal Microaneurysms Segmentation: For this task, we used the Indian Diabetic Retinopathy Image Dataset (IDRiD) [58], which contains 81 fundus images including 54 images for training and 27 images for testing. We computed the area under the precision-recall curve (AUC-PR), which was used in the IDRiD challenge, and the area under the receiver operating characteristic curve (AUC-ROC) to quantitatively evaluate the results of microaneurysms segmentation.

TABLE I
PERFORMANCE OF FIVE GLAND SEGMENTATION METHODS ON THE CRAG DATASET

Methods	O-F1(%)	O-Dice(%)	O-Hausdorff
DCAN, 2016 [10]	73.60	79.40	218.76
MILD-Net, 2019 [11]	82.50	87.50	160.14
DSE, 2019 [14]	83.50	88.90	120.13
DeepLabv3+ (base) [32]	77.89±0.38	85.01±0.83	151.51±2.85
SESV-DLab (mean±std)	**85.43**±0.35	**89.90**±0.62	**106.10**±2.70

TABLE II
PERFORMANCE OF SEVEN SKIN LESION SEGMENTATION METHODS ON THE ISIC-2017 DATASET

Methods	Jac(%)	Dice(%)	Acc(%)	Sen(%)	Spe(%)
CDNN, 2017 [59]	76.50	84.90	93.40	82.50	97.50
DDN, 2018 [7]	76.50	86.60	93.90	82.50	**98.40**
SCDC, 2020 [40]	77.10	85.90	93.50	83.50	97.60
SAL, 2019 [6]	77.14	85.16	-	-	-
SSP, 2018 [39]	77.30	85.70	93.80	85.50	97.30
DeepLabv3+ (base) [32]	77.62 ±0.12	85.83 ±0.09	93.82 ±0.04	87.84 ±0.20	96.13 ±0.17
SESV-DLab (mean± std)	**78.77** ±0.10	**86.79** ±0.08	**94.14** ±0.03	**88.35** ±0.15	95.70 ±0.05

B. Implementation Details

Due to the extreme class-imbalance issue presented in the IDRiD dataset, we chose the segmentation loss shown in Eq. 4 for this dataset. As for the CRAG and ISIC-2017 / 2018 datasets, we followed the suggestion in [14], [38] and used the cross-entropy loss and hybrid loss as the segmentation loss, respectively.

In training stage, we followed the suggestion in [14], [46] and randomly cropped 512 × 512 patches on the CRAG dataset and 640 × 640 patches on the IDRiD dataset as the input. On the ISIC-2017 / 2018 datasets, we followed the suggestion in [38] and resized the images to 224 × 224 as the input. To alleviate the over-fitting of DCNNs, we employ the online data argumentation, including the random rotation, shear, shift, zooming, and horizontal / vertical flip, to enlarge our training dataset. We optimized all networks with the Adam algorithm and empirically set the initial learning rate to 0.0001, 0.00005 and 0.0001, the batch size to 3, 16 and 3, and the hyper-parameters L_R in V_{net} to 96, 128 and 96 on the CRAG, ISIC-2017 / 2018 and IDRiD datasets, respectively.

In the testing stage, each image was segmented as follows. On the CRAG dataset, we extracted 512 × 512 patches from each test image with a stride of 256 pixels, recomposed the segmented patches, and averaged the predictions on a pixel-by-pixel basis to generate the segmentation result. We also employed the morphological opening with a 10 × 10 square structure element to smooth segmentation results. Finally, the connected component labeling operation was used to label the region of each gland instance for evaluation. On the IDRiD dataset, we extracted 640 × 640 patches from each test image with a stride of 320 pixels, recomposed the segmented patches, and averaged the predictions on a pixel-by-pixel basis to generate the segmentation result. On the ISIC-2017 / 2018 datasets, we resized each test image to 224 × 224 and fed it to the trained SESV-DLab model to generate the segmentation result, which was then resized to the original size for evaluation.

C. Experiment Results

1) Gland Segmentation: We compared our SESV-DLab model against the base segmentation model (*i.e.*, the pre-trained DeepLabv3+ [32]), MILD-Net [11], DCAN [10], and DSE model [14] on the CRAG dataset. The average performance of these models is given in Table I. It reveals that our SESV-DLab model performs better than other models in terms of all evaluation metrics. Specifically, our model remarkably improves O-F1 by 7.54%, improves O-Dice by 4.89%, and reduces O-Hausdorff by 29.97 over the base model. Even comparing to the state-of-the-art DSE model, which directly used the predicted errors to emendate segmentation results, our model also substantially improves O-F1 by 1.93%, improves O-Dice by 1.00%, and reduces O-Hausdorff by 11.68. Such results not only prove the effectiveness of the proposed SESV strategy, but also suggest our SESV-DLab model is able to partly overcome the drawbacks of the DSE model.

2) Skin Lesion Segmentation: Table II gives the performance of our SESV-DLab model, the base segmentation model [32], and five recent skin lesion segmentation methods, including the convolutional-deconvolutional neural network (CDNN) [59], dense deconvolutional network (DDN) [7], skip connection and dense convolution U-Net (SCDC) [40], fully convolutional network with a star shape prior (SSP) [39], and stacked adversarial learning (SAL) model [6] on ISIC-2017 dataset. It shows that our model achieves the highest Jaccard index, Dice score, accuracy, and sensitivity, though its specificity is slightly lower than that of other

TABLE III
PERFORMANCE OF TEN SKIN LESION SEGMENTATION METHODS ON THE ISIC-2018 DATASET, INCLUDING THE PERFORMANCE OF SEVEN TOP-RANKING SOLUTIONS IN THE 2018 ISIC CHALLENGE LEADERBOARD

Methods	Evaluation metrics (%)					
	Jac_th	Jac	Dice	Acc	Sen	Spe
Rank #1	83.6	85.2	91.5	95.4	95.6	94.1
Rank #2	83.2	85.4	91.4	95.2	93.7	94.6
Rank #3	82.5	84.7	91.1	95.1	95.2	94.1
Rank #4	80.4	84.1	90.4	94.5	92.0	94.4
Rank #5	80.4	84.4	90.8	94.7	94.2	94.1
SESV-DLab	80.3	83.3	90.2	94.6	96.2	92.5
Rank #7	80.2	82.5	89.3	94.0	90.4	96.3
Deeplabv3+ (base) [32]	78.6	82.5	89.6	94.2	96.2	92.1
ACA-net, 2020 [60]	77.1	81.9	89.1	N/A	94.3	93.2
SCDC, 2020 [40]	N/A	82.4	88.5	92.9	95.3	91.1

methods. Comparing to the base model, our model improves the average Jaccard index by 1.15%, Dice score by 0.96%, accuracy by 0.32%, and sensitivity by 0.51%, but deteriorates the average specificity by 0.43%. The substantial performance gains over the base model and five recent solutions indicate the superiority of the proposed SESV-DLab model.

We performed the ten-fold cross-validation on the ISIC-2018 dataset. Specifically, we divided the ISIC-2018 training dataset into ten folds. Each time, we used nine folds to train our SESV-DLab model and the other fold to validate it. In the testing stage, the average result of ten trained models is treated as the final result. We evaluated our SESV-DLab model against the base segmentation model (*i.e.*, Deeplabv3+), two recently published methods (*i.e.*, adaptive color augmentation-based DCNN (ACA-net) [60] and SCDC [40]), and six top-ranking solutions (except for ours) in the 2018 ISIC skin lesion segmentation challenge leaderboard.[1] The performance of these methods was listed in Table III. It shows that, excluding the top-ranking solutions in the leaderboard, our model achieves the highest Jac_th of 80.3%, highest Jac of 83.3%, highest Dice of 90.2%, highest Acc of 94.6%, highest Sen of 96.2%, and second highest Spe of 92.5%. Comparing to the solutions in the leaderboard, our model ranks sixth according to Jac_th. Since the top five solutions in the leaderboard were not published yet, our model holds the state-of-the-art segmentation performance among all published methods.

3) Retinal Microaneurysms Segmentation: On the IDRiD dataset, we compared our SESV-DLab model to the base segmentation model [32], advanced semi-supervise collaborative learning (SSCL) model [46], DRU-Net [48], and three top-ranking methods on the IDRiD challenge leaderboard[2] (*i.e.*, iFLYTEK-MIG, VRT, and PATech). The performance of these methods was shown in Table IV. It reveals that our model achieves the highest AUC-PR and AUC-ROC. Particularly, if measuring the performance in terms of the officially recommended metric AUC-PR, our model beats the base model by 2.34% and beats the top-ranking method by 0.82%, setting the new state of the art.

[1]https://challenge2018.isic-archive.com/live-leaderboards/
[2]https://idrid.grand-challenge.org/Leaderboard/

TABLE IV
PERFORMANCE OF SEVEN MICROANEURYSMS SEGMENTATION METHODS ON THE IDRID DATASET

Methods	AUC-PR(%)	AUC-ROC(%)
iFLYTEK-MIG (Rank #1)	50.17	N/A
VRT (Rank #2)	49.51	N/A
PATech (Rank #3)	47.40	N/A
DRU-Net [48]	N/A	98.20
SSCL [46]	49.60	98.28
DeepLabv3+ (base) [32]	48.65±0.15	98.91±0.25
SESV-DLab (mean±std)	**50.99**±0.14	**99.13**±0.13

The results obtained in these three experiments suggest that the proposed SESV-DLab model can effectively improve the accuracy of the base segmentation model, *i.e.*, DeepLabv3+, and can achieve advanced performance on different medical image segmentation tasks.

V. DISCUSSIONS

A. Contribution Made by Each of Four Networks

The proposed SESV framework is composed of four networks, including a segmentation network S_{net}, an emendation network E_{net}, a re-segmentation network rS_{net}, and a verification network V_{net}. To verify the contribution made by each of them, we constructed five variants of our SESV-DLab model and conducted ablation studies on the aforementioned three datasets. Model I is the base segmentation network S_{net} (*i.e.*, DeepLabv3+). Model II is the DSE model [14], which uses the segmentation error map predicted by the emendation network E_{net} to emendate the initial segmentation mask produced by the base network S_{net}. Model III consists of the base segmentation network S_{net} and re-segmentation network rS_{net}. In this model, the concatenation of the image and initial segmentation mask is fed to rS_{net} to generate a refined segmentation mask. Model IV is constructed by introducing the verification network V_{net} to Model III to verify the segmentation refinement. Model V is similar to our SESV-DLab model, except for not using the verification network V_{net}. Table V shows the segmentation performance of these five Models and our SESV-DLab model, based on which three conclusions can be drawn.

First, it shows that performing emendation directly in the inference stage (Model II) results in a lower performance than using our re-segmentation strategy (Model V), and is even worse than the base model (Model I) on the ISIC-2017 and IDRiD datasets. The poor segmentation performance is attributed mainly to the inevitable mistakes produced by E_{net}, which lead the initial segmentation mask to be emendated incorrectly, particularly on the IDRiD dataset where most microaneurysms are very small. Note that, if more than half of mis-segmented pixels predicted by E_{net} are truly mis-segmented, Model II can improve the initial segmentation result; otherwise, it deteriorates the initial segmentation result. Second, the superior performance of Model V over Model III suggests that using the predicted error map as the prior that indicates the locations of potential segmentation errors makes the re-segmentation more accurate than the initial segmentation. It also means that our re-segmentation strategy is able to

TABLE V
PERFORMANCE OF OUR SESV-DLAB MODEL UNDER DIFFERENT SETTINGS ON THREE DATASETS

Models	Networks				CRAG			ISIC-2017					IDRiD
	S_{net}	E_{net}	rS_{net}	V_{net}	O-F1	O-Dice	O-Hausdorff	Jac	Dice	Acc	Sen	Spe	AUC-PR
I	√	×	×	×	77.89	85.01	151.51	77.62	85.83	93.82	87.84	96.13	48.65
II	√	√	×	×	83.60	89.29	111.52	76.57	85.20	93.80	85.64	95.96	33.34
III	√	×	√	×	80.23	87.85	123.35	77.88	86.05	93.82	84.25	**97.52**	49.25
IV	√	×	√	√	83.27	88.17	119.84	78.11	86.13	93.89	85.86	96.57	49.87
V	√	√	√	×	84.44	89.67	108.84	78.30	86.45	94.00	88.08	95.58	50.21
SESV-DLab	√	√	√	√	**85.43**	**89.90**	**106.10**	**78.77**	**86.79**	**94.14**	**88.35**	95.70	**50.99**

TABLE VI
PERFORMANCE OF OUR SESV-DLAB MODEL ON THREE DATASETS WHEN USING DIFFERENT REGION WIDTH L_R

L_R	CRAG			ISIC-2017					IDRiD
	O-F1	O-Dice	O-Hausdorff	Jac	Dice	Acc	Sen	Spe	AUC-PR
64	84.78	89.85	107.72	78.55	86.65	94.07	88.18	95.69	50.72
96	**85.43**	89.90	**106.10**	78.63	86.69	94.10	**88.52**	95.50	**50.99**
128	84.62	**89.91**	108.79	**78.77**	**86.79**	**94.14**	88.35	**95.70**	50.76
160	84.72	89.83	108.89	78.69	86.72	94.11	88.41	95.61	50.74

TABLE VII
PERFORMANCE IMPROVEMENT OBTAINED BY APPLYING OUR SESV FRAMEWORK TO THREE SEGMENTATION MODELS ON THREE DATASETS

S_{net}	CRAG			ISIC-2017					IDRiD
	O-F1	O-Dice	O-Hausdorff	Jac	Dice	Acc	Sen	Spe	AUC-PR
PSPNet [33]	**69.47**	77.15	210.92	70.19	79.87	92.21	77.62	**97.82**	46.03
SESV-PSP	69.01	**79.70**	**192.53**	**73.00**	**82.72**	**93.10**	**83.49**	95.61	**47.54**
U-Net [15]	77.20	82.03	180.36	76.36	84.60	93.47	80.98	**98.26**	46.36
SESV-Unet	**78.17**	**86.67**	**135.08**	**78.19**	**86.14**	**94.17**	**84.40**	97.48	**48.21**
FPN [34]	73.21	81.38	196.31	75.88	84.16	93.23	80.94	**98.18**	48.42
SESV-FPN	**76.29**	**83.98**	**153.70**	**78.19**	**86.29**	**94.18**	**85.07**	96.57	**50.54**

ease the negative effect caused by the mistakes in the predicted error maps. Third, the performance gain of our SESV-DLab model over Model V and the gain of Model IV over Model III demonstrate the effectiveness of the verification network V_{net}.

Figure 5 visualizes, from left to right, seven images from three datasets, the segmentation results obtained by Model II, the initial and refined segmentation masks, the segmentation results obtained by our SESV-DLab model, and the ground truth. It shows that using rS_{net} can correct some inaccurately segmented regions (highlighted by red rectangles) and using V_{net} can filter some incorrectly emendated regions (highlighted by blue rectangles). Meanwhile, it reveals that the results produced by our model are more similar to the ground truth than the results produced by Model II. It demonstrates again the effectiveness of each network in the proposed SESV framework.

B. Size of Verification Regions

The verification network V_{net} determines whether to accept or reject the segmentation refinement on a region-by-region basis. In V_{net}, the size of verification regions $L_R \times L_R$ is a critical hyper-parameter. A smaller L_R may result in the loss of context information during the training of V_{net}, while a large L_R may lead to discontinuity in the final output. Thus, the setting of L_R represents the trade-off between reducing the discontinuity and preserving the context information. To search for an optimal L_R, we attempted to set L_R from 64 to 160 with an interval of 32 and recorded the segmentation performance obtained on three datasets in Table VI. It shows that the performance of our SESV-DLab model is relatively robust to the variation of L_R, and there is somewhat a correlation between the value of L_R and the size of to-be-segmented targets. Particularly, our SESV-DLab model obtains the highest accuracy on the CRAG, IDRiD, and ISIC-2017 / 2018 datasets when setting L_R to 96, 96, and 128, respectively. We thereby adopted such settings in our experiments.

C. Using Other Segmentation Networks

In our pervious experiments, we adopted DeepLabv3+ as the base segmentation network S_{net}. In this section, we attempted to use PSPNet [33], U-Net [15], and FPN [34] as S_{net}, respectively, aiming to justify the generality of our SESV framework. The backbone of these three networks is ResNet50 [61] pre-trained on ImageNet dataset [62]. The results are shown in Table VII. It reveals that, no matter which segmentation network or which dataset was used, using the proposed SESV framework can steadily and consistently produce more accurate medical image segmentation than using the segmentation network alone. Therefore, we believe that our SESV framework is generic and can be use to improve the performance of other medical image segmentation models.

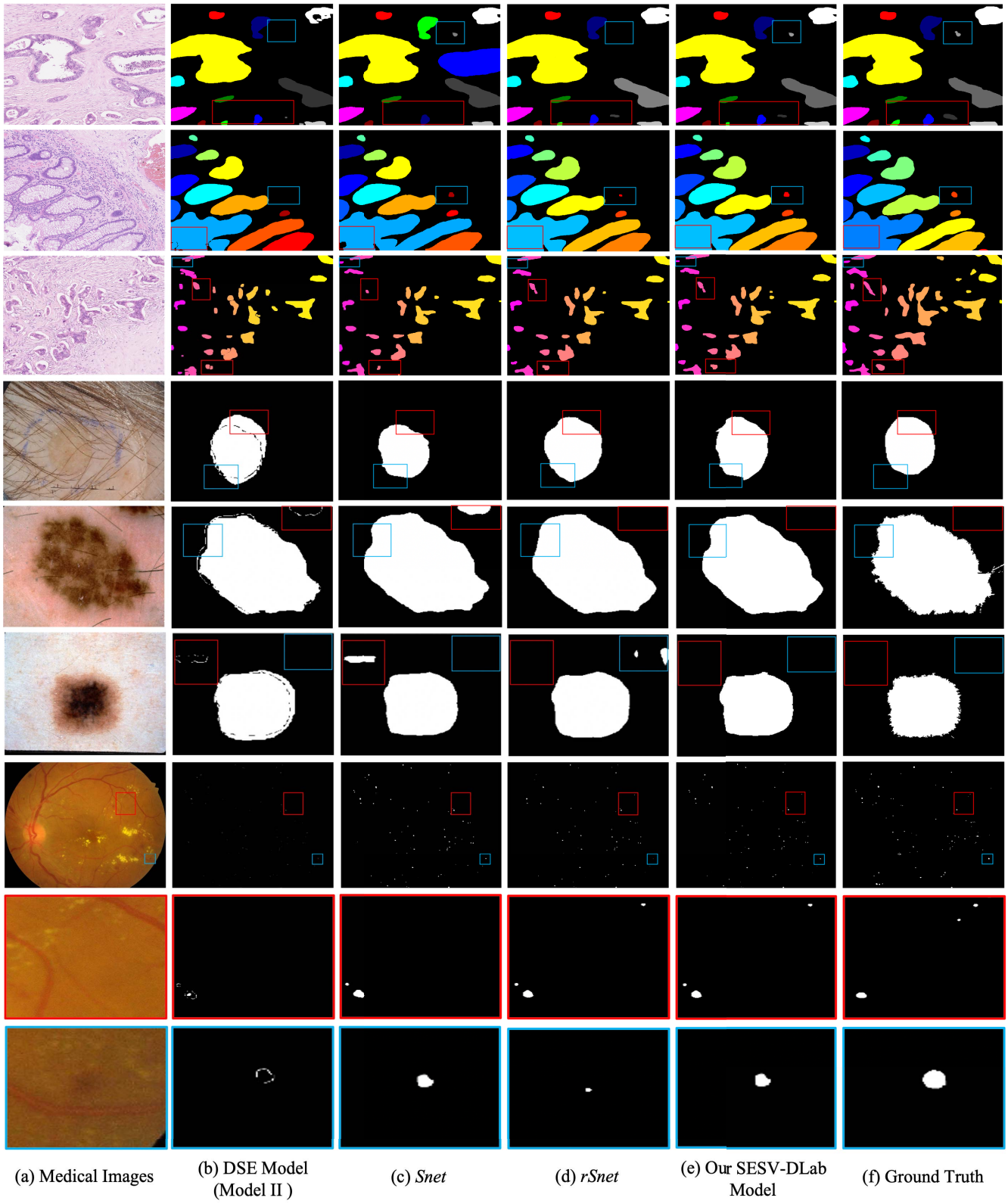

Fig. 5. Qualitative evaluation of segmentation results obtained on three datasets. (a) Medical images; (b) Segmentation results generated by Model II that is similar to the DSE model [14]; (c) Initial segmentation masks predicted by S_{net}; (d) Refined segmentation masks produced by rS_{net}; (e) Our final segmentation results, each combining the correctly segmented regions in both initial and refined segmentation masks; and (f) Ground truth. The red rectangles denote that using the re-segmentation network can correct some regions mis-segmented by the initial segmentation network. The blue rectangles denote that using the verification network can filter some regions incorrectly 'refined' by the re-segmentation network.

D. Computational Complexity

The proposed SESV-DLab model was implemented using the Keras and Tensorflow software packages, and was evaluated on the CRAG and ISIC datasets using a desktop with a NVIDIA GTX 2080 Ti GPU and on the IDRiD dataset using a workstation with a NVIDIA Tesla P100 GPU. On the CRAG dataset, it took us about 48 hours to train our model (*i.e.*, 15 hours for S_{net}, 20 hours for E_{net}, 8 hours for rS_{net}, and 5 hours for V_{net}). On the ISIC dataset, it took us about 55 hours to train our model (*i.e.*, 20 hours for S_{net}, 20 hours for E_{net}, 10 hours for rS_{net}, and 5 hours for V_{net}). On the IDRiD dataset, it took us about 40 hours to train our model (*i.e.*, 14 hours for S_{net}, 12 hours for E_{net}, 10 hours for rS_{net}, and 4 hours for V_{net}).

Due to containing four DCNNs, our SESV-DLab model has a relatively high spatial and computational complexity. Particularly, the time cost of training our model is about three times of the cost of training the base model. This is a major disadvantage of our model and should be addressed in our future work. Nevertheless, the inference of our model is very efficient, costing less than 1 second to segment an image of size 512×512. The improved segmentation accuracy and efficient inference suggests that our SESV-DLab model could be better used in a routine clinical workflow than other segmentation approaches.

VI. Conclusion

This paper proposes a novel and generic framework called SESV to improve the accuracy of DCNN-based medical image segmentation models. Taking DeepLabv3+, a state-of-the-art image segmentation model, as a case study, we demonstrate that using our SESV framework can substantially improve the accuracy of DeepLabv3+ and can achieve advanced performance in the segmentation of gland cells on the CRAG dataset, skin lesions on the ISIC-2017 and ISIC-2018 datasets, and retinal microaneurysms on the IDRiD dataset. Further studies show that, for other segmentation models such as PSPNet, U-Net, and FPN, substantial performance gains can also be achieved on those three datasets by using our SESV framework. In the future, we will extend this framework to semi- and weakly-supervised settings, and thus reduce the requirement of pixel-wise image annotations.

Acknowledgment

The authors declare that there is no conflict of interests regarding the publication of this article. Part of this work was done when Yutong Xie and Jianpeng Zhang were visiting students and Hao Lu was a postdoctoral fellow with the School of Computer Science, The University of Adelaide.

References

[1] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annu. Rev. Biomed. Eng., Annu. Rev.*, vol. 2, no. 1, pp. 315–337, 2000.

[2] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[3] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2065–2074, Sep. 2017.

[4] M. M. K. Sarker *et al.*, "Slsdeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2018, pp. 21–29.

[5] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876–1886, Sep. 2017.

[6] L. Bi, D. Feng, M. Fulham, and J. Kim, "Improving skin lesion segmentation via stacked adversarial learning," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1100–1103.

[7] H. Li *et al.*, "Dense deconvolutional network for skin lesion segmentation," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 527–537, Mar. 2019.

[8] P. Naylor, M. Lae, F. Reyal, and T. Walter, "Segmentation of nuclei in histopathology images by deep regression of the distance map," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 448–459, Feb. 2019.

[9] H. Qu, Z. Yan, G. M. Riedlinger, S. De, and D. N. Metaxas, "Improving nuclei/gland instance segmentation in histopathology images by full resolution neural network and spatial constrained loss," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2019, pp. 378–386.

[10] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "DCAN: Deep contour-aware networks for accurate gland segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2487–2496.

[11] S. Graham *et al.*, "MILD-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Med. Image Anal.*, vol. 52, pp. 199–211, Feb. 2019.

[12] Y. Xu *et al.*, "Gland instance segmentation using deep multichannel neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2901–2912, Dec. 2017.

[13] Z. Yan, X. Yang, and K.-T. Cheng, "Enabling a single deep learning model for accurate gland instance segmentation: A shape-aware adversarial learning framework," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 2176–2189, Jun. 2020.

[14] Y. Xie, H. Lu, J. Zhang, C. Shen, and Y. Xia, "Deep segmentation-emendation model for gland instance segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2019, pp. 469–477.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[16] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for, volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[17] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[18] J. Jiang *et al.*, "Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 134–144, Jan. 2019.

[19] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing, "Modified U-net (mU-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1316–1325, May 2020.

[20] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-Dense UNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[21] X. Yang *et al.*, "Towards automated semantic segmentation in prenatal volumetric ultrasound," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 180–193, Jan. 2019.

[22] H. Jia *et al.*, "3D APA-net: 3D adversarial pyramid anisotropic convolutional network for prostate segmentation in MR images," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 447–457, Feb. 2020.

[23] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "MS-net: Multi-site network for improving prostate segmentation with heterogeneous MRI data," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2713–2724, Sep. 2020.

[24] J. Zhang, A. Saha, Z. Zhu, and M. A. Mazurowski, "Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 435–447, Feb. 2019.

[25] S. S. Mohseni Salehi, D. Erdogmus, and A. Gholipour, "Auto-context convolutional neural network (Auto-Net) for brain extraction in magnetic resonance imaging," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2319–2330, Nov. 2017.

[26] J. Zhang, Y. Xie, P. Zhang, H. Chen, Y. Xia, and C. Shen, "Light-weight hybrid convolutional network for liver tumor segmentation," in *Proc. Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4271–4277.

[27] Q. Dou *et al.*, "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.*, vol. 41, pp. 40–54, Oct. 2017.

[28] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2494–2505, Jul. 2020.

[29] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution Encoder–Decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2020.

[30] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing failure prediction by learning model confidence," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 2902–2913.

[31] G. Wang *et al.*, "DeepIGeoS: A deep interactive geodesic framework for medical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1559–1572, Jul. 2019.

[32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[34] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[35] S. Wang, L. Yu, X. Yang, C.-W. Fu, and P.-A. Heng, "Patch-based output space adversarial learning for joint optic disc and cup segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 11, pp. 2485–2495, Nov. 2019.

[36] Y. Wang *et al.*, "Deep attentive features for prostate segmentation in 3D transrectal ultrasound," *IEEE Trans. Med. Imag.*, vol. 38, no. 12, pp. 2768–2778, Dec. 2019.

[37] A. Gazdar and A. Maitra, "Adenocarcinomas," in *Encyclopedia of Genetics*, S. Brenner and J. H. Miller, Eds. New York, NY, USA: Academic, 2001, pp. 9–12. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B0122270800015408

[38] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2482–2493, Jul. 2020.

[39] Z. Mirikharaji and G. Hamarneh, "Star shape prior in fully convolutional networks for skin lesion segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2018, pp. 737–745.

[40] B. Lei *et al.*, "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101716.

[41] D. Pati no, J. Avenda no, and J. W. Branch, "Automatic skin lesion segmentation on dermoscopic images by the means of superpixel merging," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2018, pp. 728–736.

[42] F. Peruch, F. Bogo, M. Bonazza, V.-M. Cappelleri, and E. Peserico, "Simpler, faster, more accurate melanocytic lesion segmentation through MEDS," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 2, pp. 557–565, Feb. 2014.

[43] E. Ahn *et al.*, "Saliency-based lesion segmentation via background detection in dermoscopic images," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 6, pp. 1685–1693, Nov. 2017.

[44] B. Antal and A. Hajdu, "An ensemble-based system for microaneurysm detection and diabetic retinopathy grading," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1720–1726, Jun. 2012.

[45] M. Niemeijer *et al.*, "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 185–195, Jan. 2010.

[46] Y. Zhou *et al.*, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2074–2083.

[47] C. Playout, R. Duval, and F. Cheriet, "A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2434–2444, Oct. 2019.

[48] C. Kou, W. Li, W. Liang, Z. Yu, and J. Hao, "Microaneurysms segmentation with a U-net based on recurrent residual convolutional neural network," *J. Med. Imag.*, vol. 6, no. 2, 2019, Art. no. 025008.

[49] M. H. Sarhan, S. Albarqouni, M. Yigitsoy, N. Navab, and A. Eslami, "Multi-scale microaneurysms segmentation using embedding triplet loss," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2019, pp. 174–182.

[50] J. H. Tan *et al.*, "Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network," *Inf. Sci.*, vol. 420, pp. 66–76, Dec. 2017.

[51] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[52] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[53] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 20–36.

[54] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[55] N. C. F. Codella *et al.*, "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 *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

[56] N. Codella *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*. [Online]. Available: http://arxiv.org/abs/1902.03368

[57] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Dec. 2018, Art. no. 180161.

[58] P. Porwal *et al.*, "Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, Jul. 2018.

[59] Y. Yuan and Y.-C. Lo, "Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 519–526, Mar. 2019.

[60] A. Saha, P. Prasad, and A. Thabit, "Leveraging adaptive color augmentation in convolutional neural networks for deep skin lesion segmentation," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 2014–2017.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

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