G2LL: Global-To-Local Self-Supervised Learning for Label-Efficient Transformer-Based Skin Lesion Segmentation in Dermoscopy Images

Fei Chen, Jiacheng Wang, Baptiste Magnier, Wei Xue, Shaohui Huang, Liansheng Wang

G2LL: GLOBAL-TO-LOCAL SELF-SUPERVISED LEARNING FOR LABEL-EFFICIENT TRANSFORMER-BASED SKIN LESION SEGMENTATION IN DERMOSCOPY IMAGES

Fei Chen[1], Jiacheng Wang[1], Baptiste Magnier[2], Wei Xue[3], Shaohui Huang[1,], Liansheng Wang[1,*]*

[1]Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China
[2]Euromov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France
[3]Open Platform Department, Internet Service Provider, Oppo

ABSTRACT

Skin lesion segmentation in dermoscopy images is highly relevant for lesion assessment and subsequent analysis. Recently, automatic transformer-based skin lesion segmentation models have achieved high segmentation accuracy owing to their long-range modeling capability. However, limited labeled data for training the lesion segmentation models results in sub-optimal learning results. In this paper, we propose a Global-to-Local self-supervised Learning (G2LL) method for transformer-based skin lesion segmentation models to alleviate the problem of insufficient annotated data. Firstly, a structure-wise masking strategy for Masked Image Modeling (MIM) is proposed to force the model to learn the reconstruction of masked structures by exploring the semantic local contexts. Instead of masking patches randomly in the whole view, it computes superpixels to divide the images into several structured regions. Then, it masks the fixed number of patches in each region, thus it allows the exploration of the structural knowledge and solves the shape variance in the meanwhile. Secondly, a self-distilling architecture is deployed to enhance global context learning where the masked images are sent to a student network and the relative unmasked images are fed to a teacher network for knowledge distillation. In this context, extensive experiments on both the ISIC-2017 and the ISIC-2019 datasets containing a total of 28 081 images show that the proposed approach is superior to state-of-the-art self-supervised learning methods.

Index Terms— skin lesion segmentation, self-supervised learning, structure-wise masking

1. INTRODUCTION

Accurate segmentation of skin lesions for dermoscopy images is significant for the diagnosis and treatment planning of melanoma, which is the most fatal skin disease. Reliable automatic segmentation algorithms are expected to surpass human experts in segmentation accuracy and computational efficiency, but they depend on a large amount of accurate pixel annotation for lesions. Additionally, low contrast between healthy and lesion areas (due to illumination or sensor problems/calibration) makes it difficult to determine the lesion boundaries. Also, hairs in dermoscopy images may destroy the lesion appearance, falsifying the segmentation.

In order to address the challenge of lesion segmentation in dermoscopic images, several approaches have been proposed. Hand-crafted features are applied in early work, producing weak and unstable performance in skin lesion segmentation [1]. With the development of Convolutional Neural Networks (CNNs), U-Net [2] and Dilated Convolution [3] exhibit excellent segmentation performance in medical images. The subsequent CNN-based methods utilize the multi-scale features enhancement, the receptive field expansion, and attention mechanisms to enhance the segmentation [4, 5, 6, 7]. Nevertheless, the receptive field of CNN-based networks is limited [8]. Inspired by the success of Vision Transformers (ViT) in the natural image domain, several studies apply transformers to skin lesion segmentation and have obtained better results than CNNs [8, 9, 10].

Unfortunately, the lack of extensive skin lesion annotations is a major obstacle to building accurate and robust transformer-based networks. To address it, Self-Supervised Learning (SSL) in medical image analysis shows promising performance based on different designs of pretext tasks, which can be categorized into two approaches. On the one hand, the discriminative method [11] applies contrastive learning to learn image-level representation similarity and instance discrimination, which have been proven effective on classification tasks yet less valid for segmentation tasks that require fine-grained features. On the other hand, the generative method [12] uses an auto-encoder to recover the original image from the distorted counterpart, showing a good capacity for capturing local context. Nevertheless, generating the whole image is computationally expensive and another group based on the Masked Image Modeling (MIM) pretext task has been proven successful [13, 14]. They mask a proportion of image patches and predict the masked parts from unmasked patches, saving a part of the computational cost.

In this paper, we propose a Global-to-Local self-supervised

*Correspondence: {hsh, lswang}@xmu.edu.cn

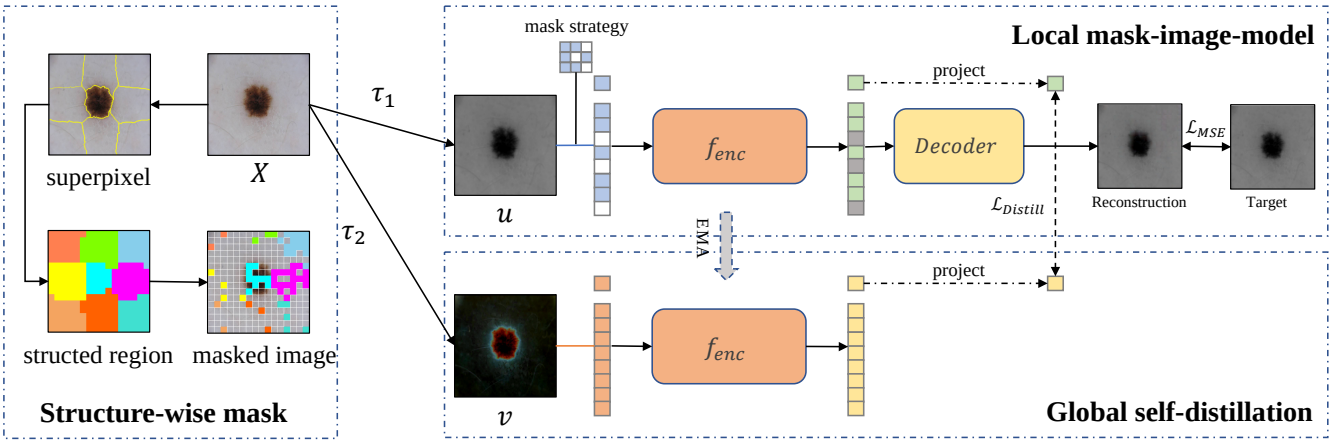

Fig. 1. An overview of G2LL for pre-training of transformer-based skin lesion segmentation model. The top part is a masked image patches recovery pretext task for local learning. The bottom part is knowledge self-distillation for global learning. The left part shows the process of structure-wise mask generation.

Learning (G2LL) method to learn universal feature representation for the transformer-based network from an unlabeled skin lesion dataset. Specially, we follow the masked image reconstruction pretext task to learn interaction in the local area [14]. To improve the learned representations that can better suit the subsequent skin lesion segmentation tasks, we propose to randomly mask the image patches in the structured region segmented by the superpixel algorithm [15] rather than the whole image. Furthermore, to enhance the learned global contexts that can be used to solve the large lesion shape and size variance at the finetuning time, we propose to utilize a teacher-student network architecture between the feature extraction of masked and unmasked images with self-distillation. The teacher network is fed with the unmasked images and the produced unmasked features are used to supervise the feature extraction of the student networks with the masked images. Extensive experiments on the ISIC-2017 dataset and evaluations by means of well-known Dice and IoU (Intersection over Union) metrics show that our proposed approach improves the segmentation performance and is superior to competing SSL methods.

2. PROPOSED METHOD

The overview learning framework of G2LL has been visualized in Fig. 1; it contains two ViT backbones, named student model and teacher model, to learn the local contexts and global features. According to the structural information computed by superpixels [15], we mask highly likely diseased image patches instead of randomly selected patches in the images (see Sec. 2.2). The student model takes unmasked image patches as input to learn local context features with a pixel-level loss for skin lesions. The teacher model is used to extract global features from the unmasked image to distill knowledge to the student model.

2.1. Basic MIM Pre-training Model

Our method is based on the classical MIM approach, MAE (Masked AutoEncoder [14]), which adds a shallow decoder after the ViT encoder to reconstruct pixels and learn local features. Hence, this subsection gives a brief description of this method and our advanced method will be introduced in the next parts.

In MAE manner, each dermoscopy image X is divided into n patches $\{X_i\}_{i=1}^{n}$ where each patch contains $16\times16\times3$ pixels. A random masking strategy is applied to mask a portion of image patches X_{mask} and the remaining patches $X_{visible}$ are fed into ViT backbone f_{enc} to encode local features $F_{visible}$.After that, a shallower decoder f_{dec}, consisting of a 2-layers transformer block with 384 hidden dimensions and 6 heads in each layer, generates $\hat{X}_{mask}$ from local features $F_{visible}$ and the learnable vector T_{mask} representing the masked token. The MSE (Mean Square Error) loss between $\hat{X}_{mask}$ and X_{mask} is utilized to optimize the reconstruction learning phase.

The mathematical description of MIM can be written as:

$$F_{visible} = f_{enc}\left(X_{visible}\right), \tag{1}$$

$$\hat{X}_{mask} = f_{dec}\left(F_{visible}, T_{mask}\right), \tag{2}$$

$$\mathcal{L}_{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{X}_{mask} - X_{mask}\right)^2, \tag{3}$$

where m is the number of masked patches.

2.2. Structure-wise Mask Strategy

MAE [14] and SimMIM (Simple framework for Masked Image Modeling [16]) apply a high mask ratio γ for randomly masking image patches, to learn meaningful feature representation. Unlike natural images, The scale of skin lesions in der-

moscopy images varies dramatically. Besides, dermoscopy images hold lower context complexity compared to natural images, so a high value of γ for dermoscopy images is very hard to learn useful representation and finally leads to unstable feature learning. However, a low mask ratio γ for MIM is too easy to train a model, so masked patches can easily be recovered by their neighbors with little meaningful feature learning.

To this end, we introduce a compromised solution for MIM, i.e., to mask more patches under the key structured regions where lesions may be present with a low γ value. Specifically, an image is divided into patches as before; then, we produce superpixels by SLIC method [15] as a meaningful label for segmenting image patches into K (empirically set to 8 as default) structured areas. Each patch is assigned the superpixel label that the most pixels hold within its area. Thus, we can mask the patches under the key structured regions. For dermoscopy images, the skin lesion usually locates in the center of the image. So, we focally mask $n \times \gamma \times \beta$ patches in the central regions, while the patches in marginal areas are less masked.

The structure-wise mask is used to improve the structure learning of MAE. The $X_{visible}$ in Eq. (1) is modified as:

$$X_{visible} = X \otimes M, \tag{4}$$

where X is the augmentation view of input dermoscopy images and $M \in \mathbb{R}^{H \times W}$ is the structure-wise masks. Here, H and W represent the height and width of the image respectively and $\otimes$ refers to element-wise multiplication.

2.3. Knowledge self-distillation Enhance Global Feature

We utilize knowledge self-distillation to enhance global feature learning. It uses siamese encoders to calculate feature representations from different augmented views u and v of the same image. The student model is MAE and the teacher model is ViT architecture which is the same as the one used in the student model. The mask strategy illustration aforementioned is applied on both u and v before sending it to the student model. We fed the non-masked view of u and v to the teacher model for knowledge self-distillation. In detail, we follow DINO (self-DIstillation with NO labels [17]) to extract the feature of the class token as the global feature with ViT, then project it with 3-layer Multi-Layer Perception (MLP) projector as $P_{\theta^s}(\cdot)$ and $P_{\theta^t}(\cdot)$. The objective Loss function for knowledge self-distillation is formulated as below:

$$\mathcal{L}_{Distill} = -\sum_{x \in \{u,v\}} \mathcal{F}_t(x, \tau_t) \cdot log(\mathcal{F}_s(x', \tau_s)), \tag{5}$$

with

$$\mathcal{F}_t(x, \tau_t) = \text{softmax}\left(\frac{P_{\theta^t}\left(h_t^{[\text{CLS}]}(x)\right) - \mathcal{C}}{\tau_t}\right), \tag{6}$$

$$\mathcal{F}_s(x', \tau_s) = \text{softmax}\left(\frac{P_{\theta^s}\left(h_s^{[\text{CLS}]}(x')\right)}{\tau_s}\right), \tag{7}$$

where x' is the masked view of x and $h_t^{[\text{CLS}]}(\cdot)$ refers to the encoded class token of x in the teacher model. $\mathcal{C}$ is the center of teacher outputs. τ_s and τ_t are temperature parameters; the setting of these hyper-parameters follow [17]. Finally, the loss for G2LL is $\mathcal{L}_{ssl} = \lambda_1 \cdot \mathcal{L}_{MSE} + \lambda_2 \cdot \mathcal{L}_{Distill}$, where λ_1 and λ_2 are the hyper-parameters to control local and global feature learning (we set $\lambda_1 = 1$ and $\lambda_2 = 1$ for better convergence).

3. EXPERIMENTS

3.1. Dataset

Two publicly available datasets, ISIC-2019 [18] and ISIC-2017 [19], are used to evaluate our method. ISIC-2019 comprises 25 331 dermoscopy images with 8 different lesion categories. ISIC-2017 contains a total of 2 750 dermoscopy images and corresponding pixel-level annotations for segmentation. It is officially divided into a train set (2 000 images), a validation set (150 images), and a test set (600 images). All the images contained in ISIC-2017 and ISIC-2019 datasets are in RGB space (Red, Green and Blue) and PNG format.

3.2. Implementation

In the pre-training stage, the images from ISIC-2019 are empirically resized to 256×256. A group of data augmentations is used to generate different input views, including color jitter, random grayscaling and Gaussian blur. We employ the small variant of ViT (ViT-S/16) with patch size 16×16 as ViT backbone. The model is optimized by an AdamW optimizer [20] with an initial learning rate of 0.00015. We train the model for 300 epochs with a batch size of 128. A cosine learning rate scheduler is adopted and warm-up for 40 epochs.

In the fine-tuning stage, the images from ISIC-2017 are also resized to 256×256. We employ an encoder-decoder network similar to TransUNet [21], in which the encoder is ViT-S/16 and the decoder is several groups of the convolutional block to produce segmentation maps. The optimizer employed in the downstream segmentation task is Adam with a learning rate of 0.0003. We load parameters of ViT-S/16 from the pre-training stage and fine-tune the segment model for 100 epochs with a batch size of 16. Finally, all experiments are implemented in PyTorch with 2 NVIDIA Geforce GTX 1080Ti GPUs.

3.3. Comparison with State-of-the-art Methods

The proposed approach is compared with several transformer-based SSL methods, MAE [14] and DINO [17]. As shown in Table 1, skin lesion segmentation results are presented with different ratios of training data for fine-tuning. All pre-trained methods attain better performance in skin lesion segmentation compared to the "Random Init." method trained from

Table 1. Comparison of skin lesion segmentation trained with 1%, 10%, 50%, and 100% of official training samples on ISIC-2017. "Random Init." means training from scratch and "Supervise" denotes the full-supervision of ISIC-2019 data. "↑" means the high value the better. The Dice (%) and IoU (%) scores are presented on the test set of ISIC-2017.

Method	1%		10%		50%		100%	
	Dice↑	IoU↑	Dice↑	IoU↑	Dice↑	IoU↑	Dice↑	IoU↑
Random Init.	68.98	58.37	76.64	66.35	79.54	69.74	81.19	71.58
Supervise	73.93	62.91	79.35	69.20	81.40	71.92	81.90	72.57
MAE [14]	73.50	62.89	79.46	69.28	81.39	72.16	81.58	72.31
DINO [17]	73.17	62.41	78.85	68.87	81.37	71.92	81.38	71.91
G2LL(Ours)	**74.43**	**63.27**	**80.28**	**70.34**	**81.98**	**72.70**	**82.46**	**72.96**

scratch. Notably, our proposed SSL method surpasses all the SSL approaches and the supervised approach. In comparison to the supervised method, which is pre-trained on ISIC-2019 with classification labels, our approach shows a higher Dice score ($0.5\% \sim 0.98\%$ improvement) and IoU score ($0.36\% \sim 1.14\%$ improvement) across different partitions of training data. Our SSL method is also superior to MAE and DINO on both metrics consistently.

We visualize some samples of dermoscopy images and segmentation results generated by different methods in Fig.2. It shows that our proposed approach produces competitive segmentation performance with supervised pre-training methods. In comparison to the SSL method, our method generates more stable segmentation which is close to the ground truth.

3.4. Ablation Study

We conduct extensive ablation experiments to demonstrate the effectiveness of different mask strategies in the MIM process. The original mask strategy is a random mask with a fixed mask ratio, it generates sub-optimal results on dermoscopy images. We introduce the semantic mask, a strategy for masking image patches under different structured areas. A further improvement is the semantic focal mask strategy, which focally masks the key structured areas where skin lesions probably exist. The mask ratio γ is 25% and focal weight β is 70% in our experiments.

Table 2. Ablation experiments on different mask strategies for MAE pretrained on the ISIC-2019 dataset. All experiments are fine-tuned on 100% of training samples on the ISIC-2017 dataset. "↑" means the high value the better.

Mask Strategy	Dice↑ (%)	IoU↑ (%)
Random Mask	81.54	71.97
Semantic Mask	81.76	72.22
Semantic Focal Mask	82.16	72.62

The fine-tuning results of the above experiments are presented in Table 2. Compared to the random mask, the semantic mask achieves a 0.22% improvement in the Dice score and a 0.25% improvement in the IoU score. It verified that structure prior guidance is beneficial for useful representation learning. Furthermore, the semantic focal mask attains improvement with 0.62% in the Dice score and 0.65% in the IoU score, respectively.

4. CONCLUSION

This paper proposed a Global-to-Local self-supervised Learning (G2LL) approach to improve the performance of skin lesion segmentation in dermoscopy images with global and local context modeling. Experiments are conducted on the downstream segmentation task, where the results have shown that the proposed method has achieved superior performance compared to the latest methods. It is also noticeable that our method has outperformed the full-supervised learning technique, indicating that the representations learned by the SSL method will be more useful for downstream lesion segmentation. The extensive ablation experiments have clearly verified the improvement of our structure-wise masking strategy. Eventually, to account for computational costs, all experimental results are reported by a single run.

In the future, we will explore the influence of semi-supervised learning in the context of label-efficient skin lesion segmentation regarding dermoscopy images.

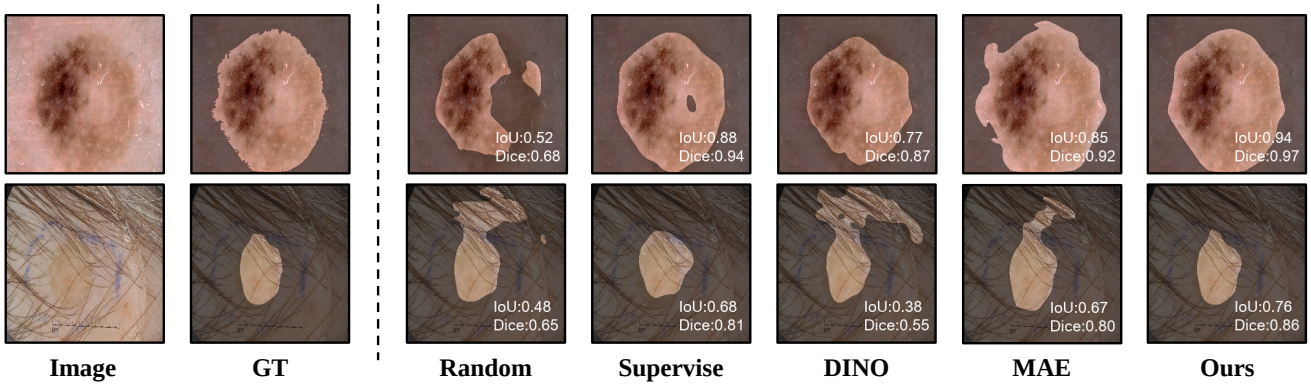

Fig. 2. Visual comparison of skin lesion segmentation results on two samples from the test set of ISIC-2017. The IoU and Dice scores are displayed in the lower right corner of each image. Note that the image on the bottom contains hairs.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [19, 18]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. REFERENCES

[1] J. Lu, E. Kazmierczak, J. H. Manton, and R. Sinclair, "Automatic segmentation of scaling in 2-d psoriasis skin images," *Transactions on Medical Imaging*, vol. 32, no. 4, pp. 719–730, 2013.

[2] 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.

[3] Y. Fisher and K. Vladlen, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations*, 2016.

[4] H. Li, X. He, F. Zhou, Z. Yu, D. Ni, S. Chen, T. Wang, and B. Lei, "Dense deconvolutional network for skin lesion segmentation," *Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 527–537, 2018.

[5] M. M. K. Sarker, H. A. Rashwan, F. Akram, S. F. Banu, A. Saleh, V. K. Singh, F. Chowdhury, S. Abdulwahab, S. Romani, P. Radeva, et al., "SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 21–29.

[6] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *Transactions on Medical Imaging*, vol. 40, no. 2, pp. 699–711, 2020.

[7] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 251–266.

[8] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, "Boundary-aware transformers for skin lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 206–216.

[9] W. Cao, G. Yuan, Q. Liu, C. Peng, J. Xie, X. Yang, X. Ni, and J. Zheng, "ICL-Net: Global and local inter-pixel correlations learning network for skin lesion segmentation," *Journal of Biomedical and Health Informatics*, 2022.

[10] J. Wang, F. Chen, Y. Ma, L. Wang, Z. Fei, J. Shuai, X. Tang, Q. Zhou, and J. Qin, "XBound-Former: toward cross-scale boundary modeling in transformers," *arXiv preprint arXiv:2206.00806*, 2022.

[11] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *IEEE International Conference on Computer Vision*, 2021, pp. 9640–9649.

[12] Z. Zhou, V. Sodha, J. Pang, M. B Gotway, and J. Liang, "Models genesis," *Medical Image Analysis*, vol. 67, pp. 101840, 2021.

[13] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *International Conference on Learning Representations*, 2021.

[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16000–16009.

[15] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Superpixels compared to state-of-the-art superpixel methods," Tech. Rep., 2012.

[16] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9653–9663.

[17] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *IEEE International Conference on Computer Vision*, 2021, pp. 9650–9660.

[18] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[19] N. CF 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 *International Symposium on Biomedical Imaging*. IEEE, 2018, pp. 168–172.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[21] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.