FEATURE FUSION FOR SEGMENTATION AND CLASSIFICATION OF SKIN LESIONS

Yue Zhang[1] *Zifan Chen*[1] *Hao Yu*[1] *Xinyu Yao*[3] *Hongfeng Li*[⋆,2]

[1] Center for Data Science, Peking University, China
[2] Institute of Medical Technology, Peking University Health Science Center, China
[3] Peking University First Hospital, China

2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761474

ABSTRACT

Automated segmentation and classification of dermoscopy images are two crucial tasks for early detection of skin cancers. Deep models trained for individual task ignore the relationship of the two tasks and lack the diagnostic proposals or explanation for diagnosis results in practice. We assume that features extracted with segmentation models and classification models on the same dataset are highly related and the two tasks have potentials to boost each other when trained together. In this paper, we propose a combined-learning network (CLNet) consisting of a classification network, a segmentation network and a feature fusion module for segmentation and classification of skin lesions. Particularly, the feature fusion module fuses features extracted by the classification branch and segmentation branch and outputs fused features for the two tasks respectively. In this way, the information shared by the two branches can be fully exploited and the performance of two tasks can be mutually improved. Experimental results demonstrate that the proposed model can achieve promising performance on the public skin disease dataset.

***Index Terms*—** Skin Lesion, Image Segmentation, Image Classification, Feature Fusion, Attention

1. INTRODUCTION

Skin cancer is one of the most common diseases worldwide [1]. Fortunately, the survival rate can be increased dramatically if a skin cancer can be diagnosed and treated in its early stage [2]. Dermoscopy images are commonly used tools to improve the diagnostic precision. However, analyzing dermoscopy images is a skillful and time-consuming work. Computer-aided diagnosis (CAD) systems have been designed to assist analyzing dermoscopy images with efficiency and reduced cost. However, diagnosing skin cancer is challenging due to a few reasons: 1) the low contrast between skin lesion areas and normal skin makes it hard to identify the region of lesion area; 2) some artifacts like hair, air bubbles and marks may influence the decision of lesion boundaries and lesion types; 3) appearance of different type of skin lesions may look similar and the appearance of same types sometimes look very differently.

Numerous methods based on convolutional neural networks (CNNs) have been proposed for skin lesion classification and segmentation. For example, Tang et al. [3] proposed a Global-Part Convolutional Neural Network (GP-CNN) model for skin lesion classification, which used the classification activation maps extracted from the Global-CNN to crop image patches to train a subnetwork. Öztürk et al. [4] proposed an improved fully convolutional network (FCN) to segment full-resolution skin lesion images without augmentation. Wu et al. [5] proposed a segmentation network with an adaptive dual attention module to enhance the feature representation in the skip connections of a U-shape FCN. However, these models are designed for classification or segmentation individually which are not feasible in practice since dermatologists not only want to know the diagnosis of a dermoscopy image, but also the region of attention. In light of this, models performing the two tasks simultaneously are studied recently. Jin et al. [6] proposed a cascade knowledge diffusion network to transfer knowledge learned from each another to simultaneously boost the performance of classification and segmentation. Xie et al. [7] proposed a mutual bootstrapping CNN which uses a coarse segmentation model to boost the classification network and combines a fine segmentation network and activation maps of the classification network to boost the performance of segmentation.

Despite the success of the models proposed in [6, 7], the subnetworks of the models are trained separately, i.e., one task needs to wait for another in the forward pass. Thus, this kind of combination is not efficient for network training and evaluation. To resolve this problem, we propose a novel model that combines the segmentation and classification tasks with a feature fusion module and can be trained end-to-end. The main contribution of this paper is twofold. First, we propose a new combined-learning network that can perform skin lesion segmentation and classification simultaneously in an end-to-end manner. Second, a new feature fusion module is built to exploit the shared information in the two tasks by fusing features extracted from the two branches and improve both the segmentation and classification performances.

⋆Corresponding author: Hongfeng Li (lihongfeng@math.pku.edu.cn)

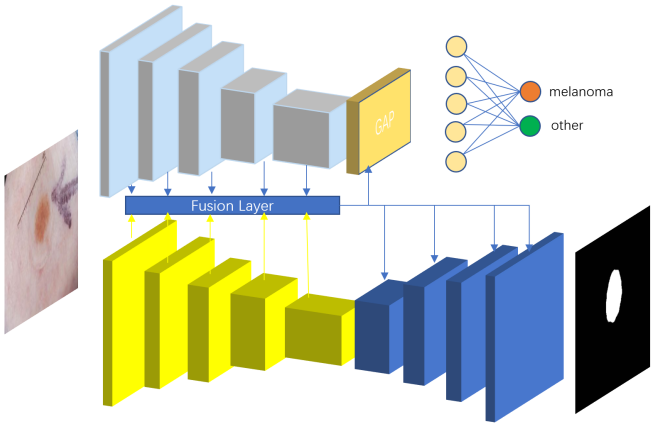

Fig. 1. The framework of the proposed CLNet. It consists of a classification branch, a segmentation branch and a feature fusion module.

2. THE PROPOSED METHOD

Skin lesion classification and segmentation are two highly related tasks. Therefore, we propose a combined-learning network (CLNet) consisting of a classification branch, a segmentation branch and a feature fusion module. Specifically, the feature fusion module fuses the features extracted by the classification branch and segmentation branch and feed them as inputs for the final classification and segmentation. The motivation is to make the two branches learn mutually to make full use of the shared information to further enhance the performance of the two tasks simultaneously. The framework of the proposed model is illustrated in Fig. 1.

We denote that EC_i is the i-th stage features of the classification subnetwork, ES_i is the i-th stage features of the segmentation encoder, De_i is the i-th block of the segmentation decoder, and FM_i is the fusion module with EC_i and ES_i as inputs. Our method can be generally formulated as below.

2.1. Classification Branch

We construct the classification branch based on the Inception V3 [8] or ResNet50 [9]. The number of nodes in the output of the fully-connected (FC) layer in the network is modified to the number of image classes. Each value in the output represents a score corresponding to the probability that a dermoscopy image belongs to a specific type of skin lesion. The input of the FC layer is the feature output by the feature fusion module. The classification branch can be formulated as:

$$\begin{aligned} &GAP(FM_1, FM_2, ..., FM_d)W^c \\ &FM_i = FM(EC_i, ES_i) \end{aligned} \tag{1}$$

where GAP is the global average pooling performed on i-th fusion module and concatenate the outputs of all fusion layer (noted as FM_i).

Class activation maps (CAMs) can coarsely visualize the concerned parts of an input image, providing some qualitative explanation for what has been learned by a network. In this paper, We use the GradCAM++ [10] to demonstrate the areas that our proposed model focuses on. We define that y_p is the prediction of a class label p and X_k is the k-th feature map before the Global Average Pooling (GAP) layer. The weight of the k-th channel of X is defined as:

$$w_k = \frac{1}{WH} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y_p}{\partial X_{i,j}^k} \tag{2}$$

where W and H is the size of feature map X^k. Then the activation map can be expressed as:

$$AM_p = \sum_k ReLU(w_k X^k) \tag{3}$$

The activation maps are generated from the convolutional features of the layer before the GAP layer. We know that the activation maps are related to the classification label, focusing on the most relative regions that can be used as clues to classify an input image. We use the ground truth label as the target label to generate activation maps.

2.2. Segmentation Branch

UNet [11] is an effective segmentation network widely used for medical image segmentation. The architecture of UNet can be divided into an encoder and a decoder. The encoder can be a common CNN that aims to extract semantic features. The decoder is to recover the resolution downsampled features and extract the border information of interested areas. Additionally, skip connections between the encoder and decoder benefit feature reuse and recover refined border of the predicted mask. Based on the original UNet structure, we replace the encoder with a ResNet50 which is more powerful for feature extraction. The decoder and skip connections keep the same as the original UNet, and we perform bilinear interpolation to the output of the decoder to make the predicted mask have the same size as the input image. We call this modified UNet as ResUNet.

To enable the UNet to use the fused features (FM_i), the skip connections are replaced with the output of the feature fusion module. The i-th decoder block (De_i) takes the upsampled features of De_{i-1} and FM_i as input and obtains output features that will be used in the next decoder block. The process is formulated as:

$$De_i = De(Concat(Up(De_{i-1}), FM_i)) \tag{4}$$

where Up is the bilinear upsample operation, and $Concat$ means concatenating features from the channel dimension.

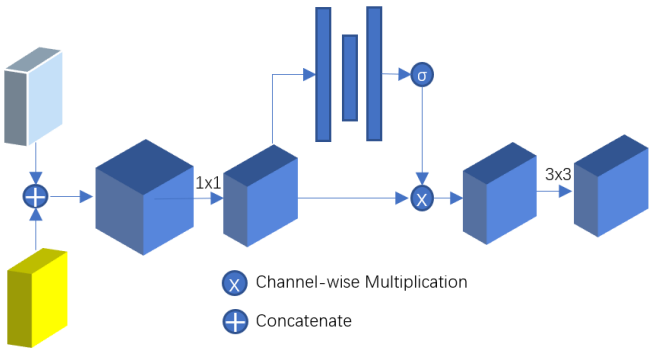

Fig. 2. The architecture of feature fusion module. Features extracted from the segmentation branch (yellow) and the classification branch (blue) are concatenated together and went through a SE attention block to select the most relevant features from the channel dimension.

2.3. Feature Fusion Module

Commonly, classification network and segmentation network are trained separately. To make full use of the information shared by the two networks, we propose a feature fusion module that fuses the features of classification branch and segmentation branch. Then the fused features are passed to the segmentation decoder and the classification head for segmentation and classification respectively. The feature fusion module is shown in Fig. 2.

As shown in Fig. 2, the features extracted by the classification and segmentation branches, represented as EC_i and ES_i, are taken as input of the feature fusion module. Note that EC_i and ES_i may have different spatial sizes. Thus, the features with smaller spatial size will be upsampled to match another before the concatenation operation. Then the two kind of features are concatenated and went through a convolution operation with kernel size of 1×1 to reduce the dimension of channels. Thereafter, the output features are fed to a SE attention block [12] to select the most relevant features from the channel dimension. Afterwards, convolutions with kernels of size 3×3 are performed.

3. EXPERIMENTS AND ANALYSIS

In this section, we evaluate the proposed model on the ISIC2017 dataset [13] which contains $2,000$ images for training, 150 images for validation and 600 images for testing. We compare the segmentation and classification performance with models that perform segmentation or classification individually to demonstrate the effectiveness of our model.

3.1. Experimental Settings

We implement our model with PyTorch [14]. The classification branch is initialized with weights pretrained on the ImageNet dataset. We use Adam as the optimizer. The initial learning rate lr_0 is 1e-4 and descends in a polynomial style with $lr = lr_0(1 - \frac{steps}{total\_steps})^{\alpha}$, where $total\_steps$ is the total iteration step, $steps$ is the step number of each iteration and we set $\alpha = 0.96$ in the experiments. The weight decay is 5e-4. The models are trained with 200 epochs in total with a single GTX 1080ti GPU. In addition, the size of the input images is 224×224 and the data augmentation of random up-down or left-right flipping is performed. The focal loss [15] is utilized as loss function for the classification task, while the addition of focal loss and dice loss [16] is utilized as loss function for the segmentation task.

3.2. Performance of the Classification Task

We train the ResNet50, Inception V3, ARLNet50 [17] on the ISIC2017 dataset for classification. In order to verify that the segmentation branch in our proposed model can boost the performance of the classification branch, we add a UNet branch to ResNet50 and Inception V3, respectively. As shown in Table 1, we can observe that the proposed method with Inception V3 and UNet outperforms Inception V3 by more than 1 point percentage in terms of AUC score, and the proposed method with ResNet50 and UNet outperforms ResNet50 by more than 3 points percentage in terms of AUC score, in both melanoma and seborrheic keratosis classification tasks. The results indicate that fusing features learned from the segmentation branch can enhance the feature representation of classification task and lead to better performance for the classification task.

Additionally, we calculate CAMs during the evaluation stage to visualize the learned representations of different classification models. We choose the convolution features to calculate CAMs for the ground truth labels. As shown in Fig. 3, we can find that the CAMs of the proposed models (third and last columns) are generally more accurate than those of the compared models (ResNet50 and Inception V3). It should be noted that the activation maps are more concentrated with the help of segmentation branch. However, wrongly located areas (especially third and forth row of the third column) may be presented by wrong classification results. Thus, we can conclude that the proposed feature fusion module can assist the learning process of classification branch.

3.3. Performance of the Segmentation Task

Similarly, we train a UNet [11] and a ResUNet (UNet with ResNet50 as the encoder) on the ISIC2017 dataset for segmentation. For comparison, we employ a ResNet classification branch to assist the segmentation task. The segmentation performance is reported in Table 2. We can see that the segmentation performance of the proposed ResUNet outperforms that of UNet by more than 1% in terms of Jaccard index (JA).

In addition, we show several examples of the predicted masks in Fig. 4. From the figure we can observe that the

Method	Mel Classification				SK Classification				AUC (avg)
	AUC	ACC	SEN	SPC	AUC	ACC	SEN	SPC	
ResNet50	81.34	81.33	52.99	88.2	90.15	88.67	53.33	94.90	85.745
Inception V3	82.96	82.17	64.10	86.54	90.99	87.83	75.56	90.00	86.975
ARLNet50	80.56	81.50	52.99	88.41	91.81	84.50	84.51	84.44	86.185
Inception V3 + UNet (Ours)	84.85	75.83	79.49	74.95	92.17	85.50	86.67	85.29	88.510
ResNet50 + UNet (Ours)	84.91	82.33	59.83	87.78	94.51	86.67	90.00	86.08	89.710

Table 1. Classification performance comparison. Mel classification is the task of differentiating melanoma with other skin lesions. SK classification is the task of differentiating seborrheic keratosis with other skin lesions. AUC, ACC, SEN, SPC are short for area under curve, accuracy, sensitivity, specificity, respectively.

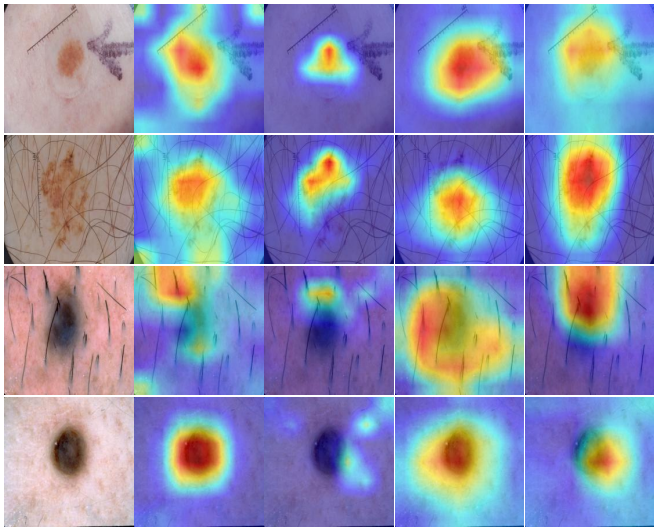

Fig. 3. CAMs for 4 randomly selected dermoscopy images. The first column contains original images, and the second column to the last column indicate the activation maps of ResNet50, ResNet50+UNet, Inception V3, Inception V3+UNet, respectively.

Method	JA	Dice
UNet [11]	71.63	83.47
ResUNet	74.13	85.14
ResNet50 + UNet (Ours)	73.25*	84.56*
ResNet50 + ResUnet (Ours)	74.54*	85.41*

Table 2. Segmentation performance comparison. '*' means that the value is averaged by the segmentation performance with classification label of two different tasks.

segmentation masks of our model are more accurate than its counterparts.

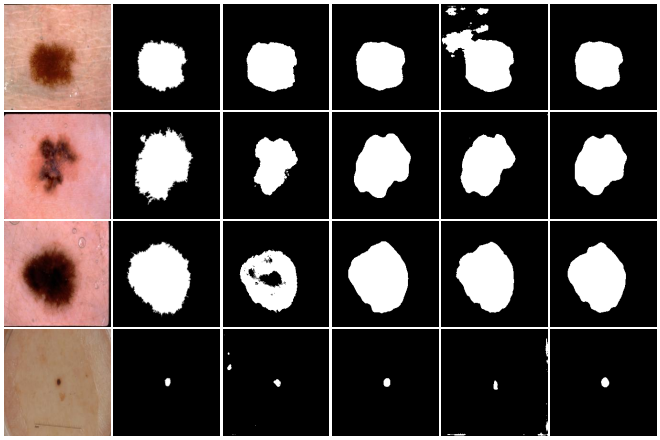

Fig. 4. Several examples of segmentation results. From left to right are original images, ground truth and predicted masks of UNet, ResNet50+UNet, ResUNet, ResNet50+ResUNet.

4. CONCLUSION

In this paper, we propose a new model called CLNet that can perform skin lesion classification and segmentation at same time and can be trained in an end-to-end manner. Particularly, a feature fusion module is constructed in our model to fuse features extracted by the classification branch and segmentation branch. Then the fused features are fed back to the two branches for final segmentation and classification. In this way, the two tasks can be mutually boosted. To verify the effectiveness of our model, we compare it with individual classification or segmentation models on a public dataset. Experimental results demonstrate the superiority of our model with different evaluation metrics.

Acknowledgement. This work was supported by the National Key Research and Development Program of China under Grant 2019YFC0840706 and 2019YFC0840700, and the National Natural Science Foundation of China 11701018 and 12090022.

5. REFERENCES

[1] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal, "Cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[2] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[3] Peng Tang, Qiaokang Liang, Xintong Yan, Shao Xiang, and Dan Zhang, "Gp-cnn-dtel: Global-part cnn model with data-transformed ensemble learning for skin lesion classification," *IEEE journal of biomedical and health informatics*, vol. 24, no. 10, pp. 2870–2882, 2020.

[4] Şaban Öztürk and Umut Özkaya, "Skin lesion segmentation with improved convolutional neural network," *Journal of digital imaging*, vol. 33, no. 4, pp. 958–970, 2020.

[5] Huisi Wu, Junquan Pan, Zhuoying Li, Zhenkun Wen, and Jing Qin, "Automated skin lesion segmentation via an adaptive dual attention module," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 357–370, 2020.

[6] Qiangguo Jin, Hui Cui, Changming Sun, Zhaopeng Meng, and Ran Su, "Cascade knowledge diffusion network for skin lesion diagnosis and segmentation," *Applied Soft Computing*, vol. 99, pp. 106881, 2021.

[7] Yutong Xie, Jianpeng Zhang, Yong Xia, and Chunhua Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2482–2493, 2020.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[12] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[13] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, 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 *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[17] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen, "Attention residual learning for skin lesion classification," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2092–2103, 2019.