Microscopy Slide Image Segmentation of Invasive Melanoma

Franklin Wang
Mike Wang
Avideh Zakhor

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-10
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-10.html

January 16, 2023

Microscopy Slide Image Segmentation of Invasive Melanoma

by

Franklin Wang

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Avideh Zakhor, Chair
Professor Joseph Gonzalez
Professor Dorit Hochbaum

Fall 2022

The thesis of Franklin Wang, titled Microscopy Slide Image Segmentation of Invasive Melanoma, is approved:

Chair  Date 1/16/23

Date

Date

University of California, Berkeley

Microscopy Slide Image Segmentation of Invasive Melanoma

Abstract

Microscopy Slide Image Segmentation of Invasive Melanoma

by

Franklin Wang

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Avideh Zakhor, Chair

Melanoma is an aggressive form of skin cancer, where survival rates are high when caught early. Breslow thickness is a measure of the depth of tumor into the skin, which provides a metric on how far the melanoma has metastasized into the deeper regions of the skin. Traditionally, the Breslow thickness measurement is used to determine the stage and severity of melanoma even though it does not take into account cross-sectional area, which has been shown to be more useful for prognostic and treatment purposes. We propose to use computer vision based methods to estimate cross-sectional area of invasive melanoma by segmenting it out in whole-slide images (WSIs) from microscopes. We present two transformer-based methods to segment invasive melanoma. First, we design a custom segmentation model from a transformer backbone for classification pretrained on breast cancer WSIs, and adapt the architecture accordingly to perform melanoma segmentation. In our second approach, we utilize a segmentation backbone pretrained on natural images and finetune it for the melanoma segmentation task. Both proposed approaches outperform existing work in terms of mean intersection over union by up to 9% and 12% respectively, while also being more memory efficient and easier to train. Analysis of our segmentation results from a board-certified dermatologist reveals that our models perform well compared to the trained human eye.

To my parents and my girlfriend, Yang.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Professor Avideh Zakhor for her support and her many hours spent advising me. Thank you for teaching me so many things about academic research and for all the guidance on my journey. Thank you to Professor Joseph Gonzalez and Professor Dorit Hochbaum for being the second and third readers for my thesis. I would also like to thank my brilliant collaborator Mike Wang, for offering so much advice from the medical side of this project and for his countless hours spent labelling segmentation data. Lastly, thank you to Aman Shah and Amal Mehta who helped me get started with this project.

Chapter 1

Introduction

According to the Center for Disease Control [17], 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 [22]. Melanoma originates in melanocytes or pigment-producing cells. 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 has the potential to spread to lymph nodes and distant organs in a deadly process called metastasis. Invasive melanoma represents around 50% of all melanoma cases on an annual basis [22] Detecting melanoma early in the in-situ stage yields a high survival rate, around 99% [22] . The survival rate decreases to 68% when invasive melanoma spreads to the lymph nodes, and further decreases to 30% when metastatic melanoma is present [23].

The primary invasive 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 such as sentinel lymph node biopsy, whole-body PET/CT, and immunotherapy, 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 [1]. 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 hematoxylin and eosin (H&E) 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). This thickness is 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 [5]. This measurement is used to classify invasive melanoma into T stages within the tumor (T), node (N), and metastasis (M) TNM staging system, which is correlated to patient survival. Figure 1.1 shows an example of a BT measurement on a microscopic image.

BT's main shortcoming is that it is a simple distance measurement in one dimension and

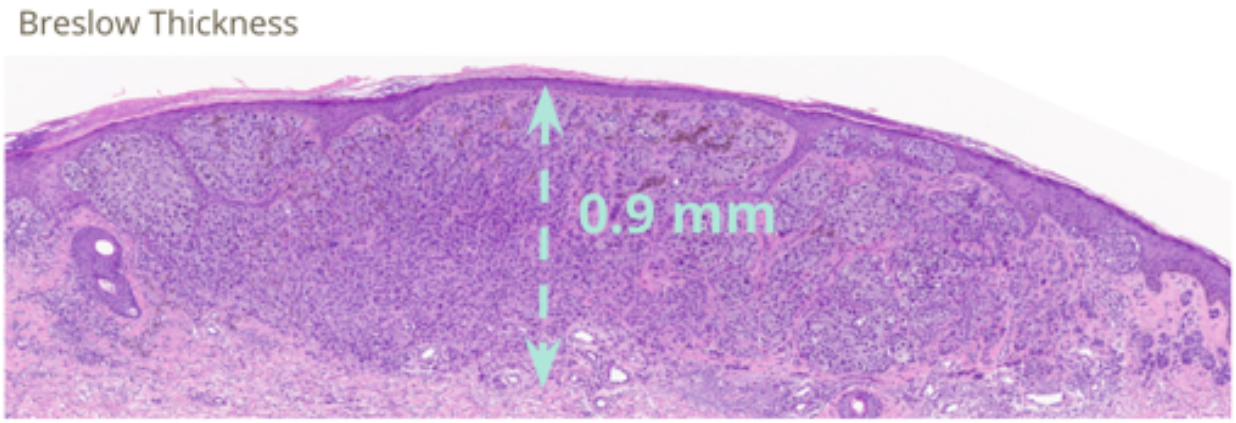


Figure 1.1: Breslow thickness is a measurement from the surface of the epidermal granular layer to the point of maximum tumor thickness perpendicular to the epidermis where the tumor originates.

cannot accurately describe a 3-dimensional tumor burden. It fails to account for variation in epidermal thickness, tumor diameter and density. The BT is overestimated when the tumor is deep but has a narrow diameter, a dispersed distribution, or a thick epidermis. It is underestimated when the tumor is shallow but wide. [30] 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 [21].

To overcome the limitations of BT, [30] proposed a manual method that estimates the invasive tumor cross-sectional area, which better predicted mortality than BT. Figure 1.2 shows a visualization of this manual area estimation process from a pathologist's perspective. As seen from the many regions in Figure 1.2 the pathologist has to pay attention to, this manual method is time-intensive with high inter-observer variability, thereby limiting its clinical utility and adoption [21]. Given that tumor-cross section evaluation is a segmentation exercise, we hypothesize that computer vision based approaches can be utilized to great effect. In particular, semantic segmentation, the task of classifying every pixel of an image, is a proven tool for automated image analysis in the medical domain. 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.

This methodology has the potential to be generalized to other invasive malignancies, such as small cell lung cancer, breast cancer, and squamous cell carcinoma. For these cancers, we could not identify previous publications that modeled the correlation between automated tumor area and survival time while differentiating between in-situ and invasive cancer, which

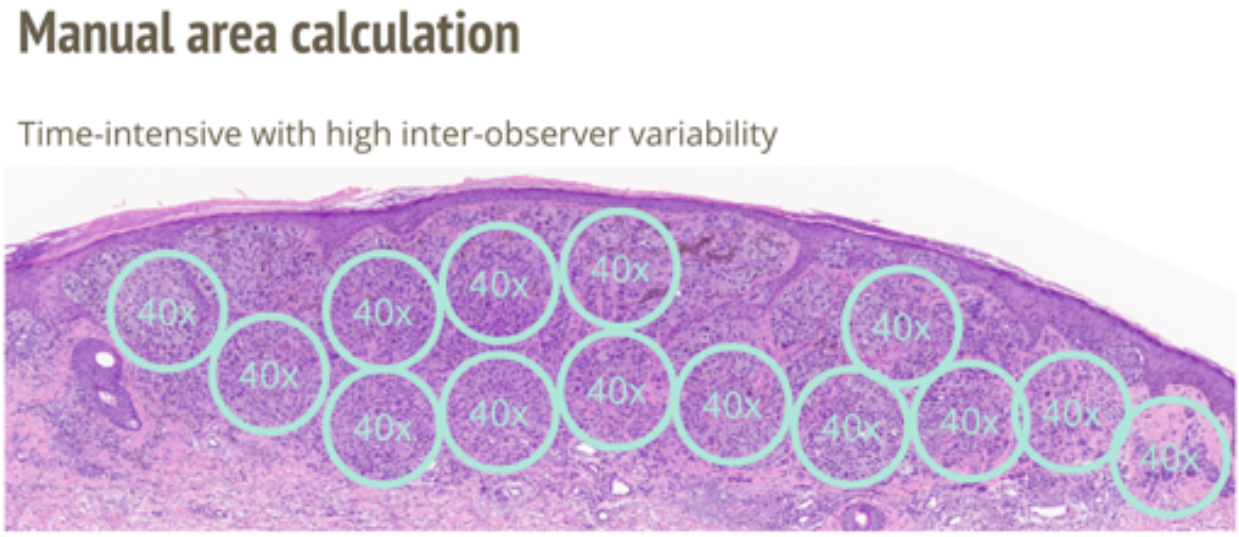


Figure 1.2: Manual calculation of tumor area is time-consuming, difficult, and potentially inaccurate.

are the highlights of our work for melanoma.

The rest of this thesis explores the development of such a semantic segmentation model for identifying and quantifying invasive melanoma, and the challenges that come with analyzing pathology images of skin tissue from a computer vision perspective. The outline of the thesis is as follows. In Chapter 2, we describe existing work related to our problem along with the potential shortcomings and areas for improvement in these works. Chapter 3 describes our two proposed models in detail, and Chapter 4 presents the experimental results and discussion of our methods. Lastly, Chapter 5 draws conclusions from our work and describes directions for future work. Appendix A contains segmentation results for our entire test set.

Chapter 2

Related Work

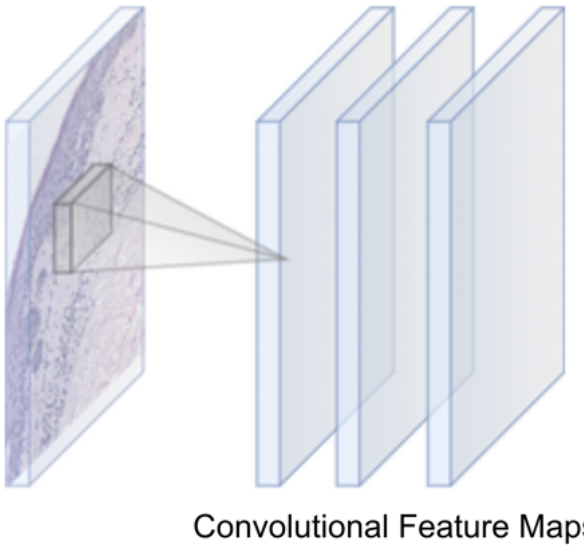
In this chapter, we will go over existing work in melanoma segmentation in the medical and traditional computer vision space. The outline of this chapter is as follows. In Section 2.1 and 2.2, we examine segmentation approaches in mainstream computer vision and segmentation approaches for medical computer vision respectively. In Section 2.3, we review methods that segment melanoma in WSIs. Lastly in Section 2.4, we summarize work in deep learning in the context of WSI analysis.

2.1 Semantic Segmentation in Computer Vision

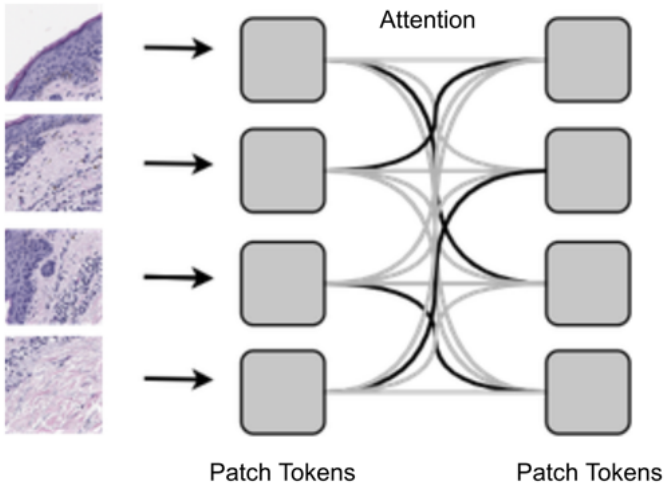
Architectures in traditional computer vision can be approximately divided into three major categories: convolutional models (CNNs), transformer models, and hybrid models. Figure 2.1 visualizes the fundamental building block operations of convolutional and transformer networks. Convolutional models use local convolutions as the building block of their backbones and construct multi-scale feature maps. UNet [29] is an early, lightweight semantic segmentation model with a symmetric encoder-decoder and has propagated to tasks outside of traditional computer vision. PSPNet [35] uses a more powerful ResNet based backbone and aggregates multi-scale features in its decoder. The decoder design from this approach is present across many different segmentation approaches now. DeepLab [8] and its successors use dilated convolutions to perform upsampling to obtain a pixelwise output.

Hybrid architectures blend convolutional backbones mixed with attention operations from transformer models. HRNet-OCR [34] is a segmentation model with a convolutional backbone that uses cross-attention between features of different scales between the encoder and decoder to "mix" features of multiple contexts and scales.

Pure vision transformers have recently made large advancements in computer vision [12]. Pure vision transformer models use only attention layers and MLPs as their backbones, and treat image patches as a token sequence with positional encodings similar to word tokens in natural language processing [12]. Unlike convolutional models, they do not share the inductive biases and translation invariance assumed by the convolution operation [12].



(a) Convolutional networks



(b) Transformer networks

Figure 2.1: Convolutional networks such as U-Net operate on limited receptive field per layer, making global-context modelling harder until deeper layers of the networks. Transformer networks divide an image into patches and model relationships of pixels within and between patches as well via self-attention. Self-attention operations can model global contexts as early as the very first layer of a network.

Self-attention layers in vision transformers allow for simultaneous global and local context modelling in even shallow layers. In contrast, convolutional networks need many cascaded convolutional layers to properly model long-range visual features due to receptive field limitations [12]. These properties give vision transformers superior performance on a wide variety of computer vision tasks, but the caveat is that they need larger amounts of training data because they lack inductive biases about images. SETR [36] adopts the original pretrained vision transformer backbone from [12] to perform semantic segmentation. Although it achieves great performance on the community benchmark ADE20K dataset [37], it suffers from high computational costs and only has single-scale, low-resolution internal representations. Pure transformer models with self-attention layers do not naturally construct multi-scale, hierarchical feature maps. To address these problems, Swin Transformer [20], another pure transformer model, generates hierarchical multi-scale feature maps with two key architectural changes. It only computes self-attention in certain local spatial 'windows' to model visual features of different scales, and it merges medium resolution features together at deeper layers of the network to generate lower resolution feature maps. SegFormer [33] uses a pure transformer method to model multi-scale contexts with lightweight architectures. SegFormer generates multiscale features with a patch merging scheme similar to Swin Transformer, but with a more lightweight decoder and without windowed attention.

2.2 Medical AI Approaches to Segmentation

Medical computer vision approaches have taken different directions than mainstream, traditional computer vision ones by focusing on simple models that are easier to train and generalize. Before deep learning, medical segmentation models used Markov random-field approaches like [15] and [28]. With the advent of deep learning, medical semantic segmentation methods have found large success in using the U-Net architecture for segmentation tasks, since it was originally designed for biological segmentation tasks [29]. For example [38] is an approach that builds U-Net with dense MLP skip connections between encoder and decoder layers. [31] segments WSIs using two parallel U-Nets that operate on different resolutions of data. TransUNet [7] is an approach that uses the structure of U-Net while replacing many of the convolutional layers with self-attention operations from transformers. [6] builds on top of Swin Transformer for medical image segmentation tasks by converting the Swin Transformer backbone into a symmetric U-Net encoder decoder. While these approaches feature different architectural building blocks and different mechanisms to learn on multi-scale information, their core basis tightly follows structures based on U-Net.

2.3 Segmenting Slide Images of Melanoma

This section reviews approaches for melanoma segmentation for our dataset in Section 2.3.1 and other datasets in Section 2.3.2.

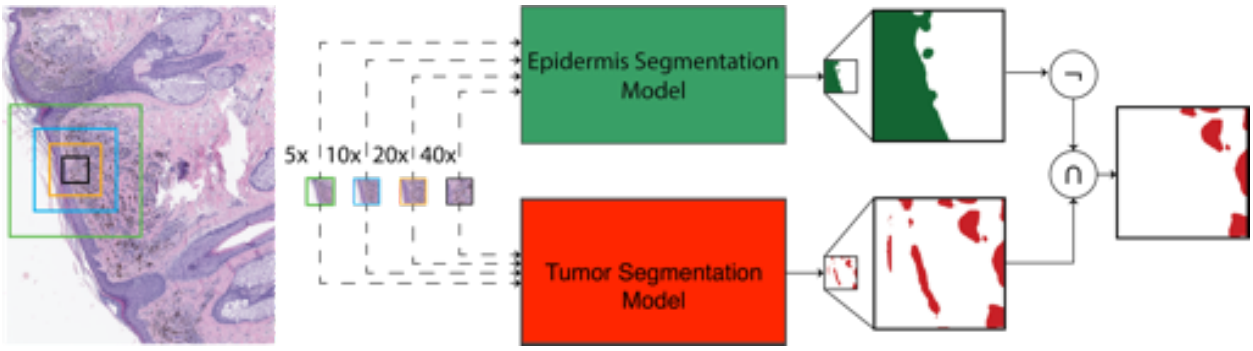


Figure 2.2: Two-stage method [24]. These method uses two HookNet or HRNet-OCR models to separately segment out epidermis and melanoma. The outputs of both of those models are combined in the end to segment out only the invasive melanoma.

2.3.1 Existing Work Designed for Our Dataset

Existing work at invasive melanoma segmentation with our dataset used a two-stage multi-resolution convolutional model [24]. Figure 2.2 visualizes the two-stage approach from [24]. One stage of the model is used to segment the epidermis, while the other is used to segment all melanoma, which includes both in-situ and invasive melanoma. This two-stage method exploited the fact that the in-situ melanoma is visually similar to the invasive one. [24] hypothesized the model might not be able to distinguish between the two types of melanoma, so it first segments all melanoma and then rules out the in-situ melanoma using the epidermis predictions to obtain the invasive melanoma predictions. The models used in [24], HRNet-OCR [34] and HookNet [31], were massive convolution-based models selected for their multi-scale and context modelling properties.

There are several problems with this approach. First, there were no labels for the in-situ melanoma, so the epidermis annotations were used as a coarse and semi-inaccurate proxy target for in-situ melanoma. This was done because all in-situ melanoma is located in the epidermis. However, this creates noisy supervision because the model learns to associate healthy epidermis tissue and melanoma as the same class. The second issue with [24] is that there are two parallel segmentation models which are both overparameterized for the small dataset, since the models contained 80-100 million parameters for a training set of 43 samples. In addition, [24] has to train both models, thus doubling 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 saving half the training time and reducing the overparamerization. This way, we can reduce the problem to one three-class segmentation task, instead of two binary segmentation tasks.

2.3.2 Existing Methods Designed for Other Datasets

We now review existing work in the segmentation of microscope slide images of melanoma and related skin tissues that do not use the dataset in this thesis. We only consider works involving microscopy slides, since macroscopic images of skin cancer taken with normal cameras contain fundamentally different image features.

[25] uses a multi-stage method to fuse melanoma segmentation networks trained on labels with different coarseness or fineness due to differences in annotation style. Both stages of the approach are simple convolutional U-Nets. [39] is segments both melanoma and nevus or skin moles using a simple 3-layer U-Net architecture. To mitigate false positives from a class-imbalanced dataset, the authors apply hard-negative mining by sampling regions where underrepresented classes are. [27] segments melanoma in a single-stage that fuses convolutional feature maps of different scales to obtain the final segmentation. The authors use multi-stride upsampling in their segmentation decoder to obtain feature maps at $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{32}$ scales, which results in superior segmentation performance compared to approaches not uinge multi-scale information. [2] segments melanoma in a different annotation setting than the other works discussed. In contrast to labelling whole regions of melanoma, it works with data where only the cancerous or non-cancerous nuclei are labelled. Older SegNet [4] and U-Net [29] convolutional models are used to perform segmentation. [26] does not segment cancer, but rather segments the epidermis portion of the skin tissue sample. Similar to many of the aforementioned works, this work uses a convolutional U-Net architecture and a sampling procedure to oversample minority classes or epidermis in this case.

The current corpus of work on segmenting melanoma generally uses older, simple convolutional models such as U-Net [29]. While these models are well-studied and are simple to train, they suffer from inferior performance on almost all traditional computer vision tasks. Basic architectures such as U-Net have been superseded by advancements in convolutional architectures and vision transformers. Our work aims to design segmentation models in ways that fuse both advancements in general architectures and techniques specifically for microscopy tissue segmentation.

The other commonality in these aforementioned works is that they do not differentiate between dangerous invasive melanoma, and harmless in-situ melanoma. In prognosis, only the invasive melanoma that has penetrated beyond the epidermis is a factor for survival. Differentiating between the invasive melanoma and harmless in-situ melanoma is a harder problem than merely segmenting all melanoma in an image. Our work focuses on segmenting only invasive melanoma while ignoring the in-situ melanoma.

2.4 Deep Learning and Whole-Slide Images

Several key characteristics of whole-slide image (WSI) data make deep learning approaches challenging. First, WSIs are high resolution with up to 150000×150000 pixels, making conventional models that operate on smaller natural images less directly feasible for modelling

WSIs. Yet simple downsampling is not a viable option for these images because downsampling procedures destroy many discriminative and important visual features. The second challenge is that independent of the tissue type, the diseased portion of an image is usually very small. This creates class imbalance issues in segmentation or patch-level classification, since the diseased class is usually highly underrepresented compared to classes representing normal, healthy tissue. The last and arguably most challenging issue is that because WSIs have large dimensions, highly accurate pixel-level annotations are rarely available. Rather, patch-level or slide-level annotations are more common, and it might not be possible to provide pixel-level labels due to the time-consuming nature of such annotations.

The most common task in WSI analysis is classification. [16] studies how CNNs can be used to classify cancer in WSIs by fusing patch-level predictions. This patch-level aggregation scheme effectively models large resolutions while still using standard CNNs for a smaller resolution patch data. [19] introduces self-supervised contrastive learning to WSI data for the whole-slide classification problems. This approach learns joint representations between patches of a WSI and the entire WSI itself. Using this approach, the multi-scale CNN model learns which patches are the most important in determining cancer phenotypes. [9] introduces student-teacher distillation to learn multi-scale representations of breast cancer tissue WSIs for tissue classification problems. This approach uses two cascaded transformers that operate at different scale hierarchies. The first lower-level transformer operates at $256x256$ patches, and the second higher-level transformer operates at 4096×4096 resolution. Since WSIs operate at fixed magnifications and thus scale, each hierarchy represents biological features at a different scale. For example, 256×256 patches represent cell-level features and 4096×4096 patches represent tissue-level features according to [9].

Since models pretrained on WSI data are quite relevant for semantic segmentation of invasive melanoma, in Chapter 3 we leverage the public transformer models pretrained on WSI data of breast cancer tissue.

Chapter 3

Proposed Approaches

In this chapter, we describe the two proposed approaches to the problem of semantic segmentation of invasive melanoma in WSIs. Now that transformer models have proven to be the state-of-the-art for dense prediction tasks such as semantic segmentation, we design transformer models for this specific segmentation problem at hand. We leverage backbones pretrained on mass data and construct multi-scale feature extraction and decoder mechanisms. Figure 3.1 outlines the general architecture of both approaches. The first method described in Section 3.1 is based on HIPT [9], and uses a classification backbone pretrained on WSIs of breast cancer tissue to extract single-scale features and transform them into multi-scale representations. The second method SegFormer [33], described in Section 3.2, uses a multi-scale backbone designed for segmentation pretrained on ImageNet [11].

3.1 Transformer Backbones Pretrained on Whole-Slide Images

Hierarchical Pyramid Transformer (HIPT) [9] released vision transformer models pretrained on whole-slide images of breast tissue via student-teacher distillation. These models are used as a backbone component of the network as visualized in Figure 3.1. To our knowledge, these are the only transformer networks pretrained with whole-slide images 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 whole-slide images 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 [12] , which only contains single-scale low resolution representations. As discussed in Section 2.1, multi-scale representations of various resolutions have proven to perform best for segmentation tasks since objects can exist at multiple scales. Therefore, a decoder mechanism that constructs multi-scale hierarchical features is a logical design to use with these pretrained HIPT models.

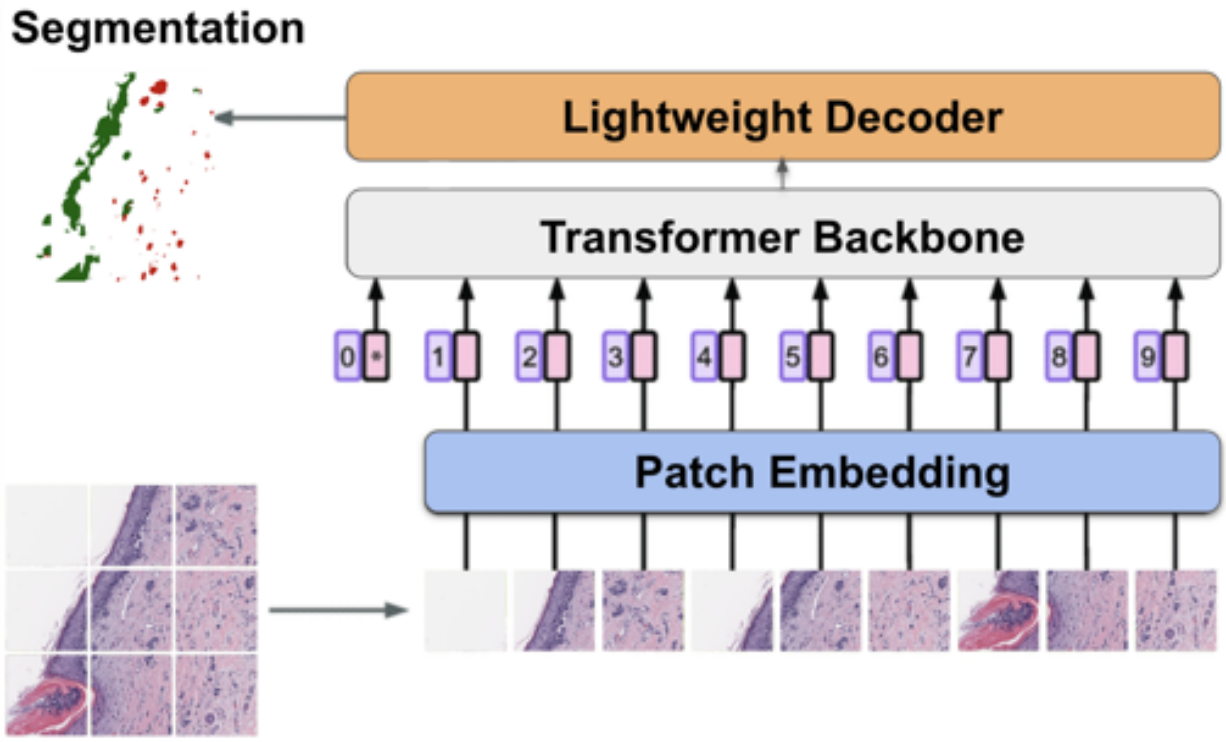


Figure 3.1: The two proposed models follow this type of transformer network architecture.

3.1.1 Backbone Description

We describe the operations of the backbone from [9]. Let H, W, and C be the respective height, width, and channels of the input image. During patch-based tokenization of the input image, a patch embedding layer splits the input image into patches of size $P \times P$, visualized in Figure 3.1. After, we obtain a feature vector z_0 of size $N \times D$, where $N = HW/P^2$ is the number of patches obtained by this splitting procedure and D is the size of the intermediate embedding dimension. Before the first self-attention layer, a learnable positional embedding vector of size $N \times D$ is added to the input vector z_0.

Let there be L self-attention blocks in series in the transformer. Each self-attention block consists of the exact same operations, visualized in Figure 3.2. Self-attention (SA) layers are defined by the following formula, where Q, K, V are different learnable linear projections of the input vector z_l:

$$V = V_{SA} z_l \tag{3.1}$$

$$K = K_{SA} z_l \tag{3.2}$$

$$Q = Q_{SA} z_l \tag{3.3}$$

$$\mathrm{SA}(z_l) = \mathrm{softmax}(QK^T)V. \tag{3.4}$$

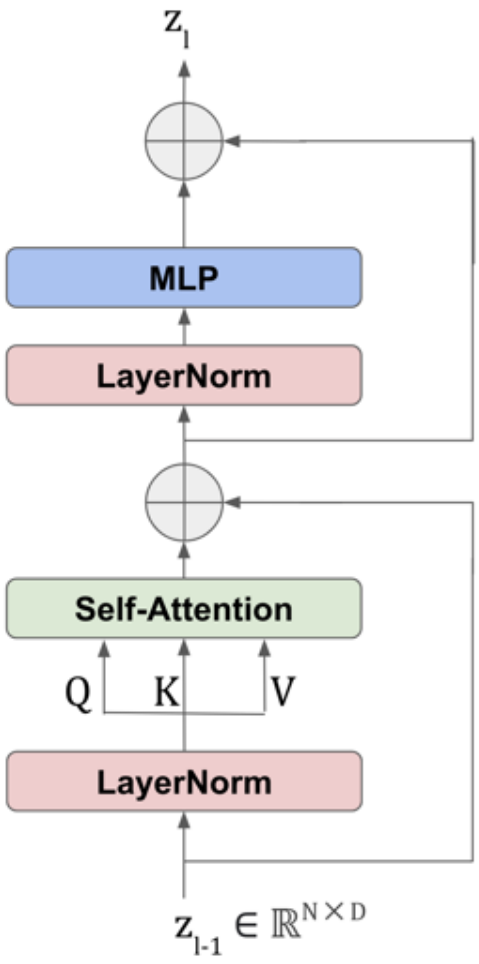


Figure 3.2: A single self-attention block of HIPT.

A complete self-attention block that accepts an input vector z_{l-1} at the layer l consists of the following:

$$z_{\text{attention}} = \text{LayerNorm}(SA(z_{l-1})) \tag{3.5}$$

$$z_{res} = z_{\text{attention}} + z_{l-1} \tag{3.6}$$

$$z_l = \text{MLP}(\text{LayerNorm}(z_{res})) + z_{res} \tag{3.7}$$

where LayerNorm stands for the layer normalization module proposed in [3]. Here the MLP consists of a fully connected layer, a GeLU activation layer [14], and another fully connected layer. The first fully connected layer expands the embedding size D to $4D$, while the second fully connected layer shrinks the embedding size from $4D$ back to D

$$\text{MLP}(z) = \text{FC}_{4D\to D}(\text{GeLU}(\text{FC}_{D\to 4D}(z))) \tag{3.8}$$

3.1.2 Decoder Designs

Since HIPT models were originally designed for classification, we have to design a decoder for segmentation to this end. We investigate three different decoder designs for HIPT models. We denote the first, second, and third decoder as *baseline*, *adapter*, and *all-MLP* respectively. A high-level schematic of the three decoder mechanisms can be found in Figure 3.3. The *baseline* decoder simply uses resampled feature maps plus the Uperhead [32] feature aggregation mechanism, visualized in Figure 3.4. The *adapter* decoder uses cross-attention between the HIPT backbone and convolutional feature maps to construct multi-scale features with an Uperhead segmentation head. The *all-MLP* decoder uses resampled feature maps with an all-MLP segmentation head.

3.1.2.1 Baseline Decoder

The *baseline* decoder, shown in Figure 3.3(a), spatially interpolates the feature maps from the intermediate layers of the vision transformer to 4 different scales. We take the intermediate feature maps after the third, sixth, ninth, and twelfth layers of the transformer at spatial size $\frac{1}{16}$ of H and W and resample them to $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ scales. We denote these feature maps as $F_{1/32}$, $F_{1/16}$, $F_{1/8}$, $F_{1/4}$. These feature maps are then passed to an *Uperhead* module shown in Figure 3.4, which is a segmentation head aggregating multi-scale features for the final prediction. $F_{1/32}$, $F_{1/16}$, $F_{1/8}$, and $F_{1/4}$ are passed to a pyramidal pooling module from [35], which combines the features in the following manner. It first takes the $F_{1/32}$ feature map and performs average pooling followed by upsampling to obtain the following feature maps, where the resolution is denoted by the subscript:

$$f_{1\times1} = \text{avgpool}(F_{1/32}) \in R^{1\times1\times D} \tag{3.9}$$

$$f_{2\times2} = \text{avgpool}(F_{1/32}) \in R^{2\times2\times D} \tag{3.10}$$

$$f_{3\times3} = \text{avgpool}(F_{1/32}) \in R^{3\times3\times D} \tag{3.11}$$

$$f_{6\times6} = \text{avgpool}(F_{1/32}) \in R^{6\times6\times D} \tag{3.12}$$

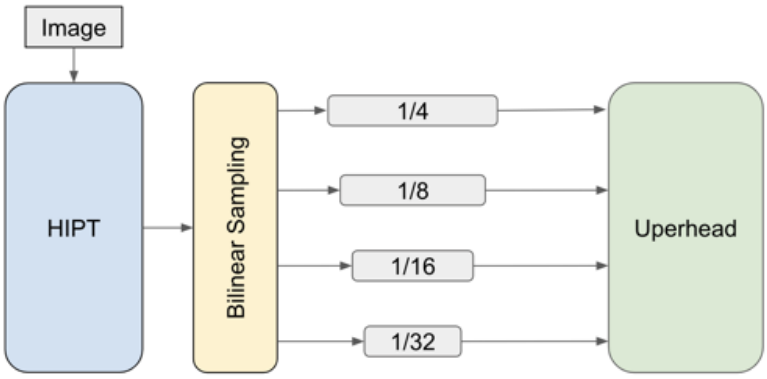
These feature maps are then upsampled to match the resolutions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ respectively. We then add these upsampled features with $F_{1/32}$, $F_{1/16}$, $F_{1/8}$, $F_{1/4}$ after a convolutional layer to obtain the next set of hierarchical feature maps $P_{1/32}$, $P_{1/16}$, $P_{1/8}$, $P_{1/4}$

$$P_{1/32} = \text{conv}(F_{1/32}) + \text{upsample}(f_{1\times1}) \tag{3.13}$$

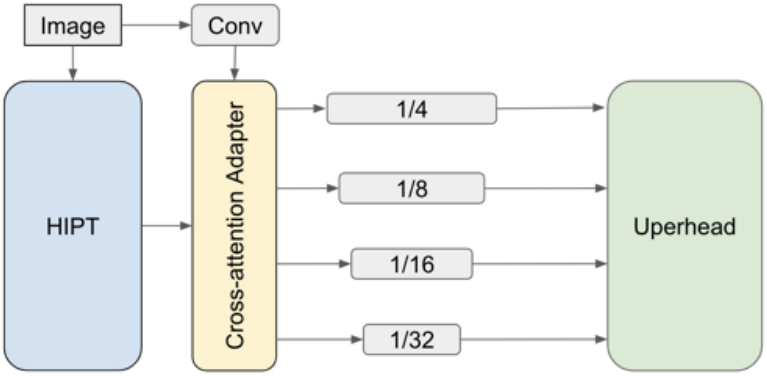
$$P_{1/16} = \text{conv}(F_{1/16}) + \text{upsample}(f_{2\times2}) + \text{upsample}(P_{1/32}) \tag{3.14}$$

$$P_{1/8} = \text{conv}(F_{1/8}) + \text{upsample}(f_{3\times3}) + \text{upsample}(P_{1/16}) \tag{3.15}$$

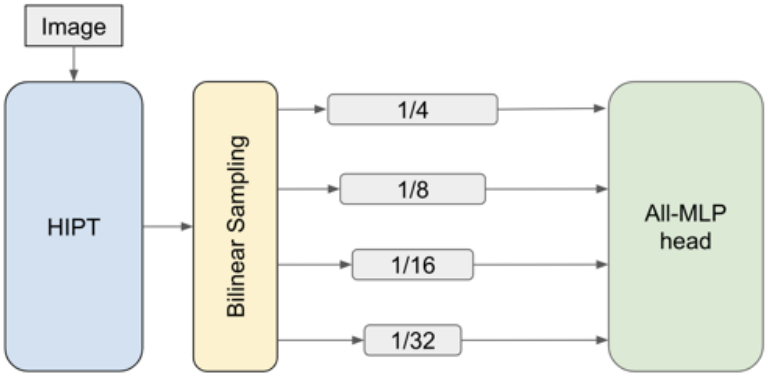
$$P_{1/4} = \text{conv}(F_{1/4}) + \text{upsample}(f_{6\times6}) + \text{upsample}(P_{1/8}) \tag{3.16}$$



(a) Baseline decoder.



(b) Adapter decoder.



(c) All-MLP decoder.

Figure 3.3: Three decoder types for HIPT models.

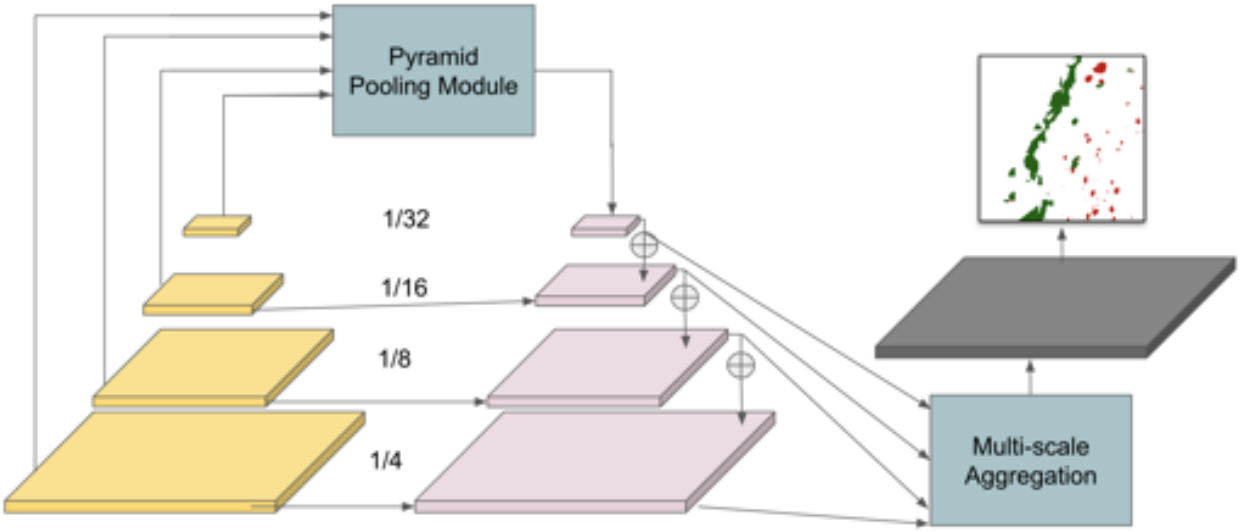


Figure 3.4: Uperhead decoder mechanism.

Finally, these features are all upsampled to $\frac{1}{4}$ resolution, aggregated by channel-wise concatenation, and then passed through the last convolutional layers to obtain the segmentation output.

$$P_{seg} = \text{conv}(\text{cat}(\text{upsample}(P_{1/32}, P_{1/16}, P_{1/8}, P_{1/4}))) \tag{3.17}$$

The final segmentation map at full resolution is then produced by bilinearly interpolating the quarter resolution output.

3.1.2.2 Adapter Decoder

The *adapter* decoder shown in Figure 3.3(b) is similar the baseline decoder from Section 3.1.2.1, except it has more intricate ways of constructing the hierarchical feature maps to be used for in the *Uperhead* module.

Since the original vision transformer architecture in HIPT is not hierarchical, we propose to not only extract but also inject hierarchical image features into the transformer network using the approach in [10]. As shown in Figure 3.3b, convolutional layers operating on the input image produce multi-scale spatial features $C_{1/4}$, $C_{1/8}$, $C_{1/16}$, $C_{1/32}$. We then reshape and concatenate $C_{1/8}$, $C_{1/16}$, and $C_{1/32}$ into a vector of size $(\frac{HW}{8^2} + \frac{HW}{16^2} + \frac{HW}{32^2}) \times D$. We denote this flattened vector of spatial features as $z_{0,sp}$, to be "injected" back into the HIPT backbone via cross-attention with the HIPT representation after the $lsupth$ layer, z_l, which is described below:

$$V = V_{\text{CA}} z_l \tag{3.18}$$

$$K = K_{\text{CA}} z_l \tag{3.19}$$

$$Q = Q_{\text{CA}} z_{l,sp} \tag{3.20}$$

$$\text{CA}(z_l, z_{l,sp}) = \text{softmax}(QK^T)V. \tag{3.21}$$

After computing cross-attention, we add this new vector back to the original HIPT representation z_l:

$$z'_l = z_l + \gamma \text{CA}(z_l, z_{l,\text{sp}}) \tag{3.22}$$

where γ controls the strength of the re-injected cross-attention features. z'_l is passed onto the next HIPT layer. For the *adapter* decoder, we also choose the every third layer of the HIPT backbone to extract hierarchical features, i.e. $l = 3, 6, 9, 12$.

Symmetrically, for every layer where we inject spatial features $z_{l,\text{sp}}$, we also extract features from the HIPT backbone and interact them with $z_{l,\text{sp}}$ via cross-attention. For the 4 layers where we extract the HIPT backbone features, we obtain the following:

$$\text{ConvFFN}(z) = \text{MLP}(\text{LayerNorm}(\text{reshape}(\text{conv}(z)))) \tag{3.23}$$

$$z'_{l+3,\text{sp}} = z_{l,\text{sp}} + \text{CA}(z_{l,\text{sp}}, z_l) \tag{3.24}$$

$$z_{l+3,\text{sp}} = z'_{l+3,\text{sp}} + \text{ConvFFN}(z'_{l+3,\text{sp}}) \tag{3.25}$$

After we extract the final cross-attention features z_{12} from the HIPT backbone, we divide the z_{12} of shape $(\frac{HW}{8^2} + \frac{HW}{16^2} + \frac{HW}{32^2}) \times D$ into three 2-D feature maps of the following sizes:

$$C'_{1/32} \in R^{\frac{HW}{32^2} \times D} \tag{3.26}$$

$$C'_{1/16} \in R^{\frac{HW}{16^2} \times D} \tag{3.27}$$

$$C'_{1/8} \in R^{\frac{HW}{8^2} \times D} \tag{3.28}$$

Finally, we construct multi-scale feature pyramid using these cross-attention features C and z'_l

$$F_{1/32} = C'_{1/32} + \text{downsample}(\text{reshape}(z'_{12})) \tag{3.29}$$

$$F_{1/16} = C'_{1/16} + \text{downsample}(\text{reshape}(z'_{9})) \tag{3.30}$$

$$F_{1/8} = C'_{1/8} + \text{downsample}(\text{reshape}(z'_{6})) \tag{3.31}$$

$$F_{1/4} = C_{1/4} + \text{downsample}(\text{reshape}(z'_{3})) \tag{3.32}$$

3.1.2.3 All-MLP Decoder

The *all-MLP* decoder shown in Figure 3.3(c) uses a more simplistic approach than the aforementioned decoders in Sections 3.1.2.1 and 3.1.2.2. It only utilizes MLP layers and upsampling operations to construct the segmentation map. We take the intermediate feature maps after the third, sixth, ninth, and twelfth layers of the transformer at spatial size $\frac{1}{16}$ of H and W. We again generate hierarchical features $F_{1/32}$, $F_{1/16}$, $F_{1/8}$, $F_{1/4}$ via bilinear resampling. First, we map all features onto a unified channel dimension C.

$$F'_{\frac{1}{32}} = FC_{D\to C}(F_{\frac{1}{32}}) \tag{3.33}$$

$$F'_{\frac{1}{16}} = FC_{D\to C}(F_{\frac{1}{16}}) \tag{3.34}$$

$$F'_{\frac{1}{8}} = FC_{D\to C}(F_{\frac{1}{8}}) \tag{3.35}$$

$$F'_{\frac{1}{4}} = FC_{D\to C}(F_{\frac{1}{4}}) \tag{3.36}$$

We then unify them by upsampling all F'_i to quarter resolution and then concatenating them.

$$F' = \text{cat}(\text{upsample}(F'_{\frac{1}{32}}, F'_{\frac{1}{16}}, F'_{\frac{1}{8}}, F'_{\frac{1}{4}})) \tag{3.37}$$

These features are then passed through 2 more fully connected layers.

$$seg = FC_{C\to Nclasses}(FC_{4C\to c})(F') \tag{3.38}$$

3.2 Segmentation Transformers Pretrained on Natural Images

In Section 3.1, we noted that it is important for segmentation transformers to have a hierarchical, multi-scale structure since it has proven to be effective for segmentation tasks. The HIPT [9] backbone was not originally designed for tasks such as segmentation, and we thus had to construct the decoder mechanisms to extract hierarchical features. We propose to directly utilize a hierarchical transformer backbone such as [33] rather than adapting other non-hierarchical pretrained models to perform segmentation.

SegFormer [33], shown in Figure 3.5, is an appropriate model for the melanoma segmentation task because it has the following characteristics. Unlike many other transformers, SegFormer has built-in hierarchical structures that produce multi-scale feature maps. It accomplishes this by merging patch tokens together several times to generate feature maps of different resolutions. It accomplishes this by using patch-embedding layers similar to the HIPT backbone in section 3.1.1. For example, after block i of the transformer block, the internal representations are of shape $h' \times w' \times D$. A patch embedding layer takes these internal representations and embeds overlapping 3×3 patches with a stride of 2, resulting in a

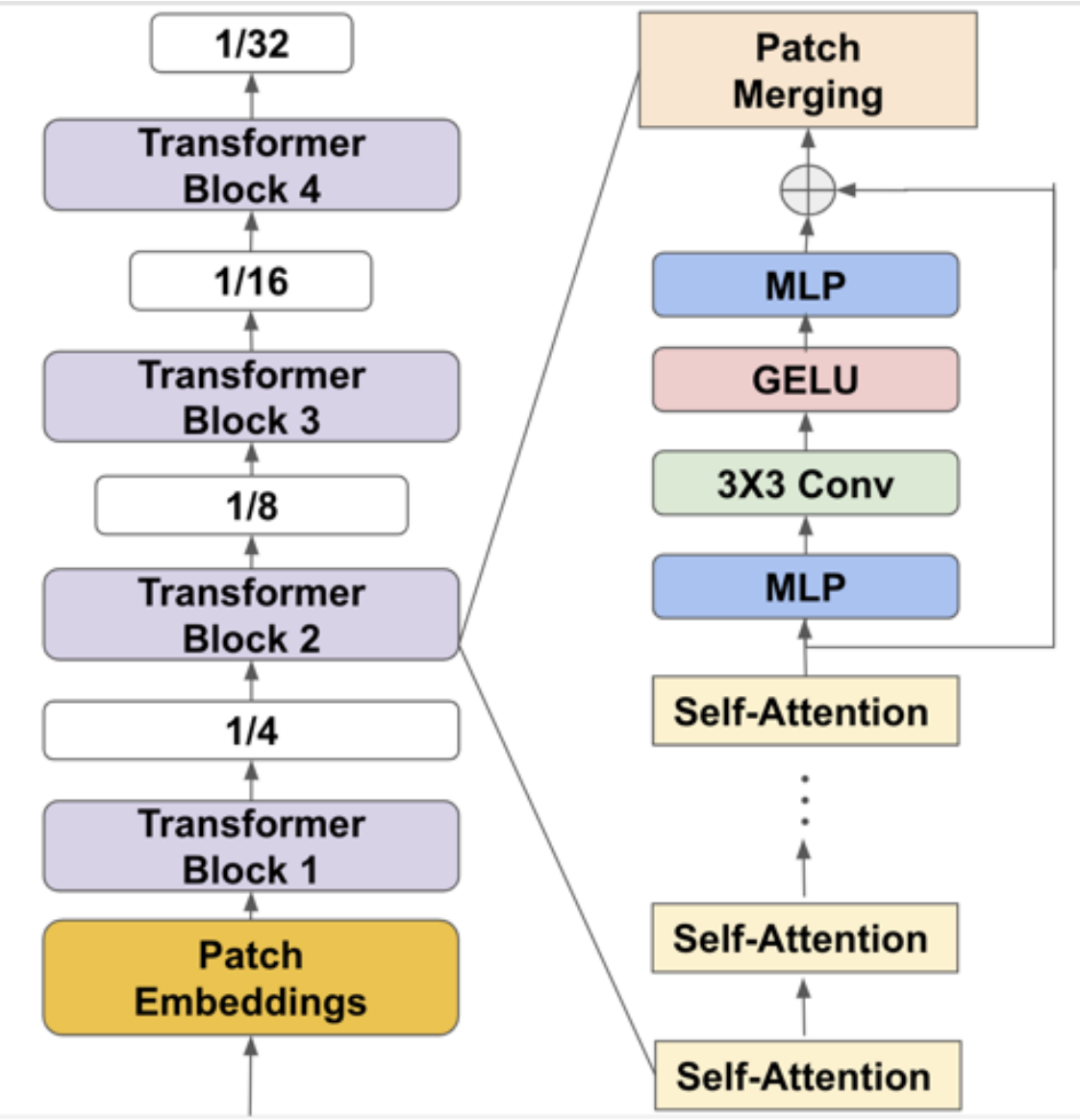


Figure 3.5: SegFormer backbone [33]. The backbone contains hierarchical feature maps produced through patch embedding operations. SegFormer also avoids positional encoding interpolation by using zero-padded convolutional layers to encode absolute position.

feature map of size $\frac{h'}{2} \times \frac{w'}{2} \times D$. Otherwise, the attention block operations in SegFormer are the same as the operations described in HIPT with Equations 3.7 and 3.4. The SegFormer models are pretrained on ImageNet-1K classification [33].

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 process allows the transformer to handle multiple resolutions, but also introduces artifacts in interpolation that usually lower performance. SegFormer [33] skips positional encodings altogether by using a convolutional layer to produce positional representations. Specifically, the convolutional layer zero-pads the spatial feature maps, and the zero-padding can be used as an implicit signal for absolute position in 2D-space [33]. SegFormer therefore has the ability to process data of different resolutions without introducing interpolation artifacts.

Chapter 4

Experimental Results

In this chapter, we describe experimental results from the methods described in Chapter 3. The outline of this chapter is as follows. In Section 4.1, we describe the data and data preprocessing steps in the experiments. In Sections 4.2 and 4.3, we then describe important metrics and implementation details of the experiments respectively. In the proceeding Sections 4.4 and 4.5, we present the experimental results of our two proposed methods in Chapter 3. In the last Section 4.6, we provide analysis and discussion of our results from both computer vision and pathologist perspectives. For the best performing model, we requested a certified dermatologist to provide their physician interpretation of the segmentation results. We include patch-based results of the physician-evaluated segmentation results in this section. We include results of the WSI segmentation in Appendix A.

4.1 Dataset and Preprocessing

Our dataset contains 55 total slide images with images as large as 10698×16846 pixels and as small as 3563×4021 pixels. The images are of skin biopsies stained with hematoxylin and eosin (H&E) [13] imaged under microscope at a magnification of 40x. 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.

In particular, we differentiate between these two stages of melanoma for two reasons: the volume of invasive melanoma is the most important prognosis factor for survival and in-situ melanoma is very low-risk. However, invasive melanoma and in-situ melanoma are visually similar, so distinguishing between these two stages of melanoma is difficult. The defining feature that separates these two stages of melanoma is their *context*, or in biological terms, what the surrounding tissues are. For invasive melanoma, the surrounding tissue is mostly dermis while for in-situ melanoma, the surrounding tissue is mostly epidermis. Therefore, correctly segmenting melanoma requires contextual understanding of the surrounding tissues.

class	training set	test set
background cells	78.06%	78.04%
fibrotic tumor	5.01%	11.38%
inflamed tumor	2.22%	3.13%
epidermis	4.27%	6.43%
uncertain tumor	0.02%	0.02%
invasive melanoma	10.41%	0.99%

Table 4.1: Class frequencies of each class in the training and test sets

The classes in our data are highly imbalanced. Table 4.1 shows that the incidence of invasive melanoma as a percentage of the total number of pixels in an image is quite small for both training and testing sets. To ameliorate this class imbalance problem, we apply weighted cross-entropy loss, where the loss-terms representing the loss from the minority classes are weighted higher. We also apply sampling based on minority classes [27]. We throw away patches that are less than 3% cellular tissue. We also oversample patches with more epidermis and more invasive melanoma. Specifically, if a patch has more than 25% melanoma and/or epidermis, we double the probability that the specific patch is sampled. We also undersample patches that are only healthy dermis tissue.

In addition, some of the classes in our dataset 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. We believe that segmenting these regions with noisy annotations is sub-optimal for learning geometric features. So for our models, we transform the data from the original 6 classes into 3 semantic classes to be learned: (a) *other*, which contains the *background cells*, *fibrotic tumor*; *inflamed tumor*; and *uncertain tumor*, (b) *invasive melanoma*, and (c) *epidermis*.

For data augmentations, we sample crops of whole-slide images at training-time. As discussed before, these sampled crops are filtered based on their semantic content, where we oversample patches with more epidermis and melanoma and undersample patches with too many semantic classes belonging to *other*. The larger slide images are sampled into $N \times N$ square patches, where $N = 512, 1024$. We also apply horizontal flipping, vertical flipping, and discrete rotations at 90°, 180°, 270°, or 360°. We avoid photometric augmentations, as they may create photometric changes which would not follow the properties of H & E stained slides [13].

We visualize all 13 images in the test set, in addition to the best predictions by mIoU for the best HIPT and SegFormer models in Appendix A.

4.2 Metrics

We train our segmentation model with pixel-wise weighted cross-entropy loss for 3 of the aforementioned classes:

$$-\frac{1}{H \times W} \sum_{i,j}^{H,W} \sum_{c=1}^{3} w_c {y_c}^{i,j} \log {p_c}^{i,j} \tag{4.1}$$

where i, j refers to the 2D position of a pixel, ${p_c}^{i,j}$ is the predicted probability of a specific class at a pixel, and ${y_c}^{i,j}$ is the one-hot encoded label of the pixel. The class weights w_c, are set to $w_{other} = 1$, $w_{melanoma} = 5$, $w_{epidermis} = 5$.

We use mean intersection over union (mIoU) as our primary evaluation metric. The mIoU is the average of the IoUs on a per-class basis. More specifically, let P be the set of all pixels of predicted for class i, and let G be the set of all ground truth pixels for class i. Then the intersection over union is defined as:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{4.2}$$

Of all the individual class IoU's, the intersection over union of the melanoma class is the most relevant since that represents the accuracy of the melanoma segmentation problem.

4.3 Implementation Details

For all HIPT models, we set the embedding patch size $P = 16$, the embedding size $D = 384$, and the number of blocks $L = 12$. For the *all-MLP* decoder, we set the hidden MLP dimension $C = 256$. For all SegFormer models, we set the embedding patch size $P = 4$ and the hidden MLP dimension $C = 256, 768$. The values of important parameters such as the patch embedding size, hidden MLP dimensions, and dimensions of the MLP decoders are listed in Table 4.2. We train our models on a machine with 3 NVIDIA Quadro RTX 8000 GPUs with PyTorch. We use the Adam optimizer [18] with a learning rate of 0.00006 and with a weight decay of 0.01. We use the commonly used linear decay learning rate scheduler with linear warmup. We apply dropout on the final segmentation head layer and also the positional embeddings for the HIPT models. Both models take approximately 1 day to train 100 epochs with a batch size of 16 per GPU.

Model	Patch Embedding Size P	Hidden MLP Dimension	MLP Decoder Dimension
HIPT	16×16	384	256
SegFormer	4×4	32, 64, 128, 320, or 512	256 or 768

Table 4.2: Table of model parameter values.

4.4 HIPT Models

In this section, we describe the experiments for HIPT models. In the following three subsections, we present results on different decoder designs, patch size, and network initializations.

4.4.1 Decoder Design

We conduct experiments on decoder mechanisms for the HIPT backbone models described in Section 3.1.2. As seen in Table 4.3, the best performing 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 more sophisticated mechanisms to construct the multi-scale feature maps from the HIPT backbone. The *baseline* architecture outperforms the *all-MLP* architecture, which may be due to MLP architectures needing more data to effectively generalize and select hyperparameters.

Model	Params	mIoU	melanoma IoU	F1 Score
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 4.3: Best results on using different decoders for HIPT. The resolution used for this experiment was 512×512.

4.4.2 Patch Size

We report the effects of using different patch sizes of our dataset on our best models in Table 4.4. The patch size is an important part of the segmentation problem since it determines the amount of contextual information for a given patch size for training and testing. We apply a shift-and-stitch reconstruction mechanism based on [29], with a shift-size of 128. We also weight the predictions by a Gaussian with a peak at the center of the patch, as pixels that are farther from the center are less likely to be accurate since they have less context.

As seen in Table 4.4, generally, smaller patch sizes result in higher performance. This is because the HIPT backbone was pretrained on images of size 256×256. Therefore, the positional encoding needs to be interpolated since the new segmentation task has a different resolution. We also experimented with using new positional embeddings of the same resolution as the input segmentation images, but found this results in a negligible difference. In addition for larger patch sizes, our GPU memory limitations prohibited us from training on large batch sizes, which might have lowered performance.

Model	Resolution	mIoU	melanoma IoU	F1 Score
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 4.4: Results on patch sizes for HIPT.

4.4.3 Initialization Experiments

We conduct experiments on finetuning of the segmentation task with different model initializations. As shown in Table 4.5, a pretrained transformer backbone is absolutely necessary to perform well on the segmentation task. Our small segmentation dataset does not have enough data to learn more general visual representations. As seen in the third row of Table 4.5, transformer models suffer from performance degradation without pretraining [12]. We see that the student and teacher models in the the first two rows of Table 4.5 perform similarly for the invasive melanoma segmentation task. Lastly, we see that the teacher model performs slightly better the student model across all metrics. This is consistent with [9], where the authors use the teacher model for finetuning tasks.

Model	Model Initialization	mIoU	melanoma IoU	F1 Score
Adapter	Student	0.679	0.367	0.534
Adapter	**Teacher**	**0.696**	**0.401**	**0.573**
Adapter	None	0.558	0.131	0.231

Table 4.5: Results on using network initialization experiments in identical settings

4.5 SegFormer Models

In this section, we describe the experiments for SegFormer models. In the following three subsections, we present results on different decoder designs, patch sizes, and network initializations.

4.5.1 Model Sizes

Table 4.6 shows experiments with three SegFormer [33] sizes: B0, B1, and B2. Table As seen, there is a general performance boost with ascending sizes. However, finetuning the large SegFormer B3 and B4 models resulted in almost zero melanoma IoU. The best model

in the entirety of this thesis is the SegFormer B2 model. SegFormer is able to achieve state-of-the-art performance on our dataset with relatively small models.

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

Table 4.6: Results on different SegFormer sizes

4.5.2 Patch Size

We conduct experiments with different patch sizes, reporting the results in Table 4.7. Larger models generally perform better, but the same trends in size do not exactly hold for patch resolution. We see that for SegFormer B0, patch sizes of 512 and 1024 have almost the same performance with a dramatic reduction in performance for a patch size of 1536. We see that for SegFormer B0, there is a dramatic reduction in performance for a patch size of 1536 compared to 512 and 1024

However for SegFormer B1, having a larger patch size of 1024 is beneficial. We notice the inverse trend for SegFormer B2, with the smallest patch size of 512 having the best performance. Overall, there are no clear trends with patch size and performance even though larger contexts naturally contain more information for segmentation.

4.5.3 Initialization Experiments

We conduct experiments on finetuning with different model initializations for SegFormer. We collect results for SegFormer B0 with no pretraining and two other different configurations shown in Table 4.8. The first configuration shown in the first row of Table 4.8 is an encoder pretrained on ImageNet classification [11]. The second configuration shown in the second row of Table 4.8 corresponds to an encoder pretrained on ImageNet classification and then the entire model, encoder and decoder, finetuned for segmentation on the ADE20K dataset [37]. We report the initialization experiments for SegFormer B0. From our experiments, pretraining the entire model results in the best performance and performs slightly better than only training the encoder alone. No pretraining results in by far the worst performance.

4.6 Comparisons and Discussion

In this section, we analyze our experimental results and visualizations from a computer vision and also a certified dermatologist's perspective.

Model	Resolution	mIoU	melanoma IoU	F1 Score
SegFormer B0	512	0.694	0.397	0.568
SegFormer B0	1024	0.695	0.398	0.569
SegFormer B0	1536	0.573	0.155	0.269
SegFormer B1	512	0.689	0.386	0.557
SegFormer B1	1024	0.717	0.441	0.613
textbfSegFormer B2	**512**	**0.719**	**0.447**	**0.618**
SegFormer B2	1024	0.708	0.424	0.595

Table 4.7: Results on different patch sizes for SegFormer.

Model Initialization	mIoU	melanoma IoU	F1 Score
Encoder Pretrained	0.689	0.386	0.557
Whole Model Pretrained	**0.694**	**0.397**	**0.569**
None	0.574	0.163	0.280

Table 4.8: Results on using teacher versus student pretrained networks in identical settings for SegFormer.

Model	mIoU	melanoma IoU	F1 Score
Multi-Scale FCN [27]	0.538	0.130	0.140
Best 2-stage [24]	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 4.9: Comparison of our proposed approach and previous approaches.

4.6.1 Computer Vision Perspective

Table 4.9 shows the best results for each type of transformer model and also the best results of the existing methods in [24] and [27]. The best SegFormer and HIPT models outperform the best model from [24] in mIoU by 12% and 9% respectively. This is because of two likely reasons. First, convolutional networks do not model global long-range contexts because of the receptive field problem. Transformers have a receptive field that 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 small, sparse, and can be present across long ranges in an image sample. Qualitatively, we see that transformer-based architectures fare much better in this long-range modelling task for scattered melanoma from Figures 4.1, 4.2, and 4.3. The second reason is that transformers exhibit superior generalization ability because they lack the inductive biases in convolutional networks. Therefore, transformers pretrained on large

Test Image No.	SegFormer mIoU	SegFormer melanoma IoU	HIPT mIoU	HIPT melanoma IoU	Qualitative Comparisons	Which Model Better Qualitatively?
0 (Fig. A.1)	**0.683**	**0.381**	0.650	0.313	Melanoma results are similar but epidermis is more accurate in SegFormer.	≈
1 (Fig. A.2)	0.519	0.053	**0.629**	**0.266**	HIPT model captures melanoma more accurately.	H
2 (Fig. A.3)	**0.757**	**0.519**	0.741	0.486	HIPT and SegFormer have comparable performance.	≈
3 (Fig. A.4)	**0.711**	**0.427**	0.647	0.299	Both HIPT and SegFormer models miss some small-scattered melanoma.	S
4 (Fig. A.5)	**0.601**	**0.205**	0.590	0.182	Both HIPT and SegFormer models miss some small-scattered melanoma.	≈
5 (Fig. A.6)	**0.667**	**0.338**	0.544	0.091	HIPT misses more small and scattered melanoma compared to SegFormer.	S
6 (Fig. A.7)	**0.596**	**0.194**	0.565	0.131	HIPT and SegFormer have comparable performance.	≈
7 (Fig. A.8)	**0.694**	**0.391**	0.663	0.328	HIPT and SegFormer have comparable performance.	S
8 (Fig. A.9)	**0.703**	**0.416**	0.570	0.150	HIPT models have more false negatives and miss some scattered melanoma.	S
9 (Fig. A.10)	0.575	0.155	**0.589**	**0.181**	HIPT slightly outperforms SegFormer.	H
10 (Fig. A.11)	**0.829**	**0.662**	0.753	0.511	SegFormer slightly outperforms HIPT in identifying small pieces of melanoma.	S
11 (Fig. A.12)	**0.769**	**0.559**	0.728	0.481	HIPT and SegFormer have comparable performance.	≈
12 (Fig. A.13)	**0.846**	**0.696**	0.696	0.403	SegFormer captures a large piece of melanoma more accurately.	S

Table 4.10: Performance comparisons on individual test samples between HIPT and SegFormer models. In the rightmost column, H stands for HIPT, S stands for SegFormer, and ≈ stands for similar when assessing which model performed better qualitatively. Visualizations of the whole-slide segmentation results for each sample can be found in Appendix A.

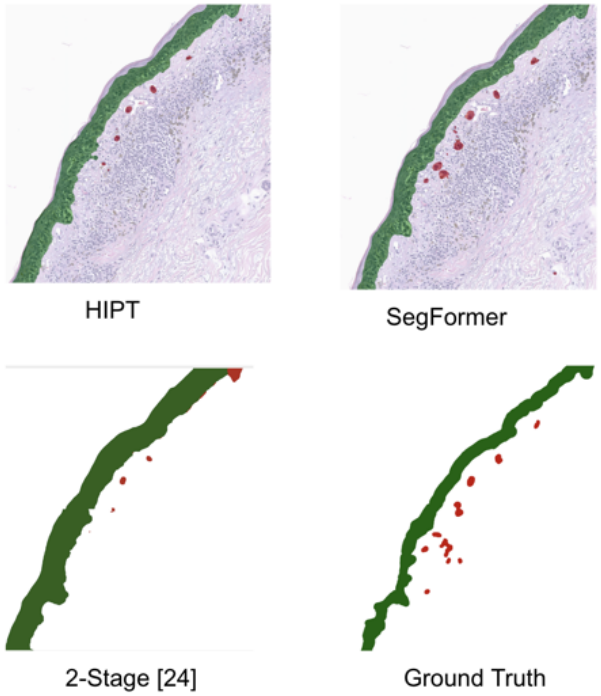


Figure 4.1: The method from [24] misses far more scattered melanoma. SegFormer is the closest to the ground truth.

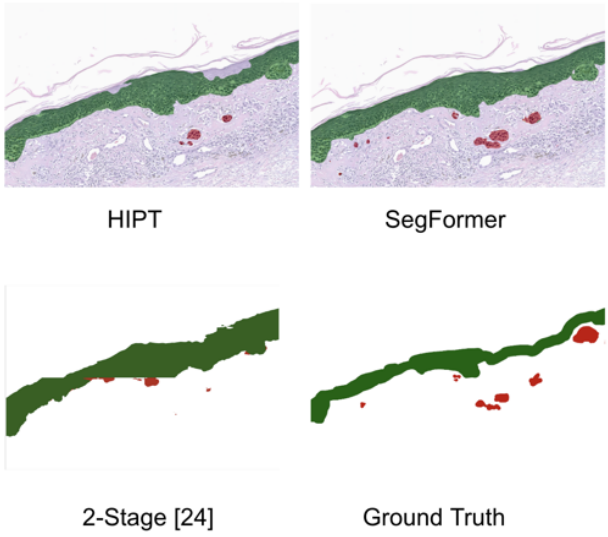


Figure 4.2: The method from [24] segments the epidermis poorly and contains lots of false positives and false negatives for melanoma. SegFormer is closest to the ground truth.

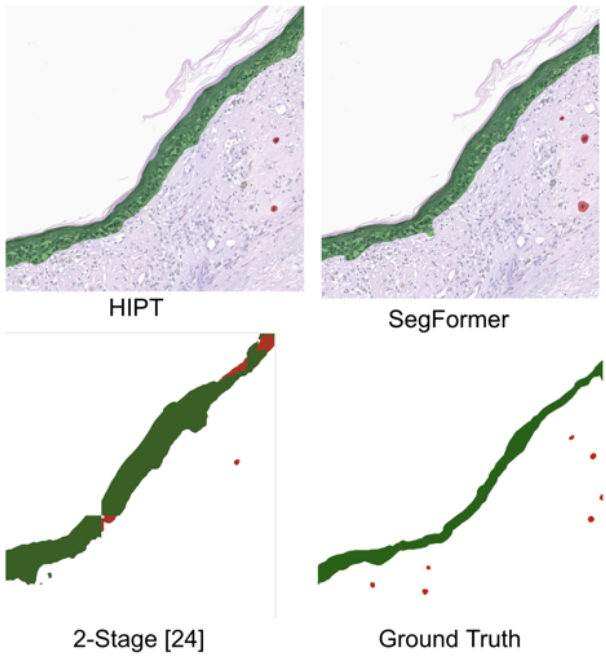


Figure 4.3: The method from [24] fails to segment the scattered melanoma and also contains artifacts at the edges of the epidermis. SegFormer is closest to the ground truth

datasets such as HIPT and SegFormer can thus generalize better than vanilla convolutional networks.

The best model from [24] contains 80 million parameters, which is significantly more than the 58.1 million and 27.5 million in the best HIPT and SegFormer models. Transformer models achieve superior performance with fewer parameters, confirming the models from [24] are overparameterized. We note the problem of class imbalance and overfitting, especially for the *invasive melanoma* and *epidermis* classes. Since the majority of the pixels in thetraining and testing sets are not invasive melanoma, there is a suboptimal amount of melanoma annotations in the training set, leading to overfitting for the melanoma class. We observe around $0.90 - 0.95$ melanoma training IoU near the end of training, whereas the highest achieved melanoma testing IoU is 0.44. While we ameliorate this class-imbalance and overfitting problem with weighted cross entropy loss and sampling for minority classes during train time, the problem still affects the training process and thus testing results.

SegFormer quantitatively performs better than HIPT in 11 of 13 test samples according to Table 4.10, resulting in an overall 0.02 higher mIoU for the whole test set. 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. From the qualitative observations along with the quantitative metrics such as melanoma IoU in Table

4.10, HIPT suffers from performance degradation in a specific setting. In particular, we notice that the segmentation maps by HIPT have more trouble detecting sparse and small melanoma, which may indicate that the internal representations have too low of a resolution. Figures A.7, A.9, and A.11 show that the HIPT models do not detect some of the small and scattered melanoma that SegFormer is able to. However, we also note that both models miss small and scattered melanoma in Figures A.4 and A.5. In one of the samples in Figure A.13, SegFormer captures a large piece of melanoma more accurately as well, which is further evidence that SegFormer's architecture allows for better multi-scale modelling. The HIPT models outperform the SegFormer models in Figure A.2. In general the qualitative performance between the two models is very similar in Figures A.1, A.3, A.7, A.8, A.10, and A.12,. The largest qualitative differences between the two models are mostly due to small and scattered melanoma.

We speculate that supervised pretraining on ImageNet has comparable quantitative effects to self-supervised pretraining on breast cancer slide images in HIPT [9] due to both settings performing substantially better than no pretraining, as seen in Tables 4.5 and 4.8. A possible reason for SegFormer to have outperformed the custom-designed HIPT models is the problem of positional encoding. HIPT was pretrained on 256×256 images at $20\times$ magnification. Therefore, to accommodate our dataset of $40\times$ images at higher resolutions, there is a mismatch in positional encodings which results in performance decrease. SegFormer on the other hand does not use positional encodings in there models, and thus resolution is not as important of a factor.

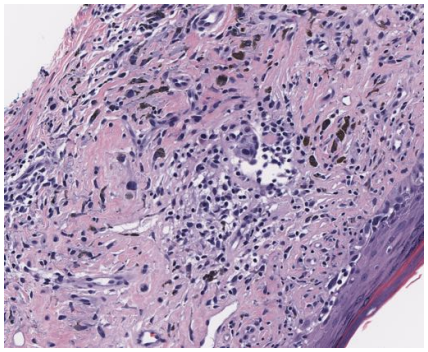
Lastly, we note that larger patch sizes in the case of HIPT tend to worsen performance due to positional encoding interpolation, while for the SegFormer models it shows inconclusive results for several SegFormer sizes. We also note that for SegFormer experiments, we only tuned hyperparameters for a single resolution and then directly used those hyperparameters for other resolutions. More experimentation with other patch sizes and more hyperparameter tuning is needed to reach more definitive conclusions.

4.6.2 Interpretations by Dermatologist

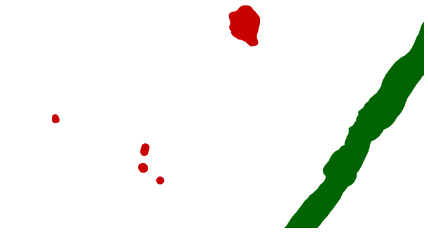
We include physician interpretations and quality assessment of our segmentation results from a certified dermatologist on selected regions of the test dataset. We provide sample images with the physician commentary in the caption of Figures 4.4 through 4.14. Overall, the model performs pretty well for prognostic standards, and some of the mistakes can be attributed to the visual similarities between in-situ melanoma and invasive melanoma such as in Figures 4.6 and 4.7, which is also hard for trained dermatologists. Other false positives in the predictions can be attributed to tissues such as sweat glands and inflammatory cells in Figures 4.4 and 4.5. False negatives in the predictions are primarily due to melanoma being inside very fibrotic regions in Figures 4.8 and 4.9, which is challenging for physicians to identify.

We note that the quantitative metrics such as mIoU are possibly not a good representation of model performance for two reasons. First, while dermatologists can generally agree

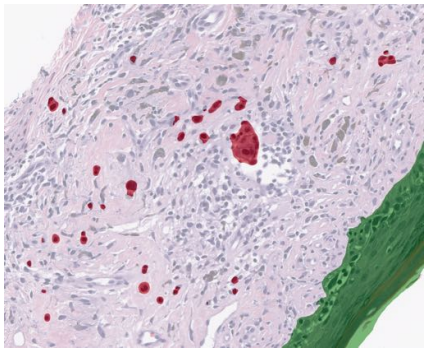
on which larger regions of a sample contain melanoma, the precise pixel annotations vary among physicians with more disagreements in labelling finer regions. Due to this physician variability in pixel-perfect ground truth, mIoU is a noisy representation of actual model performance. There tends to be higher inter-observer variability especially in small and scattered melanoma or melanoma in inflamed and fibrotic regions, as shown in Figure 4.12. The second reason is that because of the extremely high resolutions of slide images, annotators did not have time to rigorously label every single pixel. In fact, in some of the samples, the model predictions are actually more accurate than the provided ground truth such as in Figure 4.10. In fact, the SegFormer model is able to predict small sections of melanoma that annotators originally missed in Figure 4.14 and sections of epidermis that annotators originally skipped labelling in Figure 4.13. Many of the errors such as in Figure 4.11 do not actually affect the prognostic value of our model predictions from a physician standpoint. Overall, in most cases our model achieves performance that would be informative enough to use as a prognostic tool.



(a) Image patch

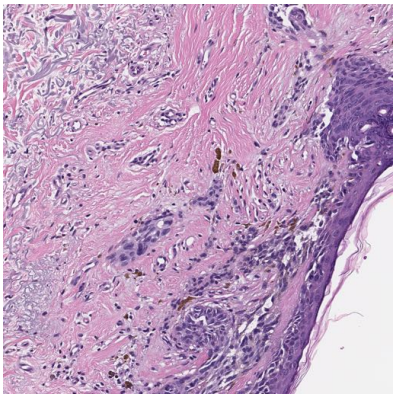


(b) Ground truth



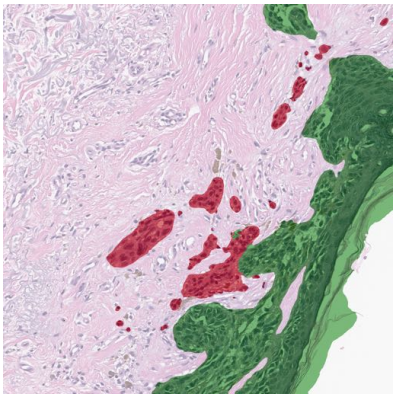
(c) SegFormer model prediction

Figure 4.4: The inflammatory cells in this sample cause the model to predict false positives. However even with the scattered false positives, this model prediction is good enough for clinical use



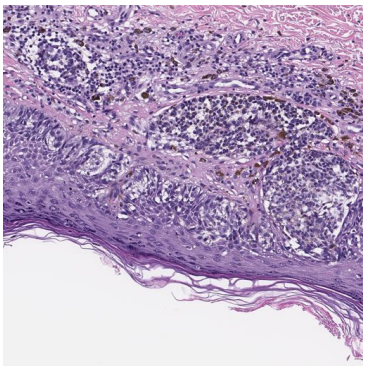
(a) Image patch

(b) Ground truth

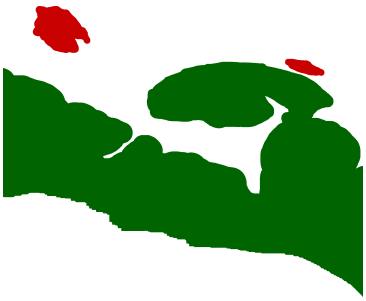


(c) SegFormer model prediction

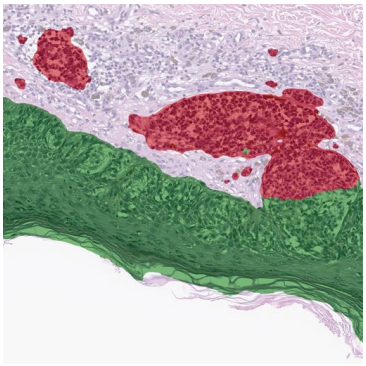
Figure 4.5: The eccrine (sweat) glands in this sample cause the model to predict false positives.



(a) Image patch

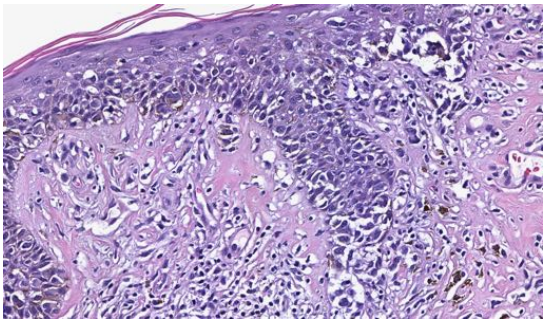


(b) Ground truth

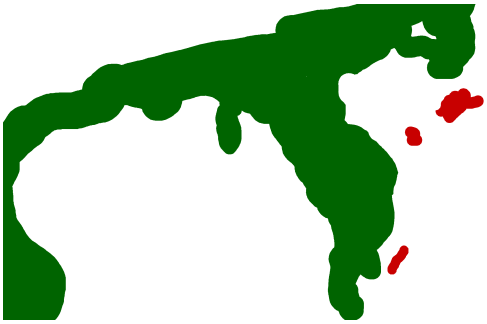


(c) SegFormer model prediction

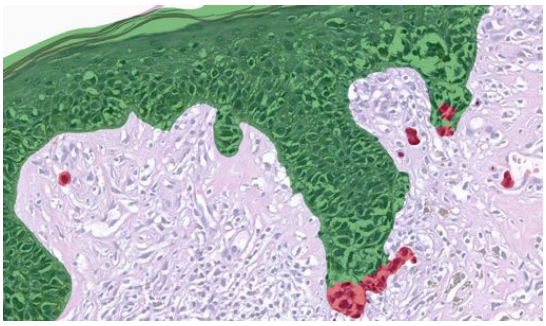
Figure 4.6: The small, dotted false positives in this sample are from capturing histiocytes (immune cells). The dominant false positive in this patch is from confusing in-situ melanoma with invasive melanoma. The false positives in this case are clinically acceptable.



(a) Image patch

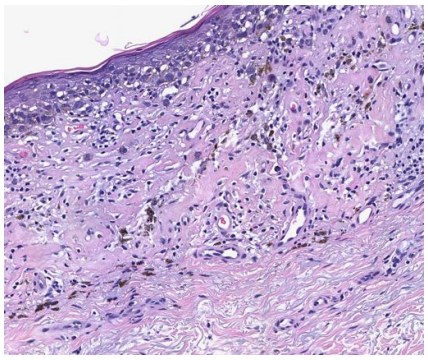


(b) Ground truth



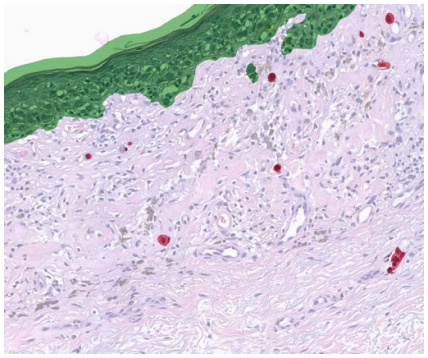
(c) SegFormer model prediction

Figure 4.7: The model impressively detects individual melanoma cells, which was actually missing from the annotations. The dominant false positive in this patch is from confusing in-situ melanoma with invasive melanoma. The false positives in this case are clinically acceptable.



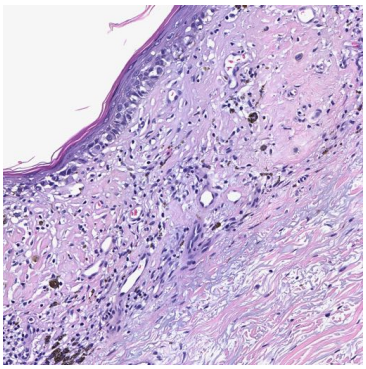
(a) Image patch

(b) Ground truth



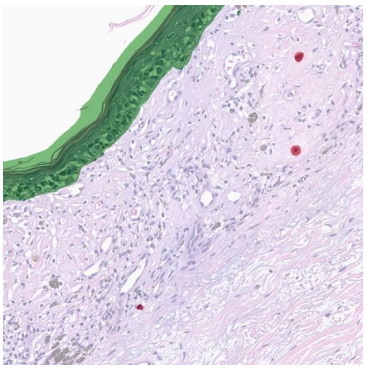
(c) SegFormer model prediction

Figure 4.8: The fibrotic nature of this sample obfuscates the melanoma, making the model predict some false negatives. This sample is difficult for physicians to annotate, so it is very impressive that the model identified individual melanoma cells.



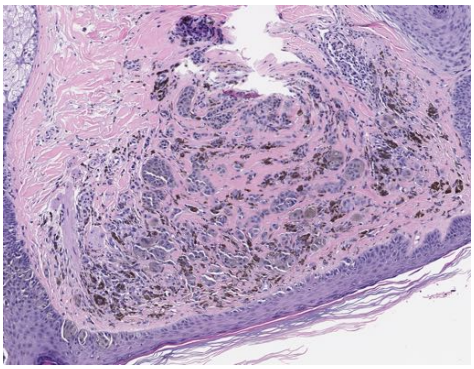
(a) Image patch

(b) Ground truth



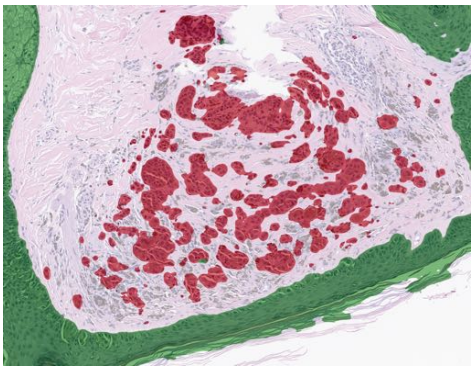
(c) SegFormer model prediction

Figure 4.9: The observations of this patch are similar to the observations in 4.8



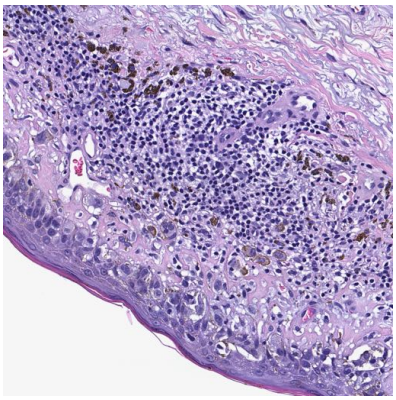
(a) Image patch

(b) Ground truth

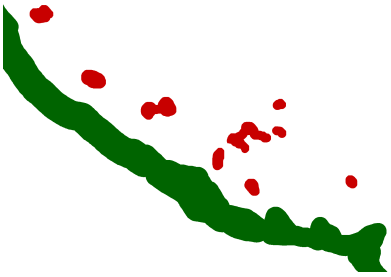


(c) SegFormer model prediction

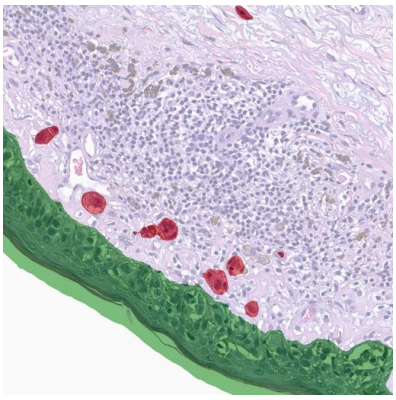
Figure 4.10: The invasive melanoma predictions are actually more accurate than the human-annotated ground truth. This is because the human-annotated ground truth is not perfectly precise due to time constraints, and metrics calculated with the ground truth undersell the performance.



(a) Image patch

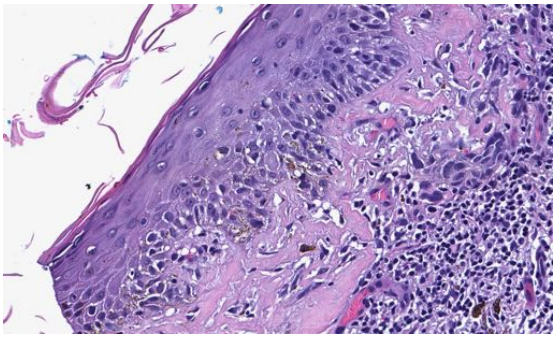


(b) Ground truth

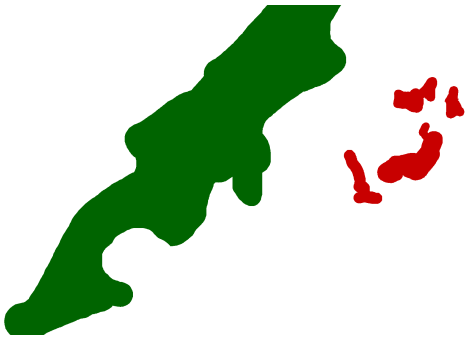


(c) SegFormer model prediction

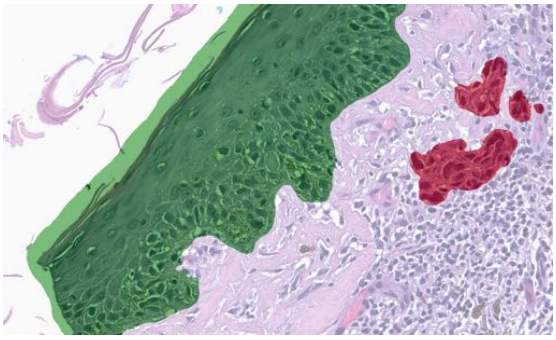
Figure 4.11: Even though there are several false negatives in this sample, the model predictions are generally so good that these errors are acceptable to use for physician prognosis.



(a) Image patch

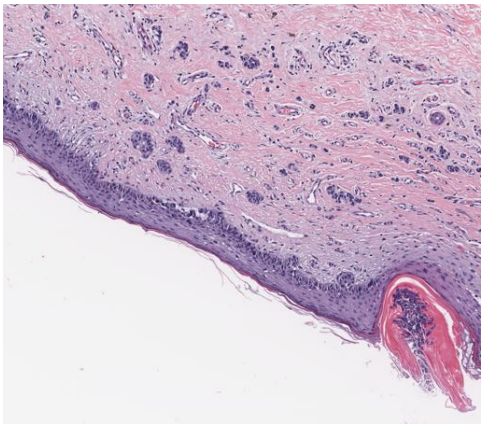


(b) Ground truth



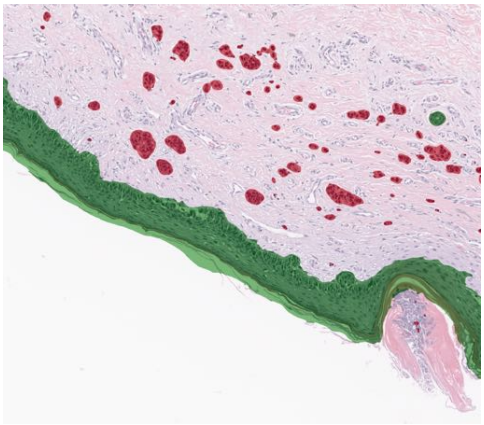
(c) SegFormer model prediction

Figure 4.12: The prediction is very accurate, and contained some tissue in between the clusters in the ground truth. This is not necessarily a mistake by the model, as there can be significant physician annotator variability with this sample.



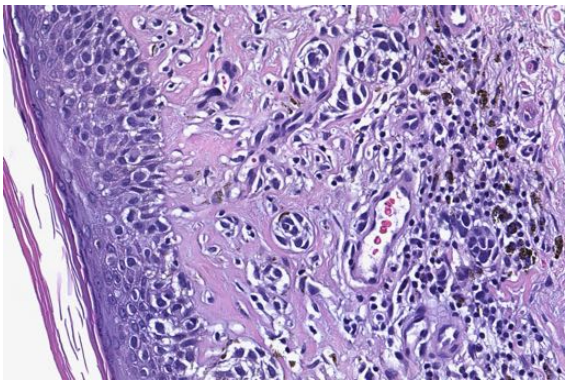
(a) Image patch

(b) Ground truth



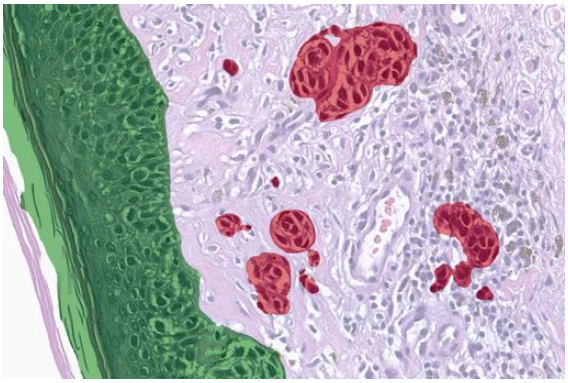
(c) SegFormer model prediction

Figure 4.13: The prediction is very accurate, and the model even detects some unannotated epidermis.



(a) Image patch

(b) Ground truth



(c) SegFormer model prediction

Figure 4.14: In the upper right corner of this sample, the model detected a small piece of melanoma that was unannotated.

Chapter 5

Conclusions and Future Work

We proposed two transformer-based methods which offer a significant improvement over the previous state-of-the-art method in [24]. We note that SegFormer models slightly outperform HIPT models due to the inherent multi-scale architectural design of SegFormer. We show that both transformer-based methods offer superior performance in the segmentation task compared to previous work with convolutional backbones [24] with less overall training time and memory. We also show that pretrained transformers are absolutely necessary to perform well on segmentation tasks. We also show that using more sophisticated methods to construct multi-scale features such as in HIPT *adapter* models result in superior segmentation performance. A board-certified dermatologist concludes that our segmentation models perform well on the majority of areas on the WSIs, and that models can exceed human performance in some regions as well.

Future work can focus on class imbalances. For example, pixel-wise cross-entropy loss can be a poor choice for class-imbalanced problems because minority classes tend to have fewer pixels and therefore fewer loss terms. Other metrics which do not behave this way can offer a better alternative towards addressing the problem of class imbalance

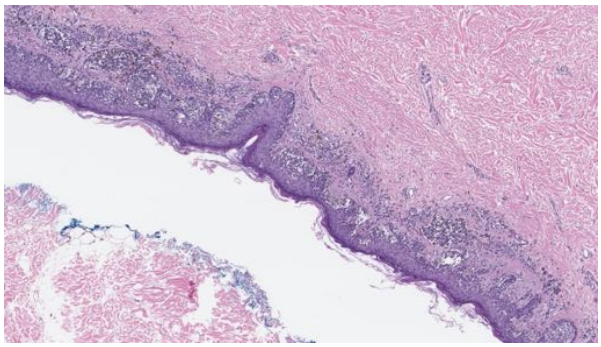
Another problem in our dataset is the lack of true pixel-perfect annotations, since our annotators took some small shortcuts due to extremely high resolution of images. Therefore methods in deep learning to deal with noisy annotations would be useful to incorporate in our training process.

Lastly, there are other architectures and paradigms for segmentation tasks that could offer performance boosts as well. Panoptic or mask prediction segmentation architectures and other hybrid convolutional-transformer architectures can be useful to investigate further.

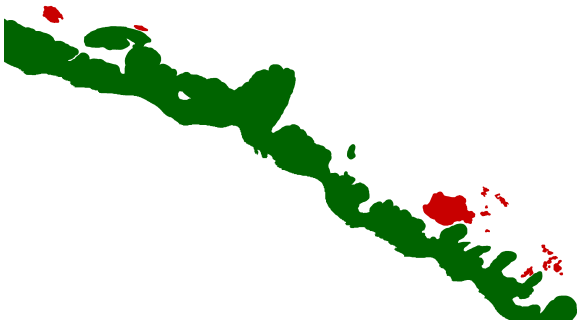
Appendix A

Whole-Slide Segmentation Results

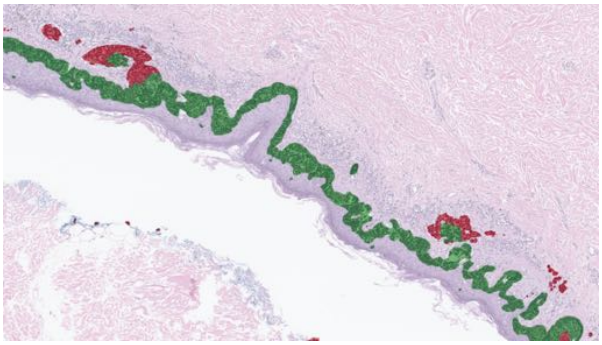
The following figures (Figure A.1-A.13) visualize whole-slide segmentation results on our test set for our HIPT models and SegFormer models. 4.10 shows quantitative results for each individual sample.



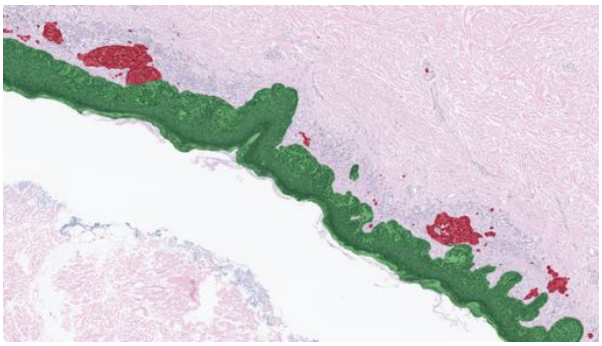
(a) Image



(b) Ground truth

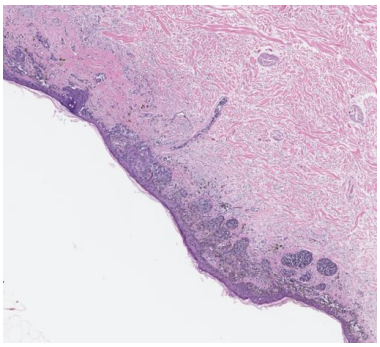


(c) HIPT model prediction

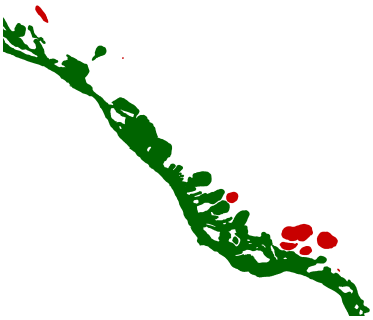


(d) SegFormer model prediction

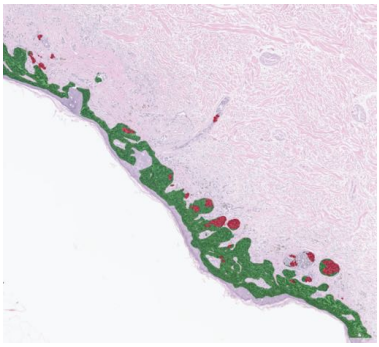
Figure A.1: Test image 0. Both models perform similarly qualitatively.



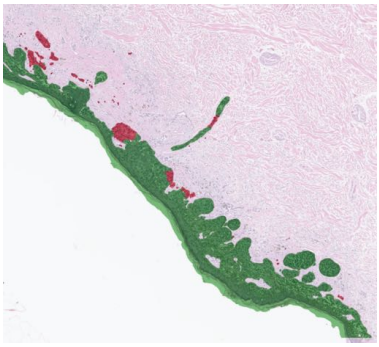
(a) Image



(b) Ground truth

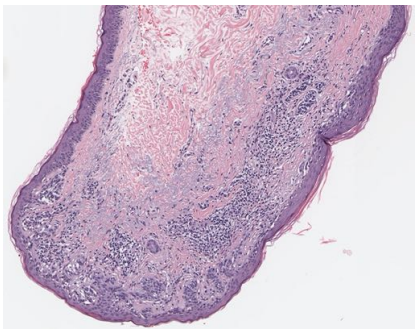


(c) HIPT model prediction

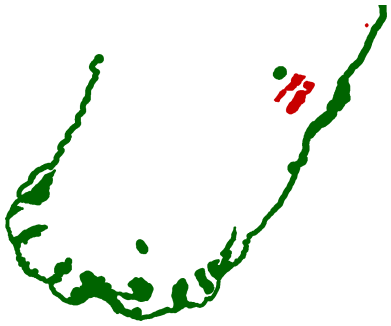


(d) SegFormer model prediction

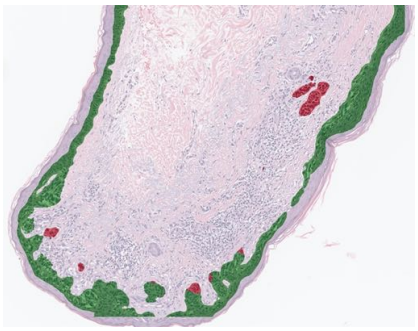
Figure A.2: Test image 1. HIPT performs better than SegFormer.



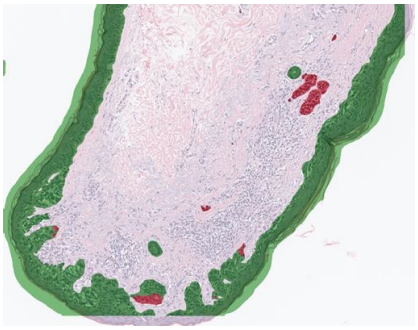
(a) Image



(b) Ground truth

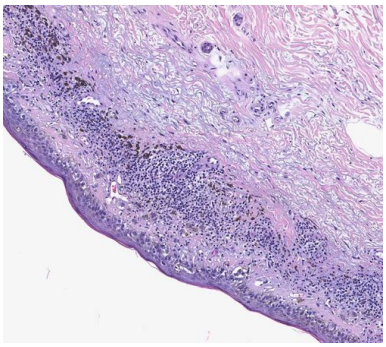


(c) HIPT model prediction

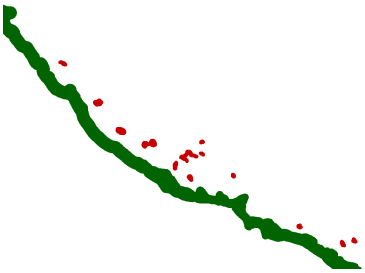


(d) SegFormer model prediction

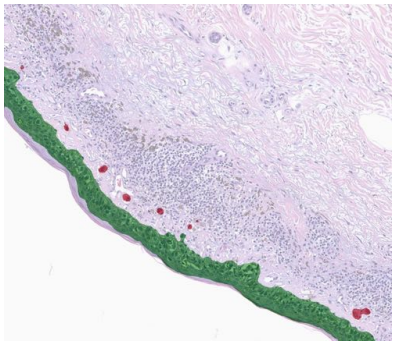
Figure A.3: Test image 2. Both models perform similarly.



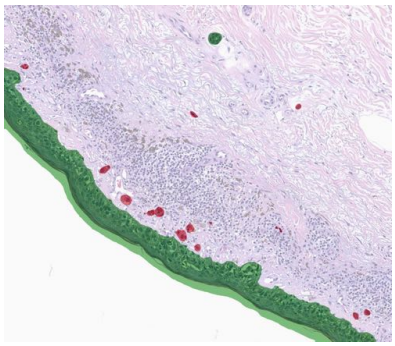
(a) Image



(b) Ground truth

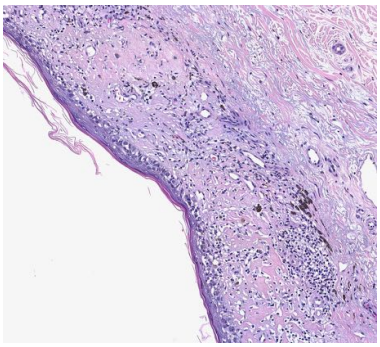


(c) HIPT model prediction

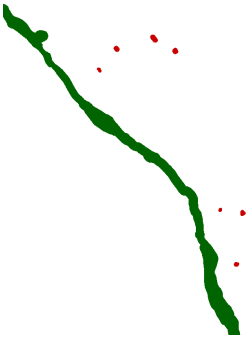


(d) SegFormer model prediction

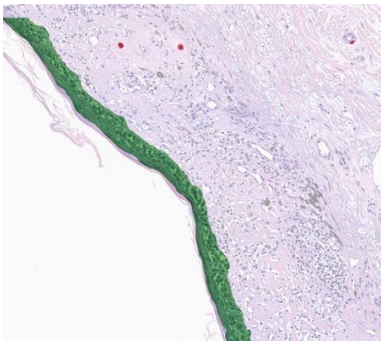
Figure A.4: Test image 3. SegFormer performs better than HIPT.



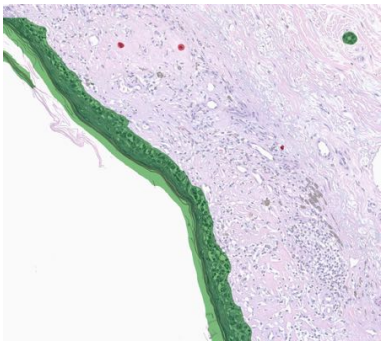
(a) Image



(b) Ground truth

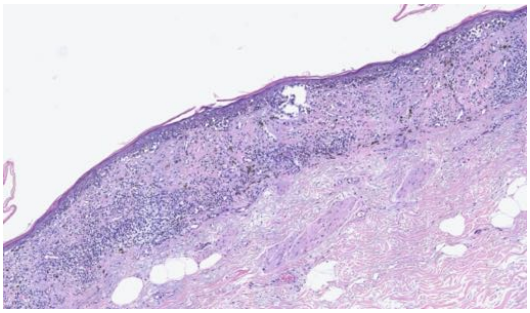


(c) HIPT model prediction



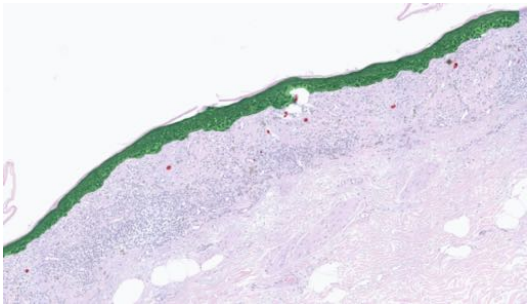
(d) SegFormer model prediction

Figure A.5: Test image 4. Both models perform similarly.

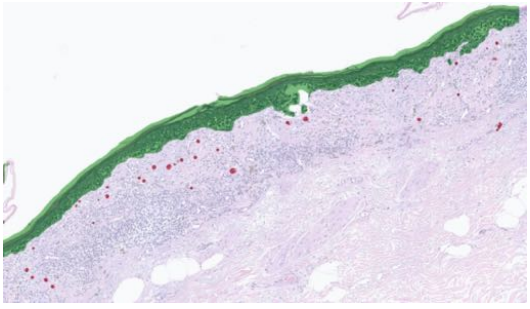


(a) Image

(b) Ground truth

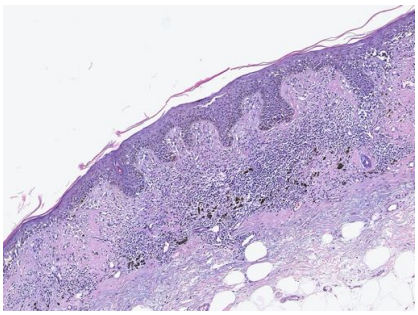


(c) HIPT model prediction

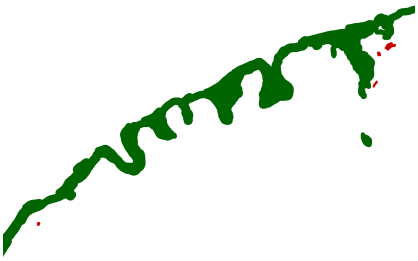


(d) SegFormer model prediction

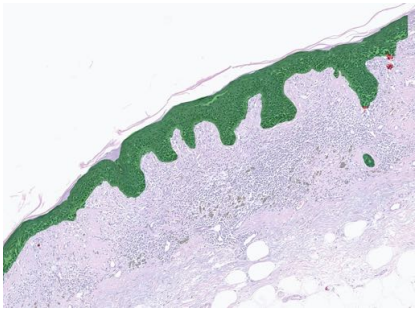
Figure A.6: Test image 5. SegFormer performs better than HIPT.



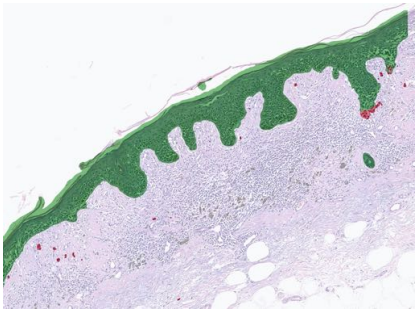
(a) Image



(b) Ground truth

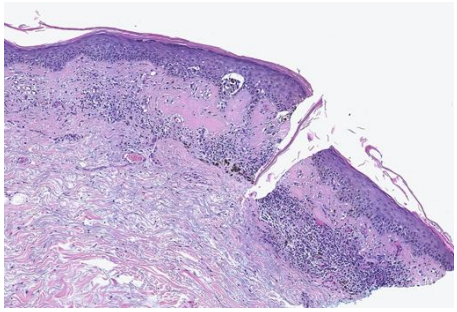


(c) HIPT model prediction



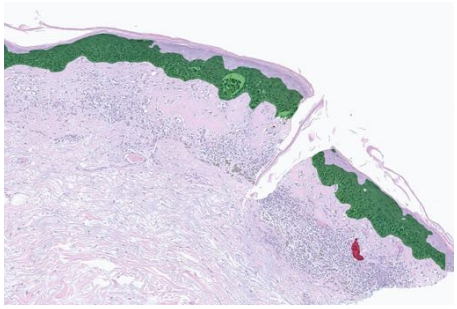
(d) SegFormer model prediction

Figure A.7: Test image 6. Both models perform similarly.

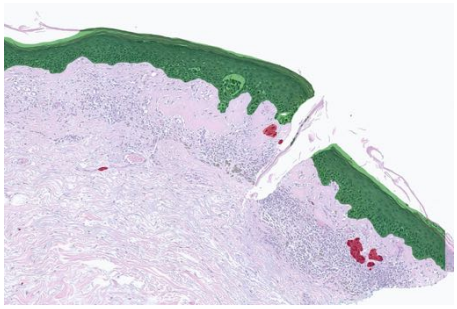


(a) Image

(b) Ground truth

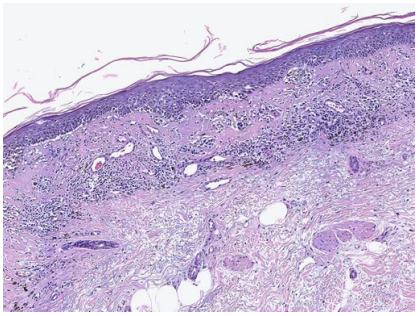


(c) HIPT model prediction



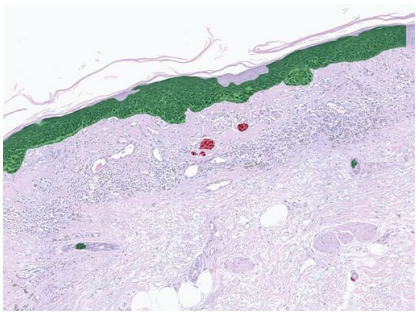
(d) SegFormer model prediction

Figure A.8: Test image 7. Both models perform similarly.

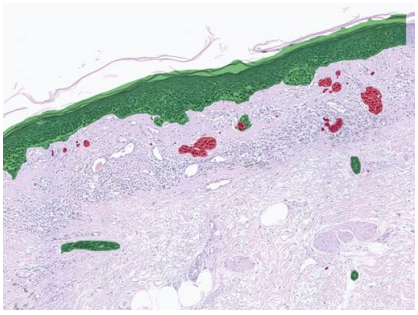


(a) Image

(b) Ground truth



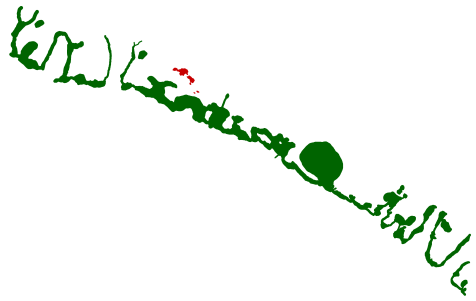
(c) HIPT model prediction



(d) SegFormer model prediction

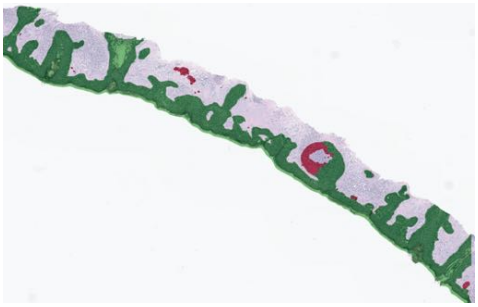
Figure A.9: Test image 8. SegFormer performs better than HIPT.

(a) Image



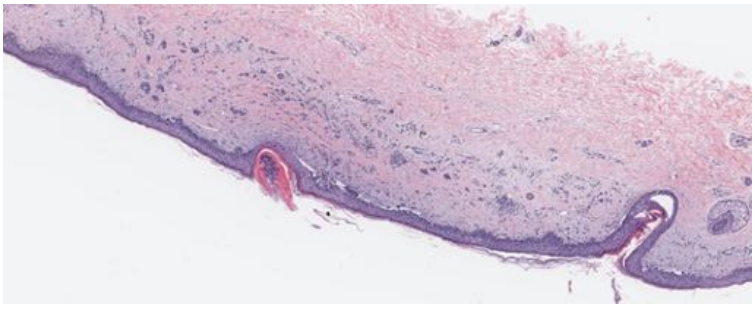
(b) Ground truth

(c) HIPT model prediction

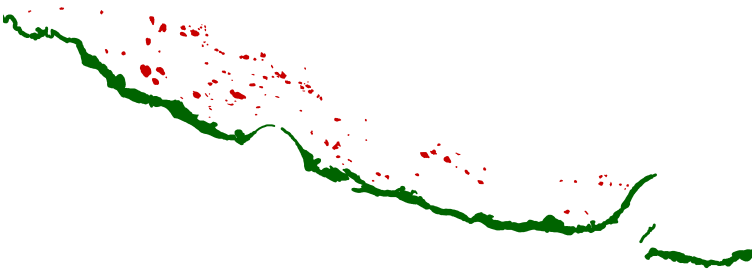


(d) SegFormer model prediction

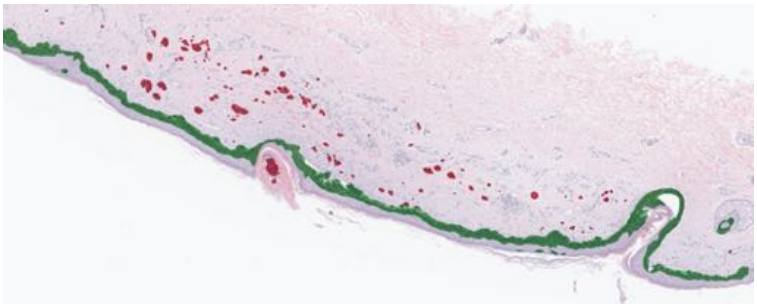
Figure A.10: Test image 9. HIPT performs better than SegFormer



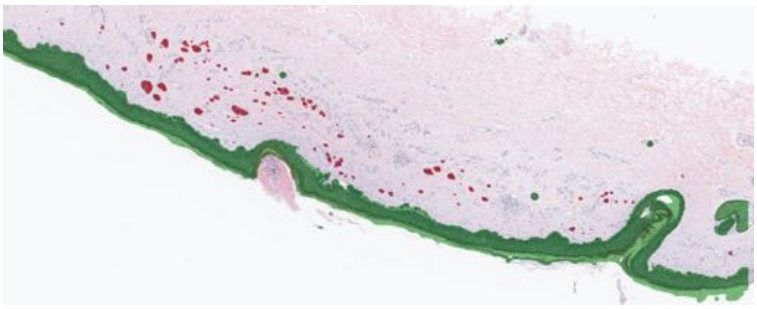
(a) Image



(b) Ground truth

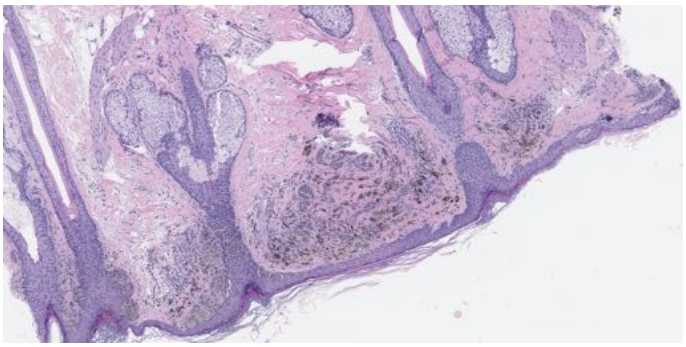


(c) HIPT model prediction



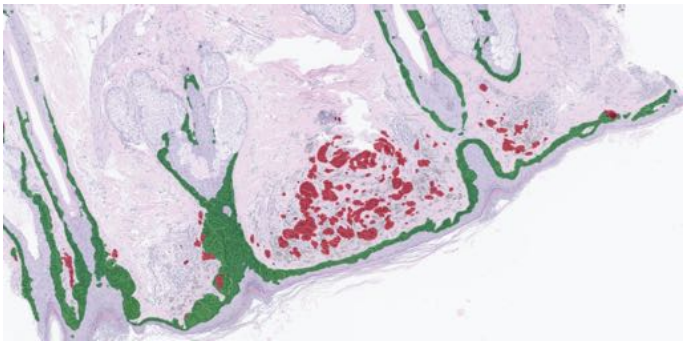
(d) SegFormer model prediction

Figure A.11: Test image 10. SegFormer performs better than HIPT.

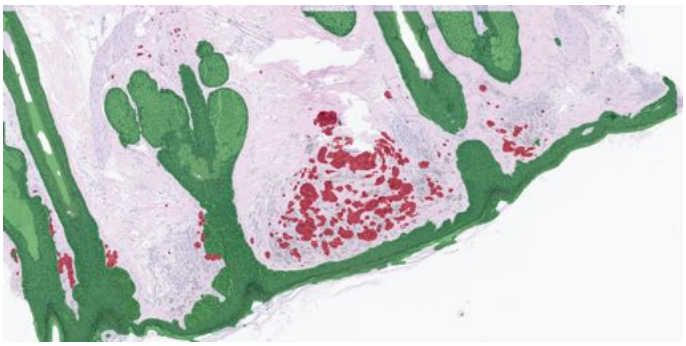


(a) Image

(b) Ground truth

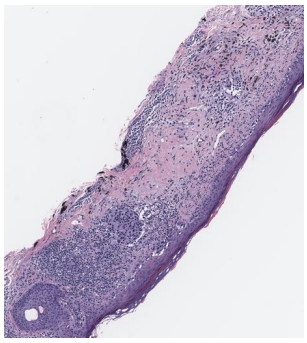


(c) HIPT model prediction

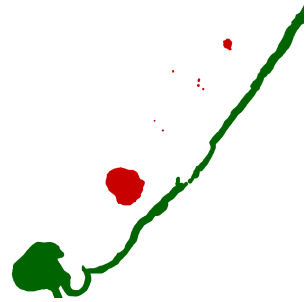


(d) SegFormer model prediction

Figure A.12: Test image 11. Both models perform similarly.



(a) Image



(b) Ground truth

(c) HIPT model prediction

(d) SegFormer model prediction

Figure A.13: Test image 12. SegFormer performs better than HIPT.

Bibliography

[1] URL: https://skincancerprevention.org/learning/melanoma-facts-statistics/.

[2] Salah Alheejawi, Mrinal Mandal, Hongming Xu, Cheng Lu, Richard Berendt, and Naresh Jha. "10 - Deep learning-based histopathological image analysis for automated detection and staging of melanoma". In: *Deep Learning Techniques for Biomedical and Health Informatics*. Ed. by Basant Agarwal, Valentina Emilia Balas, Lakhmi C. Jain, Ramesh Chandra Poonia, and Manisha. Academic Press, 2020, pp. 237–265. ISBN: 978-0-12-819061-6. DOI: https://doi.org/10.1016/B978-0-12-819061-6.00010-0. URL: https://www.sciencedirect.com/science/article/pii/B9780128190616000100.

[3] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer Normalization". In: *ArXiv* abs/1607.06450 (2016).

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2017), pp. 2481–2495. DOI: 10.1109/TPAMI.2016.2644615.

[5] Alexander Breslow. "Thickness, Cross-Sectional Areas and Depth of Invasion in the Prognosis of Cutaneous Melanoma". In: *Annals of Surgery* 172.5 (1970), pp. 902–908. DOI: 10.1097/00000658-197011000-00017.

[6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation". In: *Proceedings of the European Conference on Computer Vision Workshops(ECCVW)*. 2022.

[7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation". In: *arXiv preprint arXiv:2102.04306* (2021).

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4 (2018), pp. 834–848. DOI: 10.1109/TPAMI.2017.2699184.

[9] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. "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. DOI: 10.1109/CVPR52688.2022.01567.

[10] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. *Vision Transformer Adapter for Dense Predictions*. 2022. DOI: 10.48550/ARXIV.2205.08534. URL: https://arxiv.org/abs/2205.08534.

[11] Jia 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.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "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*. OpenReview.net, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[13] Andrew H. Fischer, Kenneth A. Jacobson, Jack Rose, and Rolf Zeller. "Hematoxylin and eosin staining of tissue and cell sections". In: *Cold Spring Harbor Protocols* 2008.5 (2008). DOI: 10.1101/pdb.prot4986.

[14] Dan Hendrycks and Kevin Gimpel. "Gaussian Error Linear Units (GELUs)". In: 2016. arXiv: http://arxiv.org/abs/1606.08415v3 [cs.LG].

[15] Dorit S. Hochbaum. "An Efficient and Effective Tool for Image Segmentation, Total Variations and Regularization". In: *Scale Space and Variational Methods in Computer Vision*. Ed. by Alfred M. Bruckstein, Bart M. ter Haar Romeny, Alexander M. Bronstein, and Michael M. Bronstein. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 338–349. ISBN: 978-3-642-24785-9.

[16] Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz. "Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[17] *Kinds of cancer*. June 2022. URL: https://www.cdc.gov/cancer/kinds.htm.

[18] Diederik Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations* (Dec. 2014).

[19] Bin Li, Yin Li, and Kevin W Eliceiri. "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14318–14328.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

[21] Timothy H. Mccalmont. "The Second Dimension—Integrating Calculated Tumor Area Into Cancer Diagnosis". In: *JAMA Dermatology* 155.8 (2019), p. 883. DOI: 10.1001/jamadermatol.2019.0609.

[22] *Melanoma*. en-US. URL: https://www.skincancer.org/skin-cancer-information/melanoma/ (visited on 11/16/2022).

[23] *Melanoma of the Skin - Cancer Stat Facts*. en. URL: https://seer.cancer.gov/statfacts/html/melan.html (visited on 11/16/2022).

[24] Neil Neumann, Michael Wang, Amal Mehta, Aman Shah, Mara Olson, Wudi Fan, Anna Weier, Avideh Zakhor, and Timothy McCalmont. "Quantifying Invasive Melanoma Volume by Deep Learning Segmentation at the Pixel-Level". In: *LABORATORY INVESTIGATION*. Vol. 102. SUPPL 1. SPRINGERNATURE CAMPUS, 4 CRINAN ST, LONDON, N1 9XW, ENGLAND. 2022, pp. 346–347.

[25] Shima Nofallah, Mojgan Mokhtari, Wenjun Wu, Sachin Mehta, Stevan Knezevich, Caitlin J. May, Oliver H. Chang, Annie C. Lee, Joann G. Elmore, Linda G. Shapiro, and et al. "Segmenting skin biopsy images with coarse and sparse annotations using U-Net". In: *Journal of Digital Imaging* 35.5 (2022), pp. 1238–1249. DOI: 10.1007/s10278-022-00641-8.

[26] Kay R. Oskal, Martin Risdal, Emilius A. Janssen, Erling S. Undersrud, and Thor O. Gulsrud. "A U-net based approach to epidermal tissue segmentation in whole slide histopathological images". In: *SN Applied Sciences* 1.7 (2019). DOI: 10.1007/s42452-019-0694-y.

[27] Adon Phillips, Iris Teo, and Jochen Lang. "Segmentation 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.

[28] Joe Qranfal, Dorit S. Hochbaum, and Germain Tanoh. "Experimental Analysis of the MRF Algorithm for Segmentation of Noisy Medical Images". In: *Algorithmic Operations Research* 6.2 (Jan. 2012), Pages 79 –90. URL: https://journals.lib.unb.ca/index.php/AOR/article/view/18234.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241.

[30] Gerald Saldanha, Jeremy Yarrow, Somaia Elsheikh, Marie O'Riordan, Hussein Uraiby, and Mark Bamford. "Development and Initial Validation of Calculated Tumor Area as a Prognostic Tool in Cutaneous Malignant Melanoma". In: *JAMA Dermatology* 155.8 (2019), pp. 890–898. DOI: 10.1001/jamadermatol.2019.0621.

[31] Mart van Rijthoven, Maschenka Balkenhol, Karina Siliņa, Jeroen van der Laak, and Francesco Ciompi. "HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images". In: *Medical Image Analysis* 68 (2021), p. 101890. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101890. URL: https://www.sciencedirect.com/science/article/pii/S1361841520302541.

[32] 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.

[33] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. "SegFormer: Simple and efficient design for semantic segmentation with transformers". In: *Advances in Neural Information Processing Systems* 34 (2021).

[34] Yuhui Yuan, Xilin Chen, and Jingdong Wang. "Object-Contextual Representations for Semantic Segmentation". In: *16th European Conference Computer Vision (ECCV 2020)*. Aug. 2020.

[35] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. "Pyramid Scene Parsing Network". In: *CVPR*. 2017.

[36] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers". In: *CVPR*. 2021.

[37] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. "Semantic understanding of scenes through the ade20k dataset". In: *International Journal of Computer Vision* 127.3 (2019), pp. 302–321.

[38] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. "UNet++: A Nested U-Net Architecture for Medical Image Segmentation". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Ed. by Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi. Cham: Springer International Publishing, 2018, pp. 3–11.

[39] 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. "Segmentation and Classification of Melanoma and Nevus in Whole Slide Images". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020, pp. 263–266. DOI: 10.1109/ISBI45749.2020.9098487.