https://doi.org/10.2352/EI.2024.36.15.COIMG-119

Transformers for Microscopy Slide Image Segmentation of Invasive Melanoma

Franklin Wang; University of California Berkeley; Berkeley, California, USA
Michael Wang; University of California San Francisco, California, USA
Avideh Zakhor, University of California Berkeley; Berkeley, California, USA
Timothy McCalmont; University of California San Francisco, California, USA

Abstract

Prognosis for melanoma patients is traditionally determined with a tumor depth measurement called Breslow thickness. However, Breslow thickness fails to account for cross-sectional area, which is more useful for prognosis. We propose to use segmentation methods to estimate cross-sectional area of invasive melanoma in whole-slide images. First, we design a custom segmentation model from a transformer pretrained on breast cancer images, and adapt it for melanoma segmentation. Secondly, we finetune a segmentation backbone pretrained on natural images. Our proposed models produce quantitatively superior results compared to previous approaches and qualitatively better results as verified through a dermatologist.

Introduction

According to the Center for Disease Control [1], skin cancer is the most common type of cancer in the United States. It is estimated that in 2022, 197,900 people have been diagnosed with melanoma, representing around 5.2 % of all cancer cases in the United States. Out of all of the types of skin cancer, melanoma is by the far the most serious [2]. Melanoma originates in melanocytes or pigment-producing cells in the epidermis. Melanoma in the epidermis is called in-situ melanoma, and it is typically low-risk. Melanoma that invades past the epidermis into the dermis is known as invasive melanoma, and is a sign of high-risk cancer. The primary invasive melanoma tumor size at the time of diagnosis is a crucial prognostic factor for survival prediction and clinical management. Over-staging a melanoma can subject patients to unnecessary risks from procedures and studies, resulting in undue financial burden on the health care system. The average annual cost of treating melanoma is estimated at $3.3 billion in the United States [3]. Therefore, accurate assessment of invasive tumor size is an early critical step in appropriate patient care and utilization of health care resources. Typically, tumor size is estimated from stained images of patient skin biopsies imaged with microscopes. The current clinical practice is to use a 50 year old prognostic metric called Breslow Thickness (BT), a one dimensional proxy for the melanoma tumor volume within the dermis. The BT is the distance from the surface of the epidermis to the deepest part of the malignant tumor within the dermis [4]. BT's main shortcoming is that it is a simple distance measurement in one dimension and cannot accurately describe a 3-dimensional tumor burden. It fails to account for variation in epidermal thickness, tumor diameter and density. [5] provides evidence that the cross-sectional area of the tumor is vital for more accurate forecasting of patient outcomes. Despite the shortcomings of BT, it is still being relied on due to its reproducibility and ease of use [6]. To overcome the limitations of BT, [5] proposed a manual method to estimate the invasive tumor cross-sectional area, which better predicts mortality than BT. However, this manual method is time-intensive with high inter-observer variability, thereby limiting its clinical utility and adoption [6]. Given that tumor-cross section evaluation is a segmentation exercise, we hypothesize that computer vision based approaches can be utilized to great effect. Segmentation maps contain detailed geometric information about invasive melanoma. These maps can then be further measured to provide metrics including BT, cross-sectional area, density, and shape. The cross-sectional evaluation provides additional information that would be invaluable for staging and management planning, and could significantly impact the standard of care.

Current work on segmenting melanoma such as [7], [8], [9], and [10] use older, simple convolutional models such as U-Net [11], like. Even though these approaches might have multi-stage models [7], different sampling methods [10], or segment different structures such as cell nuclei [9], their model designs adhere to well-studied and simple architectures, which may limit performance and model expressivity. A recent work on invasive melanoma segmentation used a two-stage multi-resolution convolutional model [12], with the first step segmenting the epidermis and the other segmenting all melanoma, i.e. in-situ and invasive melanoma. This two-stage method is inspired by the fact that the in-situ melanoma is visually similar to the invasive melanoma. To distinguish between the two [12] would segment all melanoma and then rule out the in-situ melanoma using the epidermis predictions to obtain invasive melanoma predictions. The models developed in [12] are HRNet-OCR [13] and HookNet [14], which are massive convolution-based models selected for their multiscale and context modelling properties. There are several problems with the approach in [12]. Both segmentation models are both overparameterized for a small dataset, as they contain 80-100 million parameters for a training set of 43 whole slide images (WSIs). In addition, having two models doubles the training time and computational costs. A viable alternative is to train a single network to segment both the invasive melanoma and epidermis at the same time, thus halving the training time and reducing overparamerization. This way, we can reduce the problem to one three-class segmentation task, rather than two binary segmentation tasks.

In this paper, we propose two transformer models, each for solving the above three class segmentation task. Each model has its own unique internal representation and is pretrained on a different data set. Our models achieve state-of-the-art results in melanoma segmentation without using established, simple convolutional models most commonly used in medical computer vision. We also show that multi-scale modelling and representations result in superior segmentation performance over generic single-scale transformer designs.

Dataset

Our dataset is identical to [12] and contains 55 total slide images. We partition 43 images as the training set and 12 images as the testing set. The images contain 6 labels: background cells, epidermis, invasive melanoma, inflamed tumor, fibrotic tumor, and uncertain tumor. We note that the in-situ melanoma is considered part of the epidermis in our labels, and that the only type of melanoma that is labeled is invasive melanoma. Some of the classes are not finely labelled, especially the fibrotic tumor and inflamed tumor regions. From a pathologist's perspective, the boundaries of these regions are inherently more ambiguous than other well-defined areas such as invasive melanoma and epidermis. To avoid segmenting those regions, we transform the data from the original 6 classes into 3 semantic classes. (1) other, which contains the background cells, fibrotic tumor, inflamed tumor, and uncertain tumor; (2) invasive melanoma; and (3) epidermis.

Proposed Method

In the next two subsections, we will describe our single scale and multi-scale transformer models. Figure 1 shows a high-level schematic of both models.

Single-Scale Transformer

[16] released Hierarchical Pyramid Transformer (HIPT) pretrained on WSIs of breast tissue via student-teacher distillation. These models are used as an encoder backbone component of the network. To our knowledge, these are the only transformer networks pretrained with WSIs of biological tissues. Due to the commonality in biological features between breast tissue and skin tissue and cancer in general, a model pretrained on breast cancer WSIs will likely bolster the performance of invasive melanoma segmentation. However, there are several challenges that need to be overcome with using the HIPT models. First, HIPT uses the original vision transformer backbone [15] , which only contains single-scale low resolution representations. Multi-scale representations have proven to perform best for segmentation tasks since objects can exist at multiple scales. Therefore, we propose a decoder mechanism that constructs multi-scale hierarchical features to use with these pretrained HIPT models based on [17]. Specficially, we investigate three different decoder designs for HIPT models, denoted by baseline, adapter, and all-MLP. A high-level schematic of the three decoder mechanisms can be found in Figure 2. The baseline decoder simply uses resampled feature maps plus the Uperhead [18] feature aggregation mechanism. The all-MLP decoder uses resampled feature maps with an all-MLP segmentation head. The adapter decoder uses cross-attention between the HIPT backbone and convolutional feature maps to construct multi-scale features with an Uperhead segmentation head.

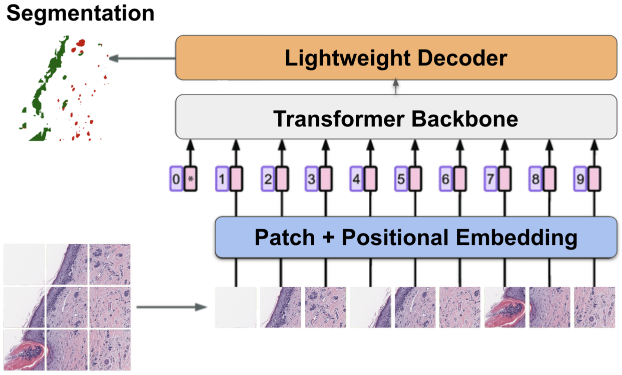

Figure 1. *Proposed pure transformer model*

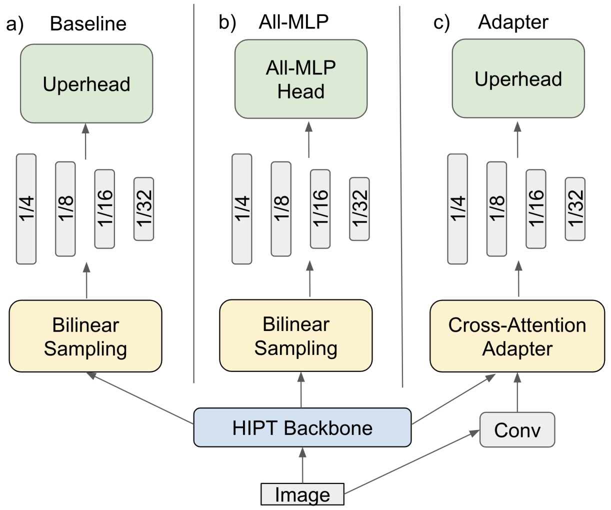

Figure 2. *Three decoder types for HIPT models.*

Multi-Scale Transformer

In this section, we propose to directly utilize a hierarchical transformer backbone rather than adapting other non-hierarchical models such as HIPT to perform segmentation [19]. SegFormer is an appropriate model for our task because unlike many other transformers, it has built-in hierarchical structures with multi-scale feature maps. Another advantageous property of SegFormer is the lack of positional embeddings. Typically, vision transformers need to interpolate positional embeddings if the resolution of the images for a finetuning task is different from the resolution the model was pretrained with. This interpolation allows the transformer to handle multiple resolutions, but also introduces artifacts that lower performance. SegFormer skips positional encodings altogether by using zero-padded convolutions to produce positional representations. Lastly, SegFormer is trained on ImageNet [20], which has repeatedly proven to confer powerful visual representations generalizable to many tasks.

Experimental Results

We train our models on a machine with 3 NVIDIA Quadro RTX 8000 GPUs with PyTorch. We use the Adam optimizer [21] with a learning rate of 0.00006 and with a weight decay of 0.01. We use the linear decay scheduler with linear warmup for the learning rate scheduler. We apply dropout on the final segmen-

Model	mIoU	melanoma IoU	F1
Multi-Scale FCN [8]	0.538	0.130	0.140
Best 2-stage [12]	0.640	0.291	0.440
Best HIPT Model	0.696	0.401	0.573
Best SegFormer Model	**0.719**	**0.447**	**0.618**

Table 1. Quantitative comparison of our proposed method and previous approaches.

tation head layer and also the positional embeddings for the HIPT models. We use pixel-wise cross entropy loss. Both models take approximately 1 day to train 100 epochs with a batch size of 16 per GPU.

Table 1 shows the best results for each type of transformer model and also the best results of the existing methods in [12] and [8]. The best SegFormer and HIPT models outperform the best model from [12] in mIoU by 12% and 9% respectively. There are two reasons behind this; first, convolutional networks do not model global long-range contexts well because of their narrow receptive field. By contrast, the receptive field of the transformer is the entire size of the image after only the first self-attention layer. Small and scattered melanoma is the most difficult to segment because it is sparse, and can be present across long ranges in an image sample. As seen in Figures 3, 4 and 5, transformer-based architectures fare better in this long-range modelling task for scattered melanoma. The second reason is that transformers exhibit superior generalization ability because they lack the inductive biases in convolutional networks. The best model from [12] contains 80M parameters, which is significantly more than the 58.1M and 27.5M in the best HIPT and SegFormer models respectively. More examples of HIPT vs. SegFormer vs. [12] are included in [22].

As shown in Table 1, SegFormer mIoU is 0.02 higher than HIPT. Unlike HIPT, SegFormer is a custom-designed architecture for segmentation with multi-scale, hierarchical feature maps. In contrast, for HIPT, we had to introduce a multi-scale feature adapter system to produce hierarchical feature maps necessary for segmentation. In particular, we notice that the segmentation maps by HIPT underperform in detecting sparse and small melanoma, which may indicate that the internal representations have too low of a resolution.

Model	Resolution	mIoU	melanoma IoU	F1
Adapter	**512**	**0.696**	**0.401**	**0.573**
Adapter	768	0.652	0.311	0.475
Adapter	1024	0.644	0.3298	0.460
Baseline	512	0.678	0.363	0.533
Baseline	1024	0.670	0.348	0.517

Table 2. Results on different patch sizes for HIPT.

Model	Resolution	mIoU	melanoma IoU	F1
Seg. B0	512	0.694	0.397	0.568
Seg. B0	1024	0.695	0.398	0.569
Seg. B1	512	0.689	0.386	0.557
Seg. B1	1024	0.717	0.441	0.613
Seg. B2	**512**	**0.719**	**0.447**	**0.618**
Seg. B2	1024	0.708	0.424	0.595

Table 3. Results on different patch sizes for SegFormer.

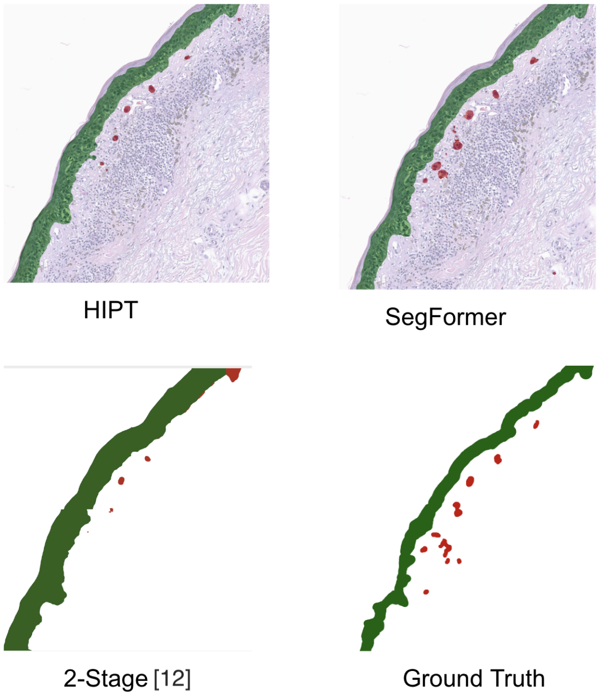

Figure 3. The method from [12] fails to detect scattered melanoma, even though the tumors exist at approximately the same scale. Meanwhile, the transformer-based models are able to detect most of the melanoma nodules, with Segformer achieving the best performance

A possible reason for SegFormer to have outperformed the custom-designed HIPT models is positional encoding. HIPT was pretrained on 256×256 images at 20× magnification. Therefore, to accommodate our dataset of 40× images at higher resolutions, there is a mismatch in positional encodings which results in performance decrease. As seen in Table 2, larger patch sizes for HIPT tend to worsen performance, hinting that positional encoding interpolation negatively impacts performance. SegFormer on the other hand lacks positional encodings, so resolution is not as important of a factor for segmentation performance. For the SegFormer models, Table 3 shows no clear trends with patch size and performance even though larger contexts contain more information for segmentation.

Model	Params	mIoU	melanoma IoU	F1
Adapter	**58.1M**	**0.696**	**0.401**	**0.573**
Baseline	33.0M	0.678	0.363	0.533
All-MLP	25.0M	0.652	0.314	0.478

Table 4. Best results on using different decoders for HIPT. The resolution used for this experiment was 512×512.

As seen in Table 4, the best performing HIPT model is the adapter decoder design. We speculate this to be due to the differences in constructing hierarchical feature maps. Specifically, the baseline and all-MLP decoders resample feature maps of a fixed resolution to a desired resolution from $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$. The *adapter* model on the other hand uses cross-attention with convolutional features to construct the multi-scale feature maps from the HIPT backbone. The *baseline* architecture outperforms

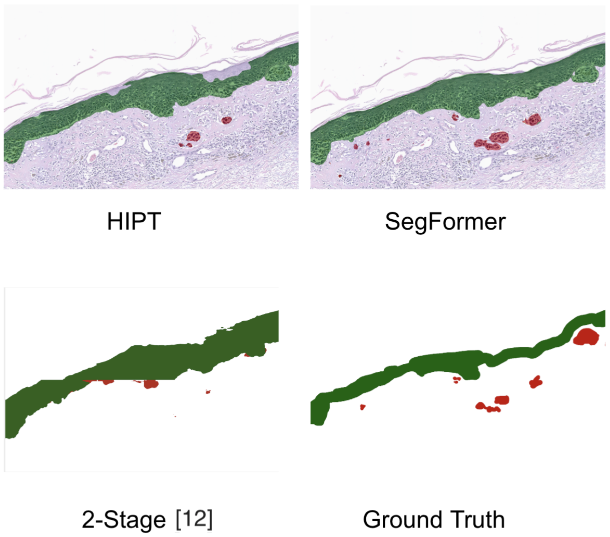

Figure 4. *The method from [12] segments the epidermis poorly and contains a large number of false positives and negatives. SegFormer is closest to the ground truth.*

the *all-MLP* architecture, which may be due to MLP architectures needing more data to generalize.

Table 5 shows a general performance boost with ascending SegFormer model sizes, with the best being the SegFormer B2. This is most likely because larger models exhibit superior capacity for large-scale pretraining which is consistent with other observations in [15]. That being said, we have also found that SegFormers B3 and B4, which are even larger models than B2, have poor segmentation performance due to non-convergence in the training. ImageNet pretraining is also a reason for SegFormer models outperforming the HIPT models. The HIPT pretraining dataset of breast cancer WSIs only has about $\frac{1}{3}$ the size of ImageNet in terms of total samples. This affects the generalization of visual representations, considering that ImageNet contains more diverse scenes than just breast cancer WSIs as well as more total samples.

Model	Params	mIoU	melanoma IoU	F1
Seg. B0	3.7M	0.695	0.398	0.569
Seg. B1	13.7M	0.717	0.441	0.613
Seg. B2	**27.5M**	**0.719**	**0.447**	**0.618**

Table 5. *Results on different SegFormer sizes*

Conclusions and Future Work

We proposed two transformer-based methods which outperform the state-of-the-art method in [12] with convolutional backbones by up to 12% in mIoU with less training time and memory. Our SegFormer models slightly outperform HIPT models due to the inherent multi-scale architectural design of SegFormer. Our HIPT adapter model uses learnable network modules rather than simple resampling to construct multi-scale features, resulting in superior segmentation performance. Future work could focus on addressing class imbalances with other types of losses or sampling strategies, as healthy tissues and cells vastly outnumber diseased tissues in living patients. Due to the high resolution nature of whole-slide images, limitations of segmentation annotation tools for physicians, and also the inherent variance in annotations between different physicians, noisy annotations are inherent to WSI segmentation. Another avenue of future work would be to incorporate methods to better handle noisy annotations. Lastly, HIPT is one of the first publicly available foundation models for histopathological data, and much further room for exploration exists for such building on top of such foundation models. In particular, vision-language models such as [23] are interesting due to the development of language-guided segmentation [24], which could offer more interpretability for physicians and fine-grained control over segmentation..

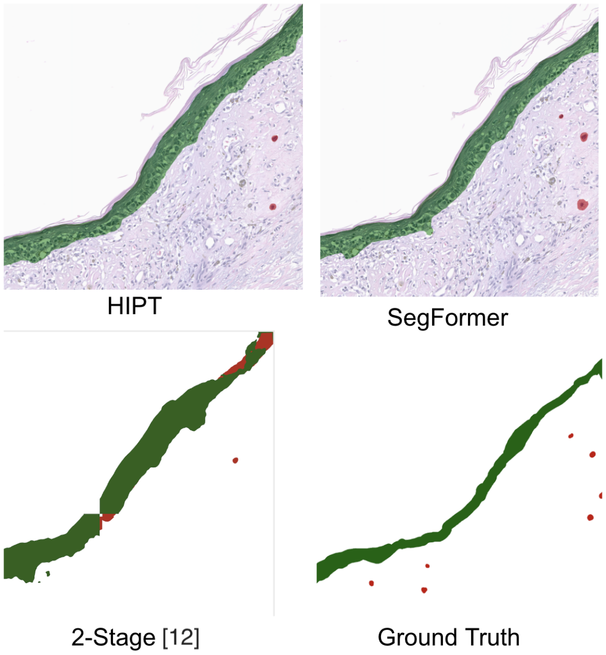

Figure 5. *The method from [12] fails to segment the scattered melanoma and also contains artifacts at the edges of the epidermis. SegFormer is closest to the ground truth.*

References

[1] Kinds of cancer," Centers for Disease Control and Pre- vention, Jun 2022.

[2] "Melanoma," The Skin Cancer Foundation.

[3] Skin cancer facts statistics," Skin Cancer Facts amp; Statistics : National Council on Skin Cancer Prevention

[4] Alexander Breslow, "Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma," Annals of Surgery, vol. 172, no. 5, pp. 902–908, 1970.

[5] Gerald Saldanha and et al., "Development and initial validation of calculated tumor area as a prognostic tool in cutaneous malignant melanoma," JAMA Dermatology, vol. 155, no. 8, pp. 890–898, 2019.

[6] Timothy H. Mccalmont, "The second dimen- sion—integrating calculated tumor area into cancer di- agnosis," JAMA Dermatology, vol. 155, no. 8, pp. 883, 2019

[7] Shima Nofallah, Linda G. et al., and et al., "Segmenting skin biopsy images with coarse and sparse annotations using u-net," Journal of

Digital Imaging, vol. 35, no. 5, pp. 1238–1249, 2022.

[8] Adon Phillips, Iris Teo, and Jochen Lang, "Seg- mentation of prognostic tissue structures in cutaneous melanoma using whole slide images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019

[9] Salah Alheejawi and et al., "10 - deep learning-based histopathological image analysis for automated detec- tion and staging of melanoma," in Deep Learning Tech- niques for Biomedical and Health Informatics, pp. 237– 265. Academic Press, 2020.

[10] Mike van Zon, Nikolas Stathonikos, Willeke A.M. Blokx, Selim Komina, Sybren L.N Maas, Josien P.W. Pluim, Paul J. van Diest, and Mitko Veta, "Segmenta- tion and classification of melanoma and nevus in whole slide images," in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 263–266

[11] laf Ronneberger and et al., "U-net: Convolutional net- works for biomedical image segmentation," in Medi- cal Image Computing and Computer-Assisted Interven- tion – MICCAI 2015, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., 2015, pp. 234–241.

[12] Neil Neumann, Michael Wang, Amal Mehta, Aman Shah, Mara Olson, Wudi Fan, Anna Weier, Avideh Zakhor, and Timothy McCalmont, "Quantifying inva- sive melanoma volume by deep learning segmentation at the pixel-level," in LABORATORY INVESTIGATION. SPRINGERNATURE CAMPUS, 4 CRINAN ST, LON- DON, N1 9XW, ENGLAND, 2022, vol. 102, pp. 346– 347

[13] Yuhui Yuan, Xilin Chen, and Jingdong Wang, "Object- contextual representations for semantic segmentation," in 16th European Conference Computer Vision (ECCV 2020), August 2020.

[14] Mart van Rijthoven and et al., "Hooknet: Multi- resolution convolutional neural networks for seman- tic segmentation in histopathology whole-slide images," Medical Image Analysis, vol. 68, pp. 101890, 2021.

[15] Alexey Dosovitskiy and et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. 2021, OpenReview.net.

[16] ichard Chen and et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16123–16134

[17] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In International Conference on Learning Representations (ICLR), 2023

[18] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun, "Unified perceptual parsing for scene understanding," in European Conference on Computer Vision. Springer, 2018.

[19] Enze Xie and et al., "Segformer: Simple and efficient design for semantic segmentation with transformers," Advances in Neural Information Processing Systems, vol. 34, 2021

[20] Kia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[21] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," International Conference on Learning Representations, 12 2014.

[22] "Microscopy slide image segmentation of invasive melanoma," Master's Thesis Department of Electrical Engineering and Computer Sciences University of California Berkeley, Jan. 2023

[23] Lu, Ming Chen, Bowen Williamson, Drew Chen, Richard Liang, Ivy Ding, Tong Jaume, Guillaume Odintsov, Igor Zhang, Andrew Le, Long Gerber, Georg Parwani, Anil Mahmood, Faisal. (2023). Towards a Visual-Language Foundation Model for Computational Pathology.

[24] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B. (2023). Segment Anything. ArXiv, abs/2304.02643.

Author Biography

Franklin Wang received his BS degree in electrical engineering from the University of California Los Angeles (2020) and his MS in electrical engineering and computer science from University of California Berkeley (2022). His work primarily focuses on practical applications of computer vision.