Skin Lesion Segmentation and Disease Classification: A Study with ISIC Challenge 2018

Randy Ardywibowo[1,*], Wuyang Chen[1,*], Kai He[1,*], Xueting Liu[1,*], Xiaoqian Jia[1,*], Qing Jin[1,*], Shuai Huang[2], Xiaoning Qian[1], Zhangyang Wang[1]

***Abstract*—Skin cancer is one of the most prevalent forms of cancer in the United States. Despite its significance, detecting skin cancer remains to be a challenging task even for domain experts using techniques such as dermoscopy. Recently, the International Skin Imaging Collaboration (ISIC) has aggregated a large amount of dermoscopic images of diseased skin, and has annually held challenges asking participants to automatically classify and segment these images. Current state-of-the-art methods to accomplish this use deep neural networks, with significant performance improvements compared to traditional image processing methods. We describe our application of these deep learning techniques on the ISIC Challenge of 2018. Specifically, we focus on Task 1: Skin Lesion Segmentation, and Task 3: Skin Disease Classification of the challenge. To accomplish Task 1, We develop a two-stage coarse-to-fine pipeline based on the PSPNet backbone. To accomplish Task 3, we implement a cascading classifier using the PNASNet backbone, as well as using various data augmentation and class re-balancing pre-processing techniques. We report the various experiments and ablation studies that led us to our final method pipelines for both tasks.**

I. Introduction

MORE than 5 million people in the United States are diagnosed with skin cancer every year [1]. As many of these diseases occur on the skin surface, visual examination techniques are appealing for their detection and diagnosis. Recent techniques such as dermoscopy magnify the skin and eliminate its surface reflection, allowing for visual inspection of skin disease images with improved diagnostic accuracy [2]. Despite so, skin disease classification remains a difficult task even for domain experts, not to mention automated algorithms.

The International Skin Imaging Collaboration (ISIC) has aggregated a large amount of publicly accessible dermoscopy images. Currently, the ISIC archive contains over 13,000 dermoscopy images collected from many different clinical centers, ensuring a large and representative sample set [3], [4]. Moreover, ISIC has held competitions to perform segmentation and classification on these datasets annually from 2016. For the current year, the 2018 ISIC challenge consists of 3 tasks: Task 1 to segment lesion regions; Task 2 to segment different attributes of lesion regions; and Task 3 to classify skin images from different disease classes.

Recently, deep neural networks have been shown to accomplish these tasks with significant performance improvements over traditional image processing methods. For example, recent deep architectures such as UNet [5], SegNet [6], ReSeg [7], MaskRCNN [8], and PSPNet [9] have achieved considerable performance progress in image segmentation. Architectures such as ResNet [10], GoogleNet [11], InceptionV3 [12], VGG [13], and NASNet [14], have achieved exceptional performance in image classification. Deep models can also be enhanced with particular robustness to real-world image degradations such as low resolution and noise [15], [16], [17].

We participated in Tasks 1 and 3 of the ISIC 2018 challenge to segment lesion regions and classify skin images with given disease labels, respectively. We adopt the latest promising deep learning-based segmentation or classification models, assisted by careful domain-specific learning and tuning procedures. For Task 1, we develop a two-stage coarse-to-fine pipeine: the first stage identifies a bounding box that encompasses the lesion region; the second stage then focuses on this candidate wound lesion for refined segmentation. For Task 3, we utilize a tree-structured cascade of deep classifiers, that first separate the difficult classes (MEL, NV, and BKL) from other easier ones, and then continue with more specialized classifiers for either group. We detail the experiments and observations below.

II. Task 1: Skin Lesion Segmentation

A. Backbone Architecture

To accomplish the task of segmenting skin disease lesion, we firstly benchmarked multiple existing deep architectures that have shown state-of-the-art performance in image segmentation: UNet [5], SegNet [6], ReSeg [7], MaskRCNN [8], and PSPNet [9]. We performed extensive experiments and identified **PSPNet** to be the backbone for our Task 1 pipeline.

B. Dividing Patches v.s. Resizing Images: Which is Better?

Due to the large resolution of ISIC images, we consider two practical strategies to fit the training process into GPU memory: i) dividing an image into smaller patches and performing segmentation on each patch, then aggregating them back; ii) resizing an image into a lower resolution and feeding the downscaled image for segmentation. While our initial hypothesis was in more favor of the former since it could better preserve fine-scale features better that are critical for pixel-level segmentation, our experiments surprisingly suggested differently: As the example shown in Figure 1, although we kept different patches to overlap, the aggregated entire image showed strong inconsistency between local regions, and the whole segmentation result was absurd. We conjecture that the global foreground-background context could be critical to

[1] Texas A&M University, College Station TX 77843, USA
[2] University of Washington, Seattle WA 98195, USA
[*] denotes equal contribution.

lesion segmentation, which was apparently destroyed when segmented into patches.

Since we need to both preserve high-resolution features, which are more conserved by using local patches, and global foreground-background context, which are only available from an entire image, we are motivated to combine the bests of both worlds using a two-stage pipeline: first localizing the foreground lesion from the entire image, **at a lower resolution** (resized); and then performing fine segmentation on this zoomed-in local patch only, **at the original resolution**.

(a) Skin lesion image (b) True segmentation (c) Prediction by aggregating patches

Fig. 1: A failure example of patch-based segmentation. Here we chunk the origin image (a) into 300 × 300 patches, making pixel-level predictions at patches, and assembling patches back into (c), which failed to capture the global context.

C. Two-Stage Pipeline: Localization then Segmentation

Our two-stage pipeline is shown in Figure 2. The first stage uses a PSPNet to identify a local patch to be focused on, that likely contains the target lesion. It starts by resizing the original image into a lower resolution (300 × 300), from which it is only responsible for predicting a coarse segmentation map; we then take a bounding box that tightly encompasses this coarsely-segmented region. Next, we "zoom in" this bounding box region, by cropping out the corresponding candidate region at the original-resolution image. We then use a second PSPNet to perform fine-scale pixel level segmentation for this local patch only. The two PSPNets are trained jointly as an end-to-end pipeline.

D. Results

The Jaccard index has been proposed as a reliable measure of segmentation accuracy compared to 0-1 accuracy. This index measures the intersection over union (IoU) of the predicted segmentation region with the ground truth mask. Specifically, the Jaccard Index is defined as follows (TP = True Positive; FN = False Negative; FP = False Positive):

$$\text{Jaccard Index} = \text{IoU} = \frac{TP}{TP + FN + FP} \tag{1}$$

The official challenge ranking metric takes a threshold of 0.65 to remove segmentation results that are insignificantly related to the ground truth:

$$\text{Score} = \begin{cases} IoU, & \text{if } IoU \geq 0.65 \\ 0, & \text{if } IoU < 0.65 \end{cases} \tag{2}$$

Table I summarized our Task 1 segmentation results. The proposed two-stage PSPNet architecture achieves more compelling results than default PSPNet and several other state-of-the-art models. Figure 3 compares the segmentation result of a test images by different models visually.

TABLE I: Segmentation Accuracy Results (Task 1)

Model	Jaccard index	Score
SegNet	0.76	0.68
ReSeg	0.78	0.72
PSPNet	0.78	0.72
Two-Stage (proposed)	**0.80**	**0.75**

III. Task 3: Skin Disease Classification

A. Backbone Architecture

In Task 3, each image is assigned with one disease label. Like Task 1, we first seek an appropriate backbone architecture by experimenting with different state-of-the-art image classification models, including ResNet [10], InceptionV3 [12], GoogleNet [18], VGG [11], InceptionResnetV2 [13], and NASNet [14]. We used their ImageNet pre-trained versions as our initialization, and further trained them on ISIC data. Among them, NASNet achieves the highest accuracy based on hold-out evaluation, thanks to its optimized design from the architecture search. With this finding, we further investigate PNASNet, which utilized the structured search space proposed in NASNet [14] and followed a simple-to-complex progressively growing procedure. After several more comparison experiments, we identify **PNASNet5large** as our Task 3 backbone.

B. Preprocessing: Color Variations and Class Imbalance

Since the varying colors of skins are found to hurt the performance, we refer to color consistency transformation to normalize the skin regions to a consistent color tone [19], [20]. Figure 4 shows the effect of such color adjustment.

The class imbalance is another factor found to degrade classification. For example, many images are prone to misclassified as Nevus (NV) since it is the dominating class. To handle that issue, we perform several data augmentations, including flipping and brightness adjustment, to increase the volume of minority classes (BCC, AKIEC, DF, and VASC). Meanwhile, we subsample the MEL, NV, and BKL classes. Our experiments show that although the overall accuracy might be slightly negatively impacted in this way, the class-wise precision, recall, and confusion matrix all get improved.

C. From Plain to Cascaded Classifiers

Initially, we try to train a 7-class classification model directly. However, the preliminary results are very unsatisfactory. We then notice the following dataset characteristic: there are three highly confusing classes: Melanoma (MEL), Nevus (NV), and Benign keratosis (BKL) that account for a large portion of classification failures. We are hence motivated to build a cascaded classifier, and to create specialized classifiers for those confusing classes.

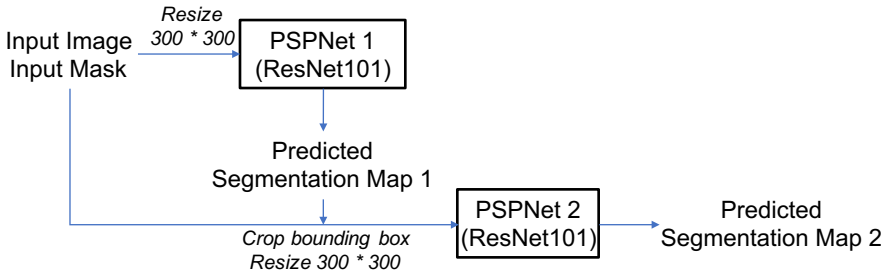
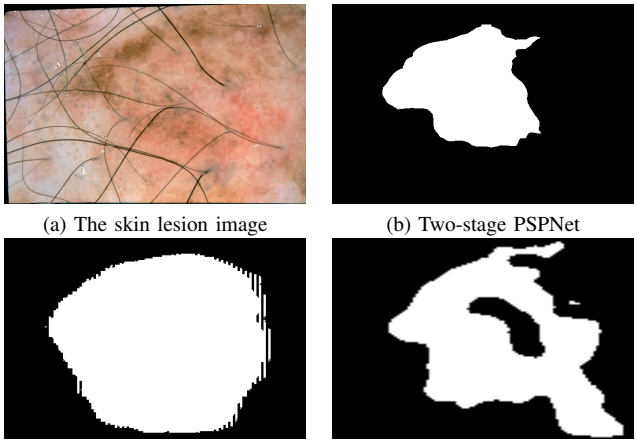


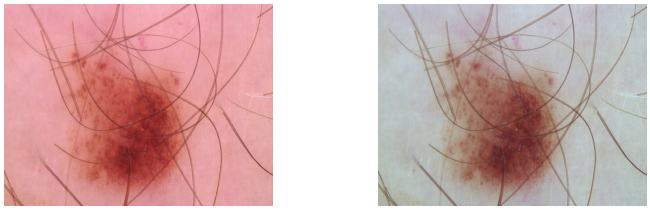
Fig. 2: The two-stage pipeline using PSPNets for Task 1.



(a) The skin lesion image (b) Two-stage PSPNet

(c) ReSeg (d) SegNet

Fig. 3: An example of segmentation results from the test set.



(a) The unprocessed image (b) After color consistency transformation

Fig. 4: Color consistency transformation

Fig. 5 demonstrates our cascaded classifier pipeline, in which the "Classifier 2" is specifically aimed at classifying the three difficult classes (MEL, NV and BKL). Each of the three classifiers adopts the PNASNet backbone. Likewise, the pipeline is trained from end to end to boost the final classification performance.

D. Results

To showcase the validity of our designed hierarchical structure, we compare the proposed methods with: i) a vanilla 7-class PNSANet classier; and ii) an alternative cascaded classifier (Alter-cascade) pipeline with the identical structure to the proposed, except that we only group MEL and NV together (rather than MEL/NV/BKL) for the Classifier 2 to proceed on. We also compare the effects of the pre-processing steps of color constancy transformation ("Color Norm") and data augmentation/subsampling ("Balance"), respectively.

Since it is an imbalanced classification problem, we use the balanced accuracy (mean recall) as the performance indicator. Table II demonstrates various results measured on the ISIC validation set: equipped with the two pre-processing techniques, the proposed cascaded pipeline achieves the highest balanced accuracy of 0.843 among all.

IV. Discussion and Conclusions

In this work, we designed effective successful pipelines for ISIC 2018 Challenge Tasks 1 and 3. Our experimental results from this challenge suggest a number of important findings: 1) For lesion segmentation, the global contextual information of foreground and background is as important as the high-resolution local detail. A coarse-to-fine pipeline seems to be a natural solution; 2) Color normalization of skin images alleviates the sensitivity to different subject skin colors, and thus has positive impacts for classification; 3) Different skin diseases have very different sample volumes; moreover, the difficulty levels to distinguish between them can vary drastically too. A structured classifier cascade, plus proper class re-sampling as pre-processing, can help resolve this challenge. Our future work will include more domain-specific data augmentation techniques, as well as explicitly incorporate robustness to the noisy segmentation labels.

Acknowledgements

The authors would like to thank Texas A&M High Performance Research Computing (HPRC) for providing computational resources. Among many colleagues at Texas A&M University that provide their valuable help, we specifically thank Weizhi Li for his help in benchmarking some of the deep models.

References

[1] H. W. Rogers, M. A. Weinstock, S. R. Feldman, and B. M. Coldiron, "Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012," *JAMA dermatology*, vol. 151, no. 10, pp. 1081–1086, 2015.

[2] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *The lancet oncology*, vol. 3, no. 3, pp. 159–165, 2002.

[3] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. 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 *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. IEEE, 2018, pp. 168–172.

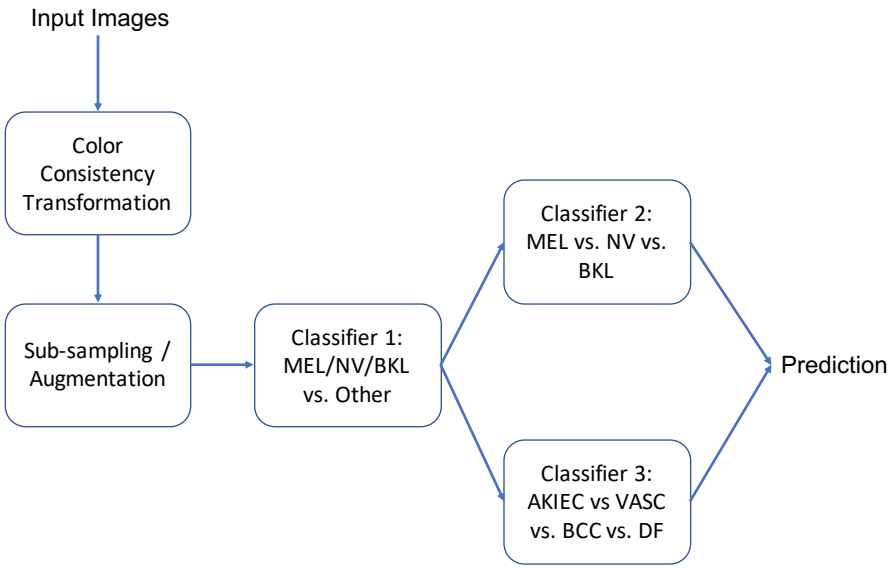


Fig. 5: The proposed cascaded classifier pipeline for Task 3.

TABLE II: Balanced Accuracy Results (Task 3)

Data Processing \ Architecture	Vanilla	Alter-cascade	Proposed
Original	0.833	0.824	0.833
Original + Color Norm	0.811	0.812	0.843
Original + Color Norm + Balance	0.764	0.824	**0.848**

[4] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *arXiv preprint arXiv:1803.10417*, 2018.

[5] O. Ronneberger, P. Fischer, and T. 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.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[7] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville, "Reseg: A recurrent neural network-based model for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 41–48.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." 2017.

[14] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," *arXiv preprint arXiv:1707.07012*, vol. 2, no. 6, 2017.

[15] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4792–4800.

[16] D. Liu, B. Wen, X. Liu, Z. Wang, and T. S. Huang, "When image denoising meets high-level vision tasks: A deep learning approach," *arXiv preprint arXiv:1706.04284*, 2017.

[17] D. Liu, B. Cheng, Z. Wang, H. Zhang, and T. S. Huang, "Enhance visual recognition under adverse conditions via deep networks," *arXiv preprint arXiv:1712.07732*, 2017.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[19] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE journal of biomedical and health informatics*, vol. 19, no. 3, pp. 1146–1152, 2015.

[20] Z. Wang, J. Yang, H. Jin, E. Shechtman, A. Agarwala, J. Brandt, and T. S. Huang, "Deepfont: Identify your font from an image," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 451–459.