Adapting Segment Anything Model to Invasive Melanoma Segmentation in Microscopy Slide Images

by

Qingyuan Liu

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 Michael Lustig

Summer 2024

The thesis of Qingyuan Liu, titled Adapting Segment Anything Model to Invasive Melanoma Segmentation in Microscopy Slide Images, is approved:

Chair Date 8/6/24

Date

Date

University of California, Berkeley

Adapting Segment Anything Model to Invasive Melanoma Segmentation in Microscopy Slide Images

Abstract

Adapting Segment Anything Model to Invasive Melanoma Segmentation in Microscopy Slide Images

by

Qingyuan Liu

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Avideh Zakhor, Chair

Melanoma segmentation in Whole Slide Images (WSIs) is useful for prognosis and the measurement of crucial prognostic factors such as Breslow depth and primary invasive tumor size. In this paper, we present a novel approach that uses the Segment Anything Model (SAM) for automatic melanoma segmentation in microscopy slide images. Our method employs an initial semantic segmentation model to generate preliminary segmentation masks that are then used to prompt SAM. We design a dynamic prompting strategy that uses a combination of centroid and grid prompts to achieve optimal coverage of the super high-resolution slide images while maintaining the quality of generated prompts. To optimize for invasive melanoma segmentation, we further refine the prompt generation process by implementing in-situ melanoma detection and low-confidence region filtering. We select Segformer as the initial segmentation model and EfficientSAM as the segment anything model for parameter-efficient fine-tuning. Our experimental results demonstrate that this approach not only surpasses other state-of-the-art melanoma segmentation methods but also significantly outperforms the baseline Segformer by 20.2% in terms of IoU.

To my beloved parents and my cherished companion, DuoDuo.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my greatest gratitude to my research advisor Professor Avideh Zakhor for her invaluable guidance and support throughout this journey. Her insightful feedback and continuous motivation have been crucial to the successful completion of this thesis. Thank you to Professor Michael Lustig for generously serving as the second reader. I would also like to express sincere thanks to Doctor Mike Wang for providing the dataset and dedicating countless hours for meticulous annotations. His commitment to details has been fundamental to the success of this research. Finally, I would like to extend my thanks to Shinwoo Choi and Michael Huang for their insightful discussions and assistance in producing results without which this thesis would not be possible.

Chapter 1

Introduction

Melanoma, one of the most serious forms of skin cancer, originates in melanocytes, the pigment-producing cells responsible for melanin production [27]. Based on its progression and location within the skin, melanoma can be categorized into two main types: in-situ melanoma and invasive melanoma. While in-situ melanoma represents cancerous melanocytes that are confined to the epidermis, invasive melanoma penetrates beyond the epidermis into the dermis, posing a significant risk of spreading to other vital organs. As invasive melanoma grows, it may invade blood vessels and lymphatic vessels, allowing cancer cells to detach from the primary tumor and cause metastatic cancer [1].

Early detection and accurate diagnosis of melanoma are crucial for improving the survival rate. While the the five-year survival rate for patients whose melanoma is detected early exceeds 99 percent, the survival rate drops to 74 percent when the melanoma spreads to the lymph nodes and plummets to as low as 35 percent when it metastasizes to distant organs [3]. Apart from early detection, timely treatment is also essential in raising the survival rate. For those with early stages of melanoma where the tumor is localized and not spread yet, patients treated more than 119 days after biopsy have a 41 percent higher risk of dying than those treated within 30 days of being biopsied [6]. The standard diagnosis practice begins with an initial examination of dermatoscopic features to determine the types of melanocyptic lesions. For suspicious and malignant lesions, a histopathologic analysis of skin biopsies stained with hematoxylin and eosin (H&E) is required [11]. While traditional approach requires examination of tissue specimens under a microscope, the advent of Whole Slide Images (WSIs) has revolutionized this process by digitizing tissue samples into high resolution images. These digital slide images enable pathologists to examine various characteristics such as the celluar architecture, breslow depth and complex histologic features that are crucial for determining the stage and aggressiveness of melanoma.

Recent studies [30, 23] have focused on utilizing deep learning technology for melanoma segmentation in whole slide images. Melanoma segmentation results with sufficient accuracy in slide images have proven highly beneficial for aiding diagnosis and assisting manual measurement of breslow depth and primary invasive tumor size, which are crucial prognostic factors. This shows the potential for a fully automatic diagnosis procedure. Phillips et al.

[23] demonstrated the effectiveness of multi-scale FCN in segmentation the dermis, epidermis and tumor in whole slide images. Wang et al. [30] utilized more advanced transformer models including Hiearachical Pyramid Transformers (HIPT) and Segformers to achieve accurate segmentation of invasive melanoma and the epidermis in microscopy slide images.

In this thesis, we propose a novel method to apply the Segment Anything Model (SAM) [18] to automatic melanoma segmentation in microscopy slide images. SAM has demonstrated great success in various computer vision tasks and has achieved state-of-the-art performance in a diverse range of image segmentation tasks, such as zero-shot instance segmentation [18, 28] and zero-shot edge detection [18, 4]. One of the key features of SAM is a prompt encoder that allows the model to adapt to diverse downstream segmentation tasks with prompt engineering. Trained on a vast visual dataset comprised of over 11 million images and 1 billion masks, SAM has demonstrated strong generalization as a foundation model to perform segmentation for a wide variety of objects.

Although SAM has demonstrated strong zero-shot generalization abilities, several studies [15, 38, 25, 8] have shown that its accuracy is limited in segmentation tasks that require specific domain knowledge. Recent studies [20, 31, 36] have proposed methods to adapt SAM for medical image segmentation, such as augmenting data with SAM's predictions or fine-tuning SAM for better performance. Despite showing impressive performance in segmenting medical images such as CT and MRI scans, these methods do not generalize well to microscopy images, especially for melanoma segmentation in microscopy slide images.

Our proposed method addresses these challenges by introducing an innovative framework that automatically generates prompts from an initial segmentation map to guide SAM. We use Segformer, an effective semantic segmentation model for melanoma segmentation, to generate an initial segmentation map. We design a dynamic prompt strategy that uses a combination of centroid and grid prompts. Additionally, we incorporate in-situ melanoma detection and low-confidence region filtering to ensure precise prompt generation. Our experimental results demonstrate that this approach not only surpasses other state-of-the-art melanoma segmentation methods but also significantly improves upon the baseline performance of Segformer by over 20%.

The structure of the thesis is as follows: Chapter 2 reviews existing work related to our problem. Chapter 3 provides a detailed description of our proposed method. Chapter 4 presents the experimental results and a comprehensive ablation study of each module in our method. Finally, Chapter 5 concludes our work and suggests potential research directions for future work. More qualitative results on our dataset are shown in Appendix A.

Chapter 2

Related Work

In this chapter, we review existing work relevant to our method. Section 2.1 delves into the advancements in foundation models within the field of computer vision. Section 2.2 analyzes the architecture of SAM along with its lightweight and efficient variants, and its adaptions for medical imaging. In Section 2.3, we explore related methods for parameter-efficient fine-tuning. Finally, in Section 2.4, we examine current deep learning techniques applied to melanoma segmentation in slide images.

2.1 Foundation Models in Computer Vision

Foundation Models, typically large-scale neural networks with billions of parameters pre-trained on vast datasets, have become the cornerstone for numerous tasks in the field of natural language processing (NLP) and computer vision. The concept of foundation models traces back to the success of large-scale models in NLP, where they demonstrate unprecedented abitlity to understand and generate human languages. With strong zero-shot and few-shot generalization abilities, these foundation models perform tasks and process data they have not explicitly encountered during training, often achieving performance levels comparable to models designed specifically for those tasks. This ability to genearalize from limited examples have made foundation models highly versatile and valuable in scenarios where labeled data is expensive to obtain. Even in cases where foundation models perform worse than expected, they provide a robust and high-performance baseline that can be adapted to specific applications with relatively minimal additional training.

In computer vision, foundation models have been explored and applied to a broad range of applications, including image classification, image generation, image segmentation, object detection and many more vision tasks. A significant advancement in this domain was the vision transformers (ViTs) designed by Dosovitskiy et al [9]. By leveraging the transformer model initially designed for NLP, ViTs process and analyze images by embedding fixed-size image patches as a sequence of tokens. This approach allows ViTs to effectively capture long-range dependencies and complex patterns in images, demonstrating impressive scalability with increasing data and adaptability across various vision tasks.

Vision Transformers have thus served as a backbone for many advanced foundation models. One notable example is CLIP (Contrastive Language-Image Pre-training) [24], which leverages a vision transformer as the core architecture for its image encoder. CLIP employs a self-supervised contrastive training approach by aligning text and visual representations within a shared latent space. By training on a vast number of image and text pairs sourced from the web, CLIP has exhibited impressive generalization capabilities across diverse vision tasks, including image classification, object detection, and zero-shot learning, all without requiring task-specific training. Despite CLIP's impressive performance, its data collection process is costly and resource-intensive, limiting its dataset size. To address this, ALIGN [16] adopts a different strategy by leveraging a noisy dataset comprising over one billion image-text pairs. ALIGN trains its foundation model by aligning visual and language representations, achieving strong zero-shot transferability to visual classification and image-text retrieval tasks. The visual representations produced by these foundation models have proven highly valuable for computer vision tasks, particularly scenarios with limited data availability.

2.2 Segment Anything Model

While the success of foundation models relies on the availability of vast amount of training data, many computer vision tasks such as image segmentation suffer from limited training data due to the high cost involved of annotation. The Segment Anything Model (SAM) [18], a promptable foundation model specifically designed for image segmentation, addresses this issue by overcoming data scarcity and has demosntrated strong generalization in zero-shot segmentation. SAM's effectiveness stems from its training on SA-1B, a huge extensive segmentation dataset co-developed with the model that contains 1.1 billion segmentation masks and 11 million images.

The development of SA-1B [18] fully leverages the two primary approaches of using SAM, one is to use SAM as an interactive model with manual prompts, the other one is to use SAM for automatic mask generation without human intervention. As illustrated in Figure 2.1, SAM's architecture includes three components: an image encoder, a prompt encoder, and a mask decoder. To ensure high scalability and adaptability, the image encoder incorporates a powerful vision transformer. The prompt encoder supports a variety of prompts including single and multiple point prompts, bounding boxes, masks, and texts. The versatility of SAM to adapt to new data distributions relies on prompt engineering through the prompt encoder. As an interactive model, SAM allows professional annotators to provide precise prompts without ambiguities, enabling accurate object segmentation. This interactive feature was crucial in the initial stage of building the SA-1B dataset where professional annotators used the initial version of SAM that was trained on common public datasets to assist with annotations.

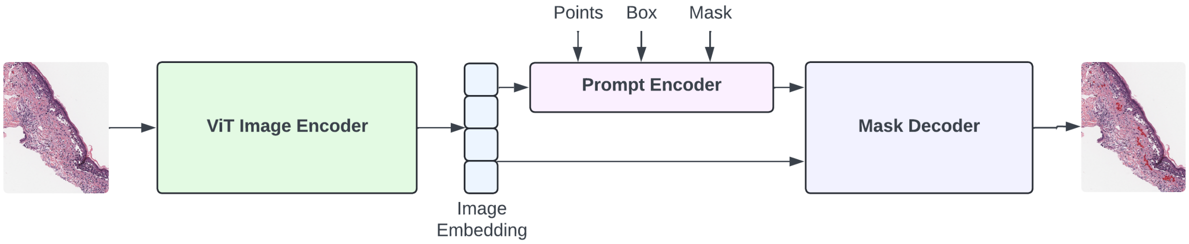


Figure 2.1: An overview of Segment Anything Model (SAM) architecture [18]. The heavy-weight ViT image encoder generates an image embedding and the mask decoder can produce various masks based on different prompts for a single image embedding.

In contrast, the automatic mask generation method [18] does not rely on manually provided precise prompts. Instead, it uses a dense regular grid of points to prompt the model. For each point, SAM predicts multiple masks with varying predicted IoU scores that indicates the model's confidence levels in predicted masks. After obtaining all masks, it filters by retaining only the confident and stable masks and de-duplicating using non-maximal suppression (NMS). This automatic segmentation approach was employed in the later stages of building SA-1B, once SAM had become more ambiguity-aware and had been refined through training on the initial stage of SA-1B dataset previously built with assistance from annotators. Although this allows for automatic segmentation using SAM, it is not designed to segment a specific category of objects, in our case the invasive melanoma. Our method addresses this limitation by allowing SAM to target invasive melanoma segmentation through automatically sampled prompts from a preliminary mask.

Lightweight and Efficient Variants of SAM

Despite SAM's impressive versatility and zero-shot generalization capabilities, its dependence on a large Transformer model incurs substantial costs for fine-tuning and inference, limiting its practical applications. To address these challenges, recent studies [37, 35, 33] have focused on reducing SAM's computational costs to enhance its usability for fine-tuning and deployment. FastSAM [37] replaces SAM's transformer backbone by YOLOv8-seg [17], a CNN-based architecture, and trains on a small portion of the SA-1B dataset for instance segmentation. MobileSAM [35] employs knowledge distillation to develop a lightweight image encoder from SAM's heavy image encoder. This method achieves faster inference speed while preserving performance levels comparable to FastSAM, making it more feasible for deployment in resource-constrained environments. EfficientSAM [33] adopts a masked autoencoders(MAE) framework for leveraging masked image pretraining. This technique enables the training of a light-weight encoder that effectively reconstructs visual representations from SAM's heavy image encoder. Unlike FastSAM and MobileSAM, which trade off a significant amount of SAM's performance for smaller model sizes and faster inference speed, EfficientSAM maintains reasonable performance comparable to the original SAM while significantly reducing complexity and computational requirements.

Adaptations of SAM in Medical Imaging

Given SAM's impressive results on various natural image segmentation tasks, recent works have explored its application to medical image segmentation, a field that stands to benefit significantly from foundation models due to the scarcity of data and the labor-intensive annotation for medical images. However, several studies [15, 38, 25, 8] have shown that SAM underperforms in medical image segmentation. This is attributed to the model's lack of domain-specific medical knowledge, the uncertain and complex object boundaries, intricate structures, and the wide-range of scales unique to medical objects [15]. Recent efforts have focused on adapting SAM for medical images, primarily through fine-tuning or adapting SAM to labeled medical dataset. Ma et al. [20] proposed to fine-tune SAM fully on labeled medical data, which is cost-ineffective due to the vast number of parameters in SAM. Wu et al. [31] proposed to integrate adapter modules into SAM, allowing efficient fine-tuning by freezing all modules except for the adapters during training.

It is important to note that these studies have primarily utilized SAM as an interactive model and evaluate its performance by providing accurate prompts based on the ground-truth annotations. This overlooks SAM's potential for automatic segmentation. Particularly in the context for microscopy whole slide image segmentation, SAM's interactive feature can assist annotators but cannot fully automate the labor-intensive process of segmenting high-resolution whole slide images. Although SAM offers automatic mask generation, this feature is designed for segmenting everything on an entire image rather than targeting specific small objects, such as scattered melanoma cells in our scenario. Zhang et al. [36] proposed to enhance medical images by adding semantic structures using SAM's automatic mask generation. This approach combines generated masks, features and stability scores to help train other image segmentation models with enhanced data. The success of this method depends on SAM's ability to generate useful structural information during the automatic mask generation process.

2.3 Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-tuning (PEFT) has emerged as an efficient strategy for adapting large foundational models to specific tasks in both NLP and computer vision. This approach involves freezing the majority of the model's weights and fine-tuning only a small subset of parameters, reducing the number of trainable parameters to as little as 0.01% of the original model in highly optimized scenarios [34, 14]. Numerous parameter- and compute-efficient methods have been developed for transformer-based models. Among the various PEFT methods, adapters have proven particularly effective for fine-tuning large vision models for downstream tasks. One of the most notable adaption techniques is Low-Rank Adaptation of Large Language Models (LoRA) [14], which reduces the number of trainable parameters by only training rank decomposition matrices injected into each layer of Transformers. While LoRA was initially popularized in NLP, recent studies have successfully applied adapter techniques to computer vision tasks, demonstrating their versatility and effectiveness. To adapt vision transformers for scalable visual recognition, Chen et al. [5] introduced lightweight modules into vision transformers. This approach enhances ViT's generalization capability to downstream tasks. Similarly, He et al. [13] proposed a method to project and decompose selected modules with specific local intrinsic dimensions in a subspace via Kronecker Adaption, enabling parameter efficient fine-tuning for vision transformers.

2.4 Melanoma Semantic Segmentation in Slide Images

Melanoma semantic segmentation can be classified into two categories: 1) skin lesion segmentation based on images captured at a macroscopic scale, providing a magnified view of the skin surface and sub-surface structures; 2) microscopy slide image segmentation that involves whole slide images (WSIs) captured at a microscopic scale, proving high-resolution, detailed views of tissue samples at the cellular level. Skin lesion segmentation involves segmenting the foreground melanoma at a macroscopic level, typically represented as a large connected component. In contrast, microscopy slide image segmentation requires segmenting melanoma at a microscopic scale, where it appears to scatter across WSIs with irregular shapes, making the task significantly more challenging. Given that these two segmentation tasks have distinct objectives and datasets, we focus exclusively on methods related to microscopic slide image segmentation. In the following sections, we categorize these methods based on their model architecture into two categorizes: Convolutional Neural Network (CNN)-based methods and Transformer-based methods.

CNN-based Methods

Many CNN-based methods have been developed for the segmenation of melanoma from microscopic WSIs. Due to the labor-intensive nature of comprehensively annotating WSIs with super large resolutions, these studies evaluate models on different datasets with varying annotation quality and standards. While some datasets have accurate annotations across entire WSIs, others have only coarse and sparse annotations for small portions of images, leading to differences in methodology. Nofallah et al. [21] proposed a two-stage segmentation pipeline using multiple U-Nets to generate separate segmentation masks for the epidermis and melanoma, which are then merged to create the final segmentation mask. Their dataset includes sparse and coarse annotations within regions of interests (ROIs) determined by a consensus panel of dermatopathologists. Phillips et al. [23] created a dataset with annotations covering full WSIs including tumor, epidermis and dermis. To deal with the gigapixel resolution of WSIs, they divided each slide into patches of 512 × 512 pixel resolution and deallt with class imbalance by undersampling patches containing mostly backgrounds and upsampling those containing tumors. Different from other works, Alheejawi [2] utilized a unique dataset that labeled nuclei only and employed CNN-based architectures including SegNet and U-Net for cell nuclei segmentation. Van Zon et al. [29] worked on a dataset including melanoma, nevus and negative WSIs, and proposed to first classify slides into melanoma and nevus, and then segment both melanoma and nevus using a 3-layer U-Net architecture. Oskal et al. [22] concentrated solely on segmenting epidermis and trained a U-net based architecture on sampled patches at 512 × 512 pixels. To handle class imbalance between epidermis regions and the rest of the slides, they upsampled epidermis and downsampled background patches based on the number of pixels belonging to each class.

The work most related to ours is the approach proposed by Shah et al [26], who used the same dataset as ours. They developed a two-stage method to identify the epidermis and melanoma separately by leveraging the fact that in-situ melanoma is confined to the epidermis while invasive melanoma penetrates beyond the epidermis. Their dataset includes detailed and accurate annotations for background cells, epidermis, invasive tumor, inflated tumor, fibrotic tumor and uncertain tumor across entire WSIs. Specifically, they used CNN-based models including HRNet-OCR and HookNet as backbones and trained two models, one for epidermis segmentation and one for tumor segmentation, to predict two separate segmentation masks for epidermis and melanoma. By removing all predicted melanoma masks inside the predicted epidermis masks, they obtained segmentation masks for invasive melanoma. Although it is feasible to train a model to segment both in-situ melanoma and invasive melanoma, this method utilizes the annotation in a problematic way by combining epidermis and invasive melanoma together as one class for tumor segmentation. As a result, this leads the model to inadvertently identify healthy epidermis tissues as tumor.

Transformer-based Methods

All the previously discussed works share a common approach: utilizing CNNs for microscopy slide image segmentation. While CNN-based architectures have demonstrated significant effectiveness in segmentation tasks, transformers have outperformed CNNs in various computer vision tasks due to their high scalability with increasing data and their ability to capture long-range dependencies and global context. Wang et al. [30] proposed to fine-tune transformer based models including the Hierarchical Pyramid Transformer (HIPT) and the Segformer [32] for segmentation of invasive melanoma in whole slide images. Specifically for HIPT, a multi-scale hierarchical decoder is used to fine-tune the HIPT pretrained on WSIs of breast tissues using student-teacher distillation, leveraging the fact that breast tissues and skin tissues share many common biological features. However, fine-tuning Segformer that is pretrained on ImageNet dataset leads to better performance than HIPT pretrained on WSIs of breast tissues. The multi-scale hierarchical feature maps of Segformer has demonstrated superior segmentation accuracy compared to the single-scale low resolution representation of HIPT.

Despite Segformer's impressive capabilities in segmenting invasive melanoma, there is still some room for improvement. The task is currently treated as a multi-class semantic segmentation problem involving three classes: melanoma, epidermis, and everything else including background cells, inflated tumor, fibrotic tumor and uncertain tumor. During inference, Segformer predicts three masks for each class and selects the class with the highest pixel-wise probability for the final segmentation mask. However, Segformer sometimes struggles in hard-to-distinguish regions where it can not distinguish between epidermis and melanoma, likely due to limited global context available in each sampled patch. As a result, Segformer predicts similar probabilities for both classes and the final segmentation masks contain areas where epidermis and melanoma are intertwined with each other. Furthermore, Segformer tends to perform better on scattered round melanoma cells while being less accurate in segmenting large, irregular melanoma clusters. From a medical perspective, melanoma cells tend to migrate and invade surrounding tissues differently, leading to different growth patterns. In Chapter 3, we will show how we leverage these observations to generate better segmentation masks for invasive melanoma.

Chapter 3

Proposed Method

In this chapter, we describe our proposed method for segmenting invasive melanoma in microscopy slide images. To begin with, we briefly introduce the architecture of Segformer and EfficientSAM, the two models used in our approach in Section 3.1. Next, in Section 3.1, we show the overall pipeline of how to use SAM together with a semantic segmentation model for automatic segmentation of invasive melanoma. In Section 3.2, we delve into the integration of adapters into SAM for parameter efficient fine-tuning. In Sections 3.3 and 3.4, we show the process of in-situ melanoma detection and low-confidence region filtering in preparation for prompt generation. In Section 3.5, we describe the details of our prompt generation method including the preparation stage and the strategy for determining prompt types. Section 3.6 describes the process of deriving the final segmentation mask. Lastly, in Section 3.7, we show how to train a SAM that best suits the purpose of our proposed method.

3.1 Preliminary

Segformer

Segformer [32] is a lightweight, fast transformer-based model designed specifically for semantic segmentation tasks. Its architecture leverages multi-resolution hierarchical structures, enhancing its capability to capture fine details at various scales. Rather than use positional encoding to embed location information, Segformer employs zero paddings in its convolution kernels to leak positional information, which avoids the reduction in accuracy when there is a discrepancy between the scaling ratios of training and testing data. In the domain of medical imaging, Segformer [30] has demonstrated exceptional accuracy in segmenting invasive melanoma in microscopy slide images. Specifically, it approaches this task as a multi-class segmentation problem by producing detailed segmentation masks for both the epidermis and the melanoma. We choose to use Segformer as our image segmentation model due to its superior performance in this specific task compared to other transformer-based models

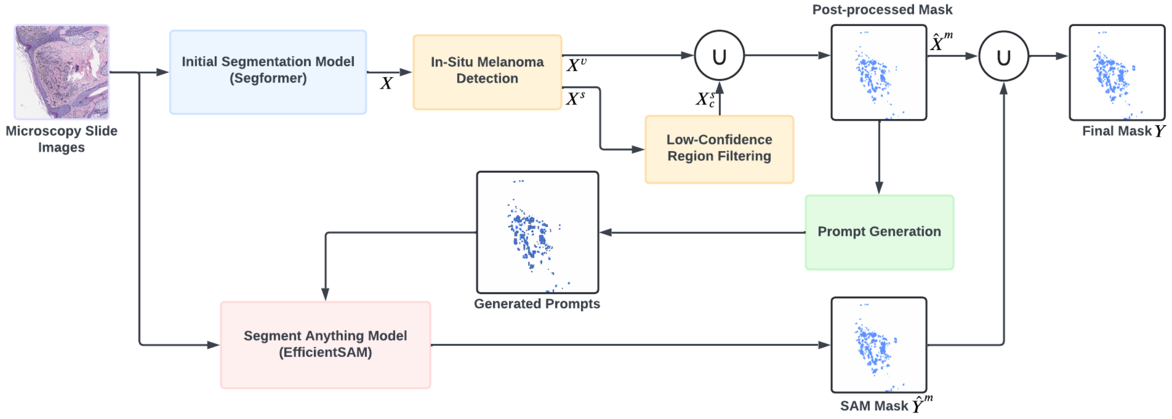


Figure 3.1: An overview of the proposed method. The initial mask X generated by Segformer is post-processed to generate the mask $\hat{X}^m$. We run SAM on the prompts generated from $\hat{X}^m$ to generate the mask $\hat{Y}^m$. The two masks are combined to create the final mask Y.

and CNN-based models. Its capability to generate segmentation masks for epidermis also significantly enhances our method's effectiveness.

EfficientSAM

EfficientSAM [33] is a light-weight SAM model with state-of-the-art quality-efficiency trade-offs. EfficientSAM shares the same architecture as SAM, which includes an image encoder, a prompt encoder, and a mask decoder. The only difference is that EfficientSAM uses a well-pretrained lightweight ViT image encoder, i.e. ViT-Tiny/-Small [9], to replace the heavier ViT-H encoder used in the original SAM model. To develop this high-quality lightweight ViT image encoder, the Masked Autoencoders (MAE) [12] pretraining method is employed. This approach involves training a masked image model with a lightweight ViT encoder to reconstruct feature embeddings generated by the heavy ViT-H encoder of SAM. The resulting lightweight image encoder combined with SAM's default mask decoder forms the EfficientSAM model. After self-supervised pretraining of its image encoder, EfficientSAM is fine-tuned on the SA-1B dataset, optimizing its performance on the segment anything task. Due to its cost-efficient fine-tuning and comparable performance to the best SAM model, we choose to use EfficientSAM as the segment anything model in our method.

Overview of the Proposed Method

Our proposed method comprises an initial semantic segmentation model and a segment anything model as illustrated in Figure 3.1. Our objective is to fine-tune SAM on our medical dataset and use it to segment invasive melanoma automatically. To achieve this, we prompt SAM with prompts generated from the mask produced by the initial segmentation model. To optimize for our task of segmenting invasive melanoma, we select Segformer as the initial segmentation model due to its superior performance compared to other models [30]. We choose EfficientSAM [33] as the segment anything model for parameter-efficient fine-tuning. The entire framework is described as follows.

Step 1. Initial Mask Generation. We run Segformer[32] to generate the initial segmentation mask X. Next, we develop an in-situ melanoma detection algorithm to separate X into the estimated in-situ melanoma regions X^s and the remaining invasive melanoma regions X^v. We filter out low-confidence regions from X_s to obtain the filtered mask X_c^s and combine it with X^v to obtain the post-processed mask $\hat{X}^m$. The details are described in Sections 3.3 and 3.4.

Step 2. Prompt Generation. We generate single point prompts from the post-processed mask $\hat{X}^m$. We determine the best prompt type for each connected component based on its shape distribution and geometric characteristics. The details of the prompt generation strategy are described in Section 3.5.

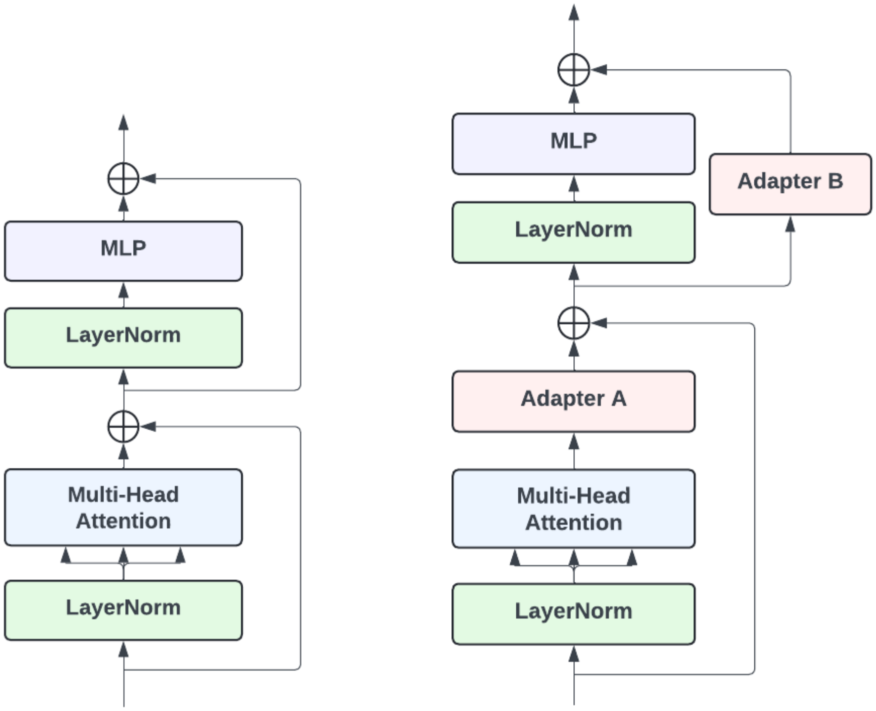
Step 3. Final Mask Generation. We run SAM on the generated prompts to produce its own invasive melanoma mask $\hat{Y}^m$. We use SAM's mask to refine the post-processed mask $\hat{X}^m$ by combining the two masks together. This aims to enhance the overall accuracy and robustness of melanoma segmentation. The details are described in Section 3.6.

3.2 Adapter Modules

Rather than perform full fine-tuning over all parameters of SAM, we adopt an adaption method as described by Wu et al. [31] to effectively fine-tune our dataset. Specifically, we integrate adapters into the ViT blocks of SAM's image encoder following Wu et al.'s approach [31]. We freeze all parameters except for the adapters in the image encoder and fully fine-tune everything else including the prompt encoder and the mask decoder. Unlike the original method proposed by Wu et al.[31], which integrates adapters into the SAM decoder with Hyper-Prompting Adapter, we choose to only integrate adapters into the image encoder. This decision is based on the fact that both the prompt encoder and the mask decoder are lightweight neural networks with significantly fewer parameters compared to the image encoder, which allows for efficient fine-tuning without the need for adapters.

We describe the adaption method from [31] as follows. Two adapters with different architectures are integrated into the ViT blocks of the SAM image encoder. As illustrated in Figure 3.2, the original ViT block consists of multiple components with residual connections including multi-head self-attention, layer normalization and a feed-forward network. Our adapters are placed into two specific positions: the first adapter is inserted after multi-head attention, while the second adapter is positioned on the residual path of the MLP, after residual connection of multi-head attention.

Both adapters include a fully connected feed-forward network consisting of two projection layers and a GeLU activation in between. The main difference between the two adapters



(a) ViT Block

(b) ViT Block with Adapters

Figure 3.2: Adatpers are integrated into two specific locations within the ViT block: one is placed after the multi-head attention, the other one is positioned on the last residual path.

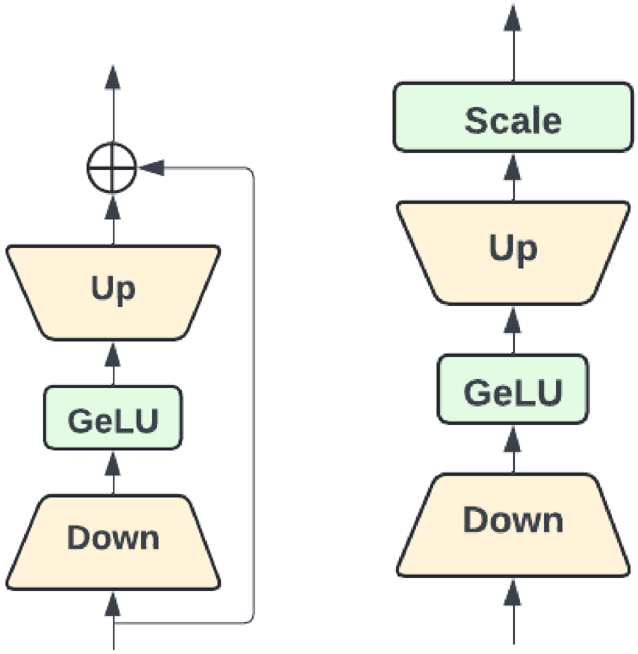
is whether they comprise residual connection. While the first adapter includes a residual connection, the second adapter is positioned on a residual path and thus does not contain a residual connection. This architectural difference of the two adapters is highlighted in Figure 3.3.

Denote the dimensionality of input and output as d_a and the dimensionality of the hidden layers as d_h. The first projection layer FC_1 projects input x to a low-dimensional space from feature dimension d_a to d_h. The second projection layer FC_2 expands the embedding back to a high-dimensional representation by increasing the dimensionality from d_h to d_a. The first adapter $Adapter_A$ includes a residual connection for its feed-forward network.

$$Adapter_A(x) = x + FC_2(GeLU(FC_1(x))) \tag{3.1}$$

Rather than contain a residual connection, the second adapter scales the output of the up-projection layer with a coefficient α.

$$Adapter_B(x) = \alpha FC_2(GeLU(FC_1(x))) \tag{3.2}$$



(a) Adapter A

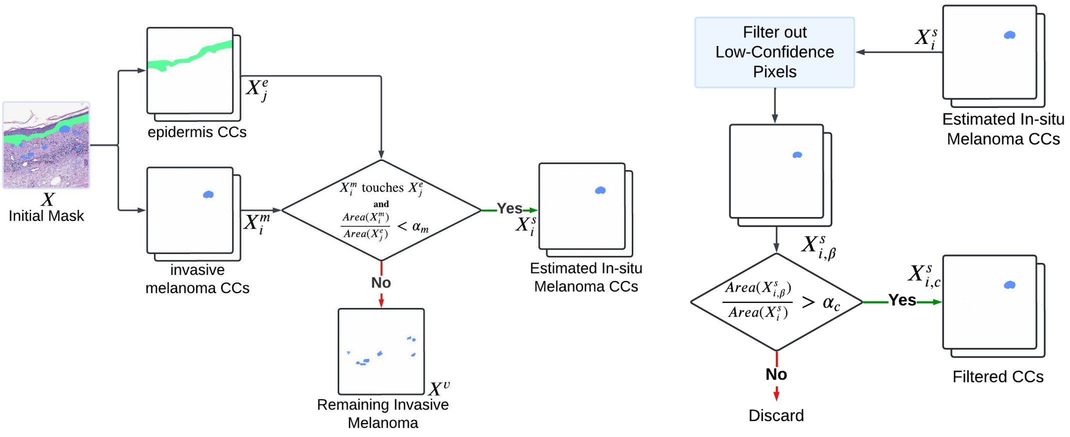
(b) Adapter B

Figure 3.3: A comparison between adapter A and adapter B: adapter A includes a residual connection, whereas adapter B replaces the residual connection with a scaling component.

In our implementation, we use adapters with an input and output dimensionality of $d_a = 768$ and set the dimensionality of hidden layers to $d_h = 1024$. For the second adapter, the scaling factor α is set to 0.5.

3.3 In-situ Melanoma Detection

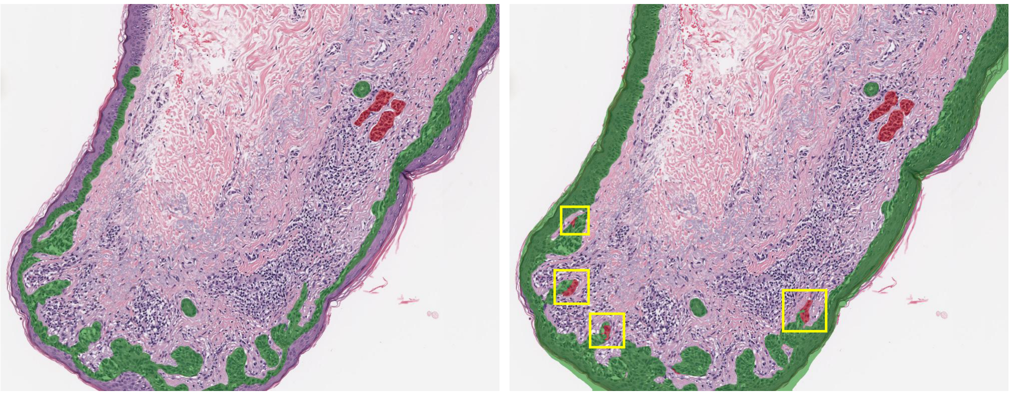
In this section, we describe our method for in-situ melanoma detection, which as seen in Figure 3.1, takes the output of initial segmentation by Segformer as input, and produces estimates of invasive melanoma mask X_v and in-situ melanoma mask X_s. The motivation is to exclude low-confidence invasive melanoma predictions that touch the epidermis predictions. From a medical perspective, melanoma confined to the epidermis are considered to be in-situ melanoma. When an in-situ melanoma component and its surrounding pixels are predicted as invasive melanoma, it appears to be an invasive melanoma component that touches the epidermis prediction. Therefore, we want to identify and exclude low-confidence invasive melanoma predictions that touch the epidermis predictions. This process is shown


epidermis CCs
X_j^e
X
Initial Mask
invasive melanoma CCs
X_i^m
X_i^m touches X_j^e and $\frac{Area(X_i^m)}{Area(X_j^e)} < \alpha_m$
Yes
X_i^s
Estimated In-situ Melanoma CCs
No
X^v
Remaining Invasive Melanoma
Filter out Low-Confidence Pixels
X_i^s
Estimated In-situ Melanoma CCs
$X_{i,\beta}^s$
$\frac{Area(X_{i,\beta}^s)}{Area(X_i^s)} > \alpha_c$
Yes
$X_{i,c}^s$
Filtered CCs
No
Discard

(a) In-situ Melanoma Detection

(b) Low-Confidence Region Filtering

Figure 3.4: In-situ melanoma detection algorithm finds the estimated in-situ melanoma regions X^s from the initial mask X. Low-confidence region filtering discards low-confidence connected components from X^s.



(a) Ground Truth

(b) Segformer-B0 Output

Figure 3.5: Motivation for in-situ melanoma detection. The highlighted areas show that the misclassified invasive melanoma components that touch the epidermis predictions have relatively small sizes compared to the epidermis.

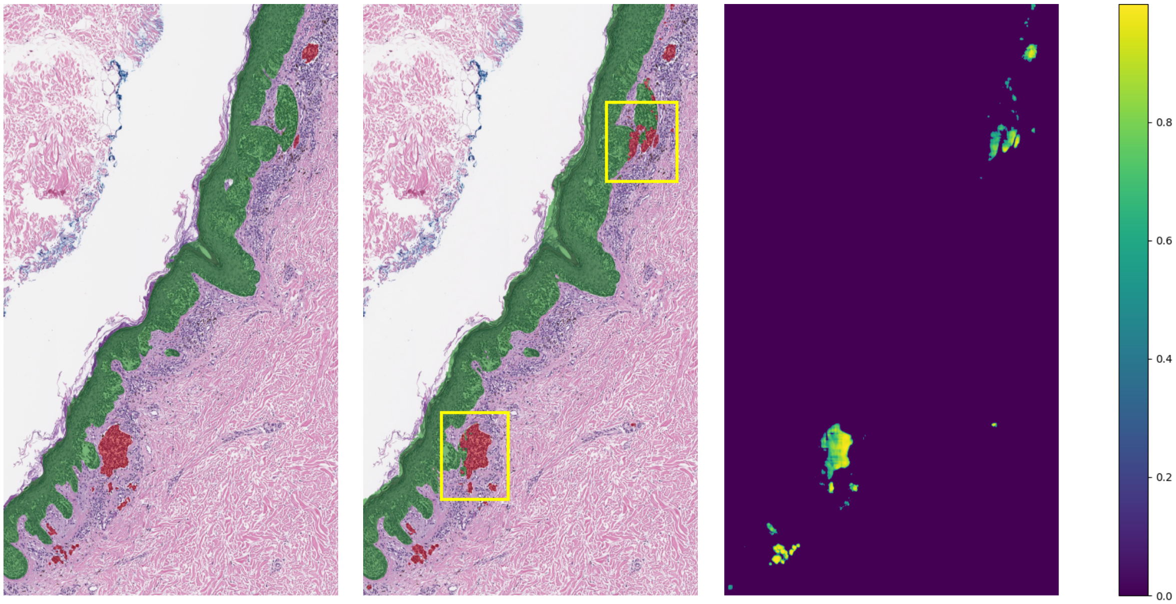
in Figure 3.4a. The initial mask X produced by Segformer contains both the epidermis mask X^e, and the invasive melanoma mask X^m. The invasive melanoma mask can be represented as a union of invasive melanoma connected components (CCs) $X^m = \bigcup X_i^m$, where i denotes the ith invasive melanoma connected component. Similarly, the mask for epidermis can be represented as a union of epidermis connected components $X^e = \bigcup X_j^e$, where j denotes the jth epidermis connected component. Since SAM heavily relies on prompts to specify the

Algorithm 1 Post-processing Algorithm

```
Require:
  The initial segmentation mask: $X$
  The epidermis mask: $X^e$
  The invasive melanoma mask: $X^m$
  The threshold for determining in-situ melanoma: $\alpha_m$
  The probability for thresholding high-confidence region: $\beta$
  The threshold for excluding low-confidence regions: $\alpha_c$
Ensure:
  The post-processed mask for invasive melanoma: $\hat{X}^m$
  $X \leftarrow \{X^e, X^m\}$
  $\hat{X}^m \leftarrow X^m$
  for each $X_i^m \subseteq X^m$ do
      for each $X_j^e \subseteq X^e$ do
          if $\textsc{Touch}(X_i^m, X_j^e)$ & $\frac{\textsc{Area}(X_i^m)}{\textsc{Area}(X_j^e)} < \alpha_m$ then
              $X_{i,\beta}^m = \{x \mid P(x) > \beta, x \in X_i^m\}$
              if $\frac{\textsc{Area}(X_{i,\beta}^m)}{\textsc{Area}(X_i^m)} < \alpha_c$ then
                  $\hat{X}^m \leftarrow \hat{X}^m \setminus X_i^m$
              end if
          end if
      end for
  end for
  return $\hat{X}^m$
```

exact objects to segment, inaccurate and ambiguous prompts can significantly degrade its performance. This can lead SAM to produce more false positives compared to the initial segmentation mask from which prompts are generated. Therefore, we filter out low-confidence invasive melanoma components to improve the accuracy of prompts so that when clicked it results in true positives as much as possible.

As shown in Figure 3.4a, we begin by iterating over all connected components X_i^m in the invasive melanoma mask X^m. For each connected component X_i^m, we determine whether it touches any epidermis component X_j^e. If such a touch exists, we compute the ratio between the area of the melanoma component X_i^m and the area of the touched epidermis region X_j^e. If this ratio falls below a specified threshold α_m, we classify it as an estimated in-situ melanoma region, This is because careful inspection of Segformer's predictions shows that in-situ melanoma that are misclassified as invasive melanoma tend to have a relatively small size compared to the epidermis, as shown in Figure 3.5. This process results in the estimated in-situ melanoma mask X^s and the remaining invasive melanoma regions X^v.



(a) Ground Truth (b) Segformer-B0 Output (c) Probability Distribution

Figure 3.6: Motivation for low-confidence region filtering. The green mask denotes the epidermis and the red mask indicates the invasive melanoma. For misclassified invasive melanoma components that touch the epidermis predictions, Segformer predicts a larger proportion of low-probability areas compared to true postivies.

3.4 Low-Confidence Region Filtering

As shown in Figure 3.1, we further filter out low-confidence estimated in-situ melanoma regions from X^s. The details of the full post-processing process including in-situ melanoma detection and low-confidence region filtering are shown in Algorithm 1. We keep high-confidence regions even if they touch the epidermis predictions, since this could result from false predictions in the epidermis and invasive melanoma that make them touch each other. We determine the confidence level based on the probability map generated by Segformer. As shown in Figure 3.4b, for each connected component X^s_i of X^s, we first find its high-confidence sub-component $X^s_{i,\beta}$ that has a probability exceeding the defined probability threshold β:

$$X^s_{i,\beta} = \{x \mid P(x) > \beta, x \in X^s_i\}. \tag{3.3}$$

We then determine the confidence level by computing the ratio between the area of $X^s_{i,\beta}$ and X^s_i. If the ratio is lower than the confidence threshold α_c, we discard this component and do not generate prompts from it. This is motivated by the empirical observation that Segformer generally predicts a larger proportion of low-probability areas for in-situ melanoma compared to invasive melanoma, as shown in the probability distribution in Figure 3.6. Therefore, if

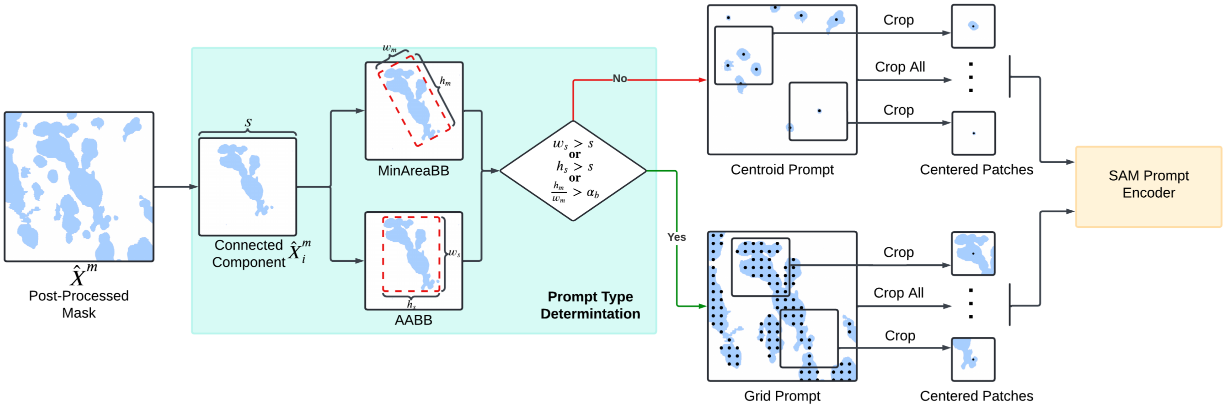


Figure 3.7: Prompt Generation Strategy. We apply the centroid prompt and the grid prompt to different invasive melanoma components based on their shape distributions and geometric characteristics.

we only keep regions with a high proportion of high probability areas, we can significantly reduce the number of false positives in X^m.

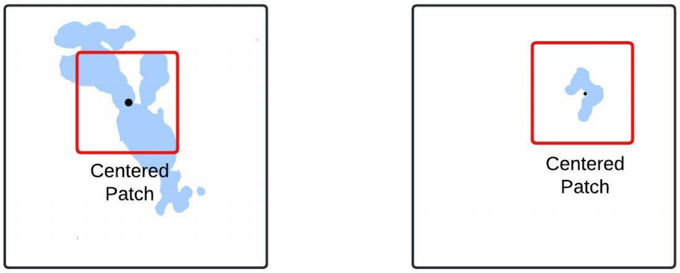
Next, as shown in Figure 3.1, we combine the filtered melanoma regions X_c^s and the invasive melanoma regions X^v obtained from in-situ melanoma detection block to obtain the post-processed invasive melanoma mask $\hat{X}^m$.

3.5 Prompt Generation

Melanoma can be in various forms in microscopy slide images, ranging from rounded clusters to jagged streaks or any combination of irregular shapes. To facilitate accurate segmentation of diverse morphologies, we design two types of point prompts: centroid and grid. Centroid prompts consist of a single point placed at the centroid of a connected component. On the other hand, grid prompts comprise a grid of points distributed within a connected component. As shown in Figure 3.7, we choose a prompt type for each connected component $\hat{X}_i^m$ and serve each point as a single point prompt for the SAM prompt encoder and feed a patch centered at the point as the input to the SAM image encoder.

A patch represents a 512 × 512 or 1024 × 1024 image segment cropped from the microscopy slide image, which is relatively small compared to the entire slide image. We choose patches centered at each point since this ideally provides maximum context, as illustrated in Figure 3.7. A patch may cover multiple melanoma components, but we only use the connected component corresponding to each point prompt as the target.

We choose prompts dynamically based on the characteristics of the connected components. While centroid prompts excel in providing optimal context for small, regularly shaped melanoma, they often fail to provide sufficient context for those covering large regions and



(a) Too Large to fit inside a patch

(b) Centroid lies outside CC

Figure 3.8: Cases where a patch centered at centroid cannot provide enough context for segmentation. In case (a), the melanoma component is too large to fit inside a patch centered at centroid. In case (b), the centroid lies outside the connected component.

those with irregular shapes, as shown in Figure 3.8a. In cases where a connected component is too large to fit inside a patch, numerous melanoma predictions are lost in a centered patch and not enough context is provided for its surroundings, as shown in Figure 3.8b. Furthermore, there are cases where the centroid lies outside the connected component and it turns out to be a point prompt not clicked on invasive melanoma, resulting in false positives.

For each connected component $\hat{X}_i^m$, we determine the type of prompts to generate based on its shape distribution and geometric attributes. The details of the prompt type determination strategy are shown in Algorithm 2. As shown in Figure 3.7, we find the width w_s and height h_s of an axis aligned bounding box (AABB) for the connected component. If either w_s or h_s exceeds the length of a patch, we opt for grid prompts. Then we find the width w_m and height h_m of an arbitrarily oriented minimum bounding box that has minimum area for the connected component and check whether the ratio between h_m and w_m exceeds a threshold α_b assuming h_m is the longer side. A high ratio implies that the shape of the component resembles a narrow rectangle or any other irregular shape where one dimension is larger than the other, in which case centroid prompt is ambiguous for segmenting the whole component. In such cases, we also choose to use grid prompts. Additionally, for each connected component chosen for grid prompts, we also include its centroid as part of the prompt if it is inside the component as a supplementary to the grid prompts. For all other cases, we use centroid prompt only.

Furthermore, if any connected component meeting the criteria for grid prompts has an area larger than $\frac{1}{4}$ of the patch, we apply grid prompts to all connected components in the slide image rather than determine prompt types individually for each component. This approach is based on the observation that neighbouring melanomas often share similar shapes, even if they don't conform to a single morphology. When a very large melanoma component with an irregular or narrow shape is found, neighboring melanomas tend to exhibit similar

Algorithm 2 Determine prompt type for a connected component

```
Require:
    A connected component of invasive melanoma: $x$
    The length of the side of a patch: $s$
    The threshold for the ratio between sides of a minimum bounding box: $\alpha_b$
Ensure:
    The generated prompts (a set of single point prompts): $P$
function GeneratePrompts(x)
    $P = \emptyset$
    $w_s, h_s =$ FindAABB($x$)
    $w_m, h_m =$ FindMinAreaBB($x$)
    if $w_s > s$ or $h_s > s$ or $\frac{h_m}{w_m} > \alpha_b$ then
        $P_{grid} =$ GenerateGridPrompts($x$)
        $P = P \cup P_{grid}$
        $center =$ FindCentroid($x$)
        if $center \in x$ then
            $P = P \cup \{center\}$
        end if
    else
        $center =$ FindCentroid($x$)
        $P = P \cup \{center\}$
    end if
    return P
end function
```

shapes. In such cases, grid prompts provide the best coverage for all melanoma components.

3.6 Final Mask Generation

After deriving the prompts, we run SAM to generate its own mask $\hat{Y}^m$ as shown in Figure 3.1. To obtain the final mask, we combine the post-processed Segformer's mask $\hat{X}^m$ and SAM's mask $\hat{Y}^m$ by performing a union between the two:

$$Y = \hat{Y}^m \cup \hat{X}^m \tag{3.4}$$

We choose to combine these two masks because single point prompts are sparse and may not comprehensively cover all regions. This sparsity can lead to ambiguity and result in missing some high-confidence predictions from Segformer. Therefore, we choose to include the post-processed mask $\hat{X}^m$ to ensure that high-confidence regions are kept in the final mask Y.

3.7 Training Strategy

We devise a training strategy involving two stages that optimizes SAM for the way we use prompts. Since we only use single point prompts, we first train our model on single point prompts for melanoma. For each patch sampled from a microscopy slide image, we first find all connected components and for each connected component we generate a single point prompt clicked at a random position inside it. This enables the model to learn to segment melanoma anywhere in a patch. After the initial training on random point prompts, we fine-tune our model on patches centered at each individual single point prompt. Each patch might contain multiple components, but we only set the component containing the point prompt as the target to reduce ambiguity. This allows the model to optimize for inference using our SAM-based method.

Chapter 4

Experiments

In this chapter, we present the experimental results from the methods described in Chapter 3. The chapter is structured as follows. Section 4.1 covers the data and preprocessing steps in our experiments. Section 4.2 discuss the implementation details for Segformer and EfficientSAM. In Sections 4.3, we showcase the main results of our method and compare it with other state-of-the-art methods. In Section 4.4, we analyze and discuss the effects of different design choices in our method. We include the full results of the microscopy slide segmentation in Appendix A.

4.1 Dataset

Our dataset comprises 101 microscopy slide images, with sizes ranging from 23,700 × 21,199 pixels to 1996 × 1679 pixels. These images are derived from skin biopsies stained with H&E [10] and captured under a microscope at 40x magnification. Detailed Annotations at this magnification level are provided by an expert dermatopathologist. Each image in the dataset is carefully selected and cropped from the original WSIs by the dermatopathologist. This is because whole slide images often contain multiple focal planes of the same tissue, leading to redundancy in the data. By selecting the most informative regions at a single focal plane, we reduce the redundancy and ensure that the dataset focuses on the most relevant tissue structures. The annotations for our microscopy images include seven classes: air, background cells, epidermis, invasive melanoma, inflamed tumor, fibrotic tumor, and uncertain tumor. It is important to note that invasive melanoma is the only type of melanoma precisely annotated. In-situ melanoma is labeled as epidermis in our dataset due to its confinement to the epidermis and due to our focus on segmenting invasive melanoma.

Segformer Dataset Generation

For preprocessing the dataset for Segformer, We follow the approach described by Wang et al [30]. Given the extremely high resolution of microscopy slide images, we divide each slide into non-overlapping patches of 512 × 512 and 1024 × 1024 pixels. We re-categorize the original annotations into three distinct classes: invasive melanoma, epidermis, and others. This reclassification is essential since the ambiguous boundaries of fibrotic and inflamed tumor make them less suitable for effective segmentation model training. By grouping these into a single category, the model can concentrate more on accurately segmenting invasive melanoma and epidermis. To address class imbalance issues, we perform selective undersampling. Patches with over 97% of air content are excluded, while those with more than 97% background cells are retained with a probability of 8%. The resulting dataset comprises 14,885 patches of around 3.9 billion pixels for 512 × 512 resolution and 4326 patches of around 4.5 billion pixels for 1024 × 1024 resolution. This preprocessing approach ensures a balanced and representative dataset for training the Segformer model.

SAM Dataset Generation

We generate the SAM dataset corresponding to the two-stage training process described in Section 3.7. For the first stage, we divide each microscopy image into non-overlapping patches and generate a single point prompt within each connected component randomly. The ground truth for each prompt is the mask corresponding solely to that particular connected component, rather than all components in the patch. This aims to minimize the ambiguity of single point prompts as much as possible and facilitates model convergence. For the second stage, we use both the centroid and randomly sampled points as the training prompts. We use patches centered at each prompt to optimize SAM specifically for our method. This stage's dataset comprises 8357 patches of 512 × 512 pixel resolution and 6170 patches of 1024 × 1024 pixel resolution.

4.2 Implementation Details

Model Settings

We use Segformer B0 and B1 [32] as the initial segmentation model and EfficientSAM-S [33] as the segment anything model in our approach. We choose EfficientSAM-S since it is the most efficient variant of SAM that reconstructs the image embeddings of ViT-H [9] in the original SAM. To further improve training efficiency, we integrate adapters into the ViT [9] image encoder following the approach described in Med-SA [31]. We use adapters with an input and output dimensionality of $d_a = 768$ and set the dimensionality of hidden layers to $d_h = 1024$. With adapters, we fine-tune $6.7M$ parameters for EfficientSAM-S to a mere of $29.3M$ parameters.

Training

For Segformer, we use AdamW [19] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a weight decay of 0.01. We use an initial learning rate of $5e-4$ and a polynomial decay scheduler. We use the weights of Segformer pretrained on ImageNet [7] as our starting point and train the model for 150 epochs. For EfficientSAM, we use the Adam optimizer with an initial learning rate of $1e-4$, an exponential decay rate of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a weight decay of 0.05. We use the initial learning rate for the first 10 epochs and apply a learning rate decay factor of 0.5 every 10 epochs. We use a batch size of 8 for patch resolution 512×512, and a batch size of 6 for patch resolution 1024×1024. The model is first trained for 100 epochs on the first-stage dataset and then trained for 150 epochs on the second-stage dataset. The model is trained with a binary cross entropy loss function with equal weights for the invasive melanoma and the backgrounds. All experiments are implemented in PyTorch and executed on 4 NVIDIA Quadro RTX 8000 GPUs.

Inference

To generate the segmentation mask with Segformer, we process each microscopy slide image into patches. Specifically, we use a sliding window to create patches by shifting the window both horizontally and vertically with a step size of 128 pixels. We apply a 2D Gaussian kernel as a weighting mechanism for each pixel within a patch. The kernel has the same size as the patch and a standard deviation of $\frac{1}{4}$ of the patch's side length. To generate the segmentation mask for EfficientSAM, we run our method with a fixed set of hyperparameters. We set the threshold for determining in-situ melanoma $\alpha_m = 0.1$, the probability threshold for high-confidence regions $\beta = 0.8$, and the threshold for excluding low-confidence regions $\alpha_c = 0.4$. Additionally, the threshold for the ratio between the sides of a minimum bounding box α_b, as described in Algorithm 2, is set to 3. The grid prompt employs a vertical and horizontal gap of 64 pixels between neighboring points.

4.3 Results

To evaluate the effectiveness of our proposed method, we compare it with state-of-the-art melanoma segmentation methods on our dataset. As shown in Table 4.1, we compare our method with melanoma segmentation methods including Multi-Scale FCN [23], HRNet & OCR [26], HIPT [30], and Segformer [32]. The results for Segformer B0 and B1 are reproduced following the methods described in Wang et al [30]. All EfficientSAM methods shown in the table use a 1024×1024 patch size. For our method, the second column in Table 4.1 refers to the patch size used for the initial segmentation from Segformer.

As shown in Table 4.1, our model outperforms all other baselines in terms of IoU and F1 score. Compared to using solely the Segformer, which is also used as the initial segmentation model in our method, our method achieves a gain of 20.2% IoU and 13.2% F1 over the state-of-the-art segmentation methods. Table 4.2 shows the number of parameters and tunable

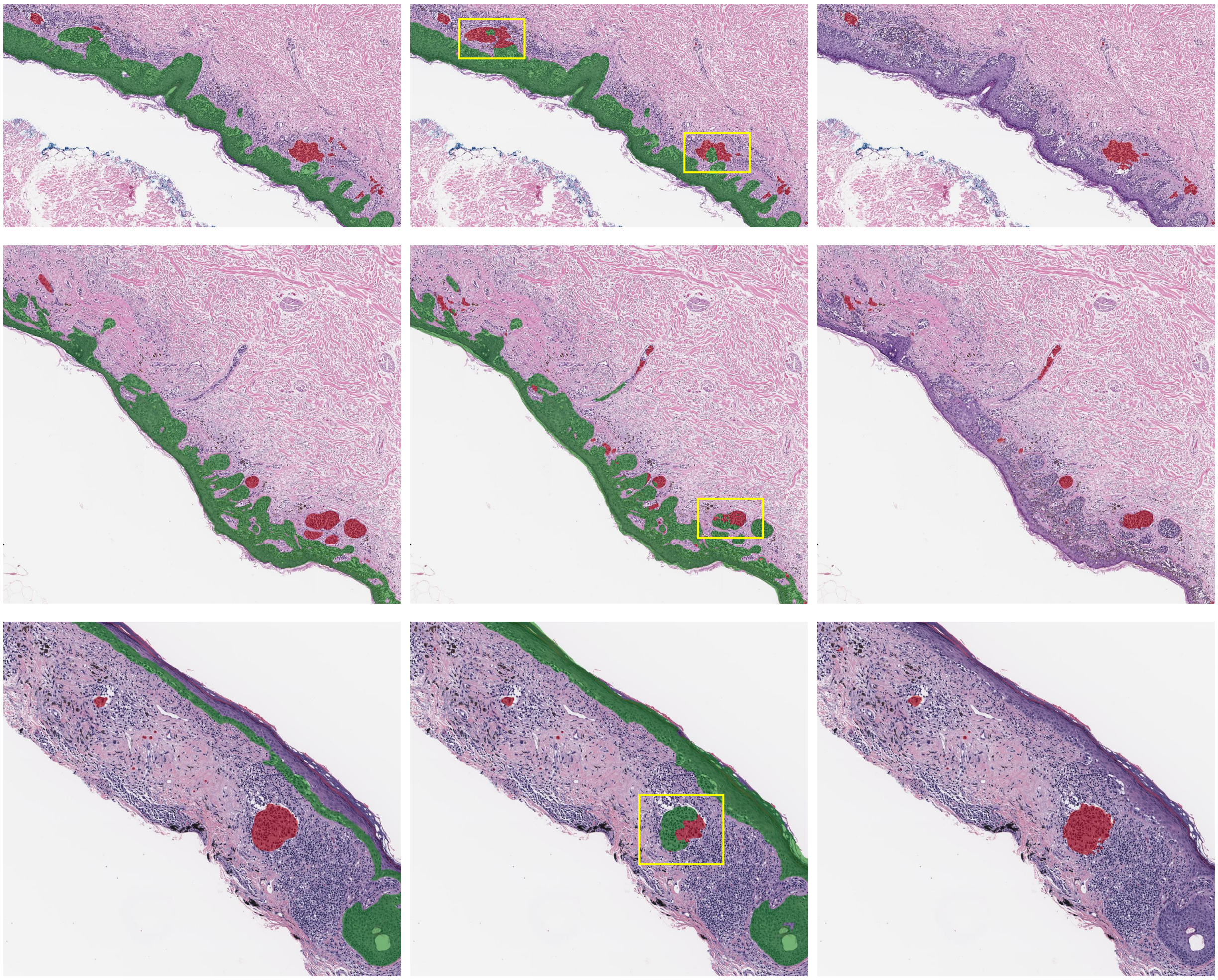
Table 4.1: The results of invasive melanoma segmentation on our dataset. We compare our proposed methods with state-of-the-art segmentation methods on patches with different resolution

Model	**Resolution**	**IoU**	**F1**	**ΔIoU**	**ΔF1**
Multi-Scale FCN [23]	512	13.0	14.0	41.1	56.2
HRNet & OCR [26]	788	29.1	44.0	25.0	26.2
HIPT [30]	512	40.1	57.3	14.0	12.9
Seg. B0 [32]	512	44.6	61.6	9.5	8.6
Seg. B1 [32]	512	42.0	59.2	12.1	11.0
Seg. B0 & EfficientSAM-S (ours)	512	**54.1**	**70.2**	-	-
Seg. B1 & EfficientSAM-S (ours)	512	47.5	64.4	6.6	5.8
HIPT [30]	1024	33.0	46.0	21.1	24.2
Seg. B0 [32]	1024	44.0	61.1	10.1	9.1
Seg. B1 [32]	1024	45.0	62.0	9.1	8.2
Seg. B0 & EfficientSAM-S (ours)	1024	**49.5**	**66.2**	4.6	4.0
Seg. B1 & EfficientSAM-S (ours)	1024	49.4	66.1	4.7	4.1

Table 4.2: The number of parameters for models used in our method.

Model	**Params (M)**	**Tunable Params (M)**
Seg. B0	3.7	3.7
Seg. B1	13.7	13.7
EfficientSAM-S	29.3	6.7

parameters for models in our method. Even though EfficientSAM-S has more parameters than the Segformers, we only fine tune approximately 23.1% of its total parameters, which is less than half of the parameters of Segformer B1. Figure 4.1 presents representative qualitative results on our dataset, where our method significantly reduces errors in in-situ melanoma regions and enhances segmentation accuracy in areas where distinguishing between invasive melanoma and the epidermis is particularly challenging. Additinoally, the best result is achieved with using EfficientSAM and Segformer B0 with 512 × 512 patches, even though this is not the best performing Segformer in terms of IoU and F1 score. A higher IoU for the initial segmentation mask does not necessarily indicate more improvement in IoU in the final mask of our method. The amount of improvement depends on the quality of sampled prompts and the room left for improvement for each prompt.



(a) Ground Truth

(b) Segformer-B0

(c) Ours

Figure 4.1: Qualitative results on our dataset. Red denotes invasive melanoma and green denotes the epidermis. Compared to Segformer, our method significantly improves the accuracy of predictions for invasive melanoma regions, especially in areas where distinguishing from the epidermis is challenging.

4.4 Ablation Studies

We now evaluate the effectiveness of our proposed method through a series of ablation studies on different components of our method, including the patch size, in-situ melanoma detection, low-confidence region filtering, the prompt determination strategy and the final mask generation. All ablation studies use a patch size of 1024 × 1024 for EfficientSAM if not specified.

Patch Size

Table 4.3: Ablation study on the effect of patch size for EfficientSAM.

EfficientSAM Patch Size	**Initial Segmentaion Model**	**Initial Model Patch Size**	**IoU**	**ΔIoU**
512	Seg. B0	512	49.4	**4.7**
1024	Seg. B0	512	**54.1**	
512	Seg. B0	1024	46.2	3.3
1024	Seg. B0	1024	**49.5**	
512	Seg. B1	512	44.9	2.6
1024	Seg. B1	512	**47.5**	
512	Seg. B1	1024	46.8	2.6
1024	Seg. B1	1024	**49.4**	

We study the impact of patch size on the performance of EfficientSAM by comparing the model's performance on 512 × 512 patches and 1024 × 1024 patches. It is important to note that the patch size used by the initial segmentation model can differ from the patch size used by SAM. While the initial segmentation model's patch size influences the quality of the prompts, EfficientSAM's patch size affects the amount of context available to the model, impacting the quality of the image embeddings generated by the ViT encoder. As shown in Table 4.3, using 1024 × 1024 patches for EfficientSAM consistently achieves better results than using 512 × 512 patches. The gains in IoU range from 5.6% to 9.5%. This demonstrates that EfficienetSAM significantly benefits from a larger patch size.

In-situ Melanoma Detection

We investigate the impact of in-situ melanoma detection described in Section 3.3 by filtering out all detected estimated in-situ melanoma regions regardless of their confidence levels. Table 4.4 shows that this consistently improves performance, with IoU gains ranging from 1.1% to 27.8%. The improvement varies based on the quality of the initial segmentation mask and the room left for improvement. In contrast, disabling in-situ melanoma detection results in a performance drop compared to the original mask produced by Segformer. This drop occurs because prompts generated from incorrect predictions lead SAM to make additional errors. Therefore, enabling in-situ melanoma detection not only mitigates errors from incorrect initial predictions but also leverages the strengths of the remaining high-confidence predictions to improve the final segmentation results.

Table 4.4: Ablation study on in-situ melanoma detection and low-confidence region filtering in preparation for prompt generation, as described in Sections 3.3 and 3.4.

Initial Model	**Segformer Patch Size**	**In-situ Melanoma Detection**	**Low- Confidence Region Filtering**	**IoU**	**IoU Gain**
Seg. B0	512			42.4	-
Seg. B0	512	✓		**54.2**	**11.8**
Seg. B0	512	✓	✓	54.1	11.7
Seg. B0	1024			41.4	-
Seg. B0	1024	✓		45.6	4.2
Seg. B0	1024	✓	✓	**49.5**	**8.1**
Seg. B1	512			44.1	-
Seg. B1	512	✓		47.5	3.4
Seg. B1	512	✓	✓	**47.5**	**3.4**
Seg. B1	1024			44.3	-
Seg. B1	1024	✓		44.8	0.5
Seg. B1	1024	✓	✓	**49.4**	**5.1**

Low-Confidence Region Filtering

We study the impact of low-confidence region filtering presented in Section 3.4. As shown in Table 4.4, our method achieves a 8.5% IoU gain for Segformer B0 and a 10.3% IoU gain for Segformer B1 when using a 1024 × 1024 patch size. However, there is a slight drop of 0.02% in IoU for Segformer B0 with a 512 × 512 patch size. This suggests that some high-confidence invasive melanoma regions that touch the epidermis predictions, which are not filtered by the algorithm, are actually in-situ melanoma. Overall, this demonstrates that low-confidence region filtering improves the robustness of our segmentation results.

Prompt Types

We study the effectiveness of different prompt types. We test prompts with centroid alone, grid alone and our proposed strategy that uses both as presented in Section 3.5. Table 4.5 shows that in most cases using both achieves the highest IoU in the final mask. Compared to grid prompts alone, using both achieves a gain as high as 8.1% IoU. Compared to centroid prompts alone, using both achieves gains up to 4.4% IoU with one case showing no improvement. It is noticeable that centroid prompts alone outperforms grid prompts alone in the final mask, but performs much worse in pre-merged mask. This shows that centroid prompts allow more accurate segmentation and thus complement the initial mask well, achieving high IoU in the final mask after merging. In contrast, grid prompts alone achieve full coverage over the initial mask, but not all points serve as effective prompts for accurate segmentation. This demonstrates that our method effectively leverages the strengths of both prompts

Table 4.5: Ablation study on the effects of using different prompts. ”Both” denotes the method that dynamically uses both centroids and grid prompts based on the shape distributions of melanoma components, as shown in Section 3.5.

Initial Model	**Prompt**	**Segformer Patch Size**	**EfficientSAM IoU (%)**	**Final Mask IoU (%)**	**Improvement in IoU (%)**
Seg. B0	Centroid	512	43.3	53.1	**9.8**
Seg. B0	Grid	512	50.6	51.7	1.1
Seg. B0	Both	512	**52.2**	**54.1**	1.9
Seg. B1	Centroid	512	44.3	47.5	**3.2**
Seg. B1	Grid	512	**46.4**	46.5	0.1
Seg. B1	Both	512	46.1	**47.5**	1.4
Seg. B0	Centroid	1024	40.8	48.5	**7.7**
Seg. B0	Grid	1024	45.5	45.8	0.3
Seg. B0	Both	1024	**48.7**	**49.5**	0.8
Seg. B1	Centroid	1024	37.7	47.3	**9.6**
Seg. B1	Grid	1024	46.0	47.2	1.2
Seg. B1	Both	1024	**47.1**	**49.4**	2.3

Table 4.6: Ablation study on final mask generation. The last row represents the result when prompts are generated from the ground truth instead of the mask produced by Segformer.

Initial Model	**Segformer Patch Size**	**Post-processed $\hat{X^m}$ IoU**	**EfficientSAM $\hat{Y^m}$ IoU**	**Final IoU**	$\Delta\hat{\boldsymbol{X}^m}$	$\Delta\hat{\boldsymbol{Y}^m}$
Seg. B0	512	48.6	52.2	**54.1**	**5.5**	1.9
Seg. B1	512	43.4	46.1	**47.5**	4.1	1.4
Seg. B0	1024	46.7	48.7	**49.5**	2.8	0.8
Seg. B1	1024	47.2	47.1	**49.4**	2.2	**2.3**
GT	-	-	63.0	-	-	-

by using grid prompts for large, irregular melanoma components and centroid prompts for small, regular melanoma components, maximizing the advantages of each prompt type.

Final Mask Generation

We study the impact of final mask generation. As shown in Table 4.6, the final mask consistently achieves the highest IoU in all cases. EfficientSAM's mask outperforms the post-prococessed mask in most cases, with the most significant improvement being a 11.3% increase for Segformer B0 using 512 × 512 patches. Merging two masks into a final mask significantly enhances accuracy. The post-processed mask shows gains ranging from 4.7%

to 11.3% IoU, while the EfficientSAM's mask gains improvements between 1.6% and 4.9% IoU. This demonstrates that EfficientSAM compliments well the post-processsed mask in generating the final mask. In addition, we evaluate the EfficientSAM's performance with prompts generated from the ground truth. Its IoU performance is 20.7% higher than the best result of using EfficientSAM alone and 16.5% higher than the best final mask. This demonstrates the upper limit of the performance of our method when using fully accurate prompts.

Chapter 5

Conclusion

We proposed a novel approach to explore the potential of SAM for melanoma segmentation in microscopy slide images. Our method utilizes Segformer to generate initial segmentation masks and subsequently prompts EfficientSAM using a dynamic selection of centroid prompts and grid prompts for automatic invasive melanoma segmentation. To ensure accurate prompt generation, we implement in-situ melanoma detection and filter out low-confidence regions. Additionally, we integrate adapters into EfficientSAM for parameter-efficient fine-tuning. Our experimental results demonstrate that this approach not only surpasses other state-of-the-art melanoma segmentation methods but also significantly improves upon the baseline performance of Segformer. Furthermore, we conduct comprehensive ablation studies to validate the effectiveness of each key component in our method.

Beyond melanoma segmentation, our framework presents a versatile approach for applying SAM to automatic semantic segmentation of various objects. Based on the proposed method, both the initial segmentation model and the segment anything model can be customized for different applications by fine-tuning on specific datasets and designing specialized prompt generation strategies. This flexibility allows for the integration of any semantic segmentation and interactive segmentation models into our framework for customized usage. Furthermore, while our current implementation primarily focus on utilizing single point prompts, we anticipate future work to explore additional prompt types supported by SAM, such as multiple point prompts, box prompts and mask prompts. This expansion would enhance the versatility and accuracy of our framework across different segmentation tasks. Lastly, our method addresses the challenge of segmenting ultra-high-resolution microscopy slide images by processing them in patches. We expect future work to handle slide images by incorporating global context beyond individual patches and thus achieve more precise and comprehensive medical image segmentation in high-resolution imaging scenarios.

Bibliography

[1] Martin D Abeloff et al. *Abeloff's Clinical Oncology E-Book*. Elsevier Health Sciences, 2008.

[2] Salah Alheejawi et al. "Deep learning-based histopathological image analysis for automated detection and staging of melanoma". In: *Deep Learning Techniques for Biomedical and Health Informatics*. Elsevier, 2020, pp. 237–265.

[3] American Cancer Society. *Cancer Facts and Figures 2024*. Accessed: 2024-06-08. 2024. URL: `https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2024-cancer-facts-figures.html`.

[4] Pablo Arbelaez et al. "Contour detection and hierarchical image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 33.5 (2010), pp. 898–916.

[5] Shoufa Chen et al. "Adaptformer: Adapting vision transformers for scalable visual recognition". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 16664–16678.

[6] Ruzica Z Conic et al. "Determination of the impact of melanoma surgical timing on survival using the National Cancer Database". In: *Journal of the American Academy of Dermatology* 78.1 (2018), pp. 40–46.

[7] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[8] Ruining Deng et al. "Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging". In: *arXiv preprint arXiv:2304.04155* (2023).

[9] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[10] Andrew H Fischer et al. "Hematoxylin and eosin staining of tissue and cell sections". In: *Cold spring harbor protocols* 2008.5 (2008), pdb–prot4986.

[11] Agnessa Gadeliya Goodson and Douglas Grossman. "Strategies for early melanoma detection: Approaches to the patient with nevi". In: *Journal of the American Academy of Dermatology* 60.5 (2009), pp. 719–735.

[12] Kaiming He et al. “Masked autoencoders are scalable vision learners”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.

[13] Xuehai He et al. “Parameter-efficient fine-tuning for vision transformers”. In: *arXiv preprint arXiv:2203.16329* 3 (2022).

[14] Edward J Hu et al. “Lora: Low-rank adaptation of large language models”. In: *arXiv preprint arXiv:2106.09685* (2021).

[15] Yuhao Huang et al. “Segment anything model for medical images?” In: *Medical Image Analysis* 92 (2024), p. 103061.

[16] Chao Jia et al. “Scaling up visual and vision-language representation learning with noisy text supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 4904–4916.

[17] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. *Yolo by ultralytics*. 2023. URL: `https://github.com/ultralytics/ultralytics`.

[18] Alexander Kirillov et al. “Segment anything”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[19] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[20] Jun Ma et al. “Segment anything in medical images”. In: *Nature Communications* 15.1 (2024), p. 654.

[21] Shima Nofallah 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.

[22] Kay RJ Oskal et al. “A U-net based approach to epidermal tissue segmentation in whole slide histopathological images”. In: *SN Applied Sciences* 1 (2019), pp. 1–12.

[23] 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 workshops*. 2019.

[24] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[25] S Roy et al. “Zero-shot medical image segmentation capabilities of the Segment Anything Model. arXiv 2023”. In: *arXiv preprint arXiv:2304.05396* ().

[26] Aman Shah et al. “Deep learning segmentation of invasive melanoma”. In: *Medical Imaging 2023: Digital and Computational Pathology*. Vol. 12471. SPIE. 2023, pp. 447–452.

[27] Skin Cancer Foundation. *Melanoma Overview: A Dangerous Skin Cancer*. Accessed: 2024-06-08. 2024. URL: `https://www.skincancer.org/skin-cancer-information/melanoma`.

[28] Koen EA Van de Sande et al. “Segmentation as selective search for object recognition”. In: *2011 international conference on computer vision*. IEEE. 2011, pp. 1879–1886.

[29] Mike Van Zon et al. “Segmentation and classification of melanoma and nevus in whole slide images”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 263–266.

[30] Franklin Wang et al. “Transformers for Microscopy Slide Image Segmentation of Invasive Melanoma”. In: *Electronic Imaging* 36 (2024), pp. 1–5.

[31] Junde Wu et al. “Medical sam adapter: Adapting segment anything model for medical image segmentation”. In: *arXiv preprint arXiv:2304.12620* (2023).

[32] Enze Xie et al. “SegFormer: Simple and efficient design for semantic segmentation with transformers”. In: *Advances in neural information processing systems* 34 (2021), pp. 12077–12090.

[33] Yunyang Xiong et al. “Efficientsam: Leveraged masked image pretraining for efficient segment anything”. In: *arXiv preprint arXiv:2312.00863* (2023).

[34] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. “Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models”. In: *arXiv preprint arXiv:2106.10199* (2021).

[35] Chaoning Zhang et al. “Faster segment anything: Towards lightweight sam for mobile applications”. In: *arXiv preprint arXiv:2306.14289* (2023).

[36] Yizhe Zhang et al. “Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 129–139.

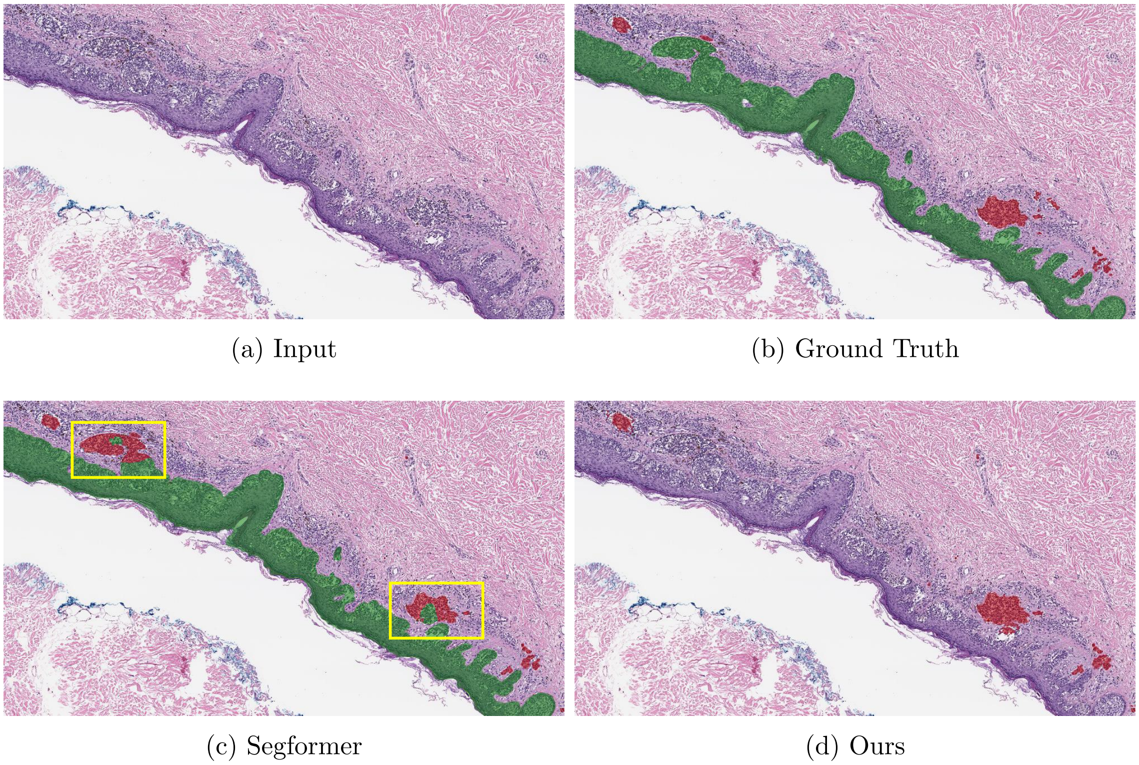
[37] Xu Zhao et al. “Fast segment anything”. In: *arXiv preprint arXiv:2306.12156* (2023).

[38] Tao Zhou et al. “Can sam segment polyps?” In: *arXiv preprint arXiv:2304.07583* (2023).

Appendix A

More Qualitative Results

We present more qualitative results on our datasest by including the input image, the ground truth, the Segformer output, and the output of our method for comparison. The results shown are produced by Segformer B0 with 512×512 patches and EfficientSAM with 1024×1024 patches.



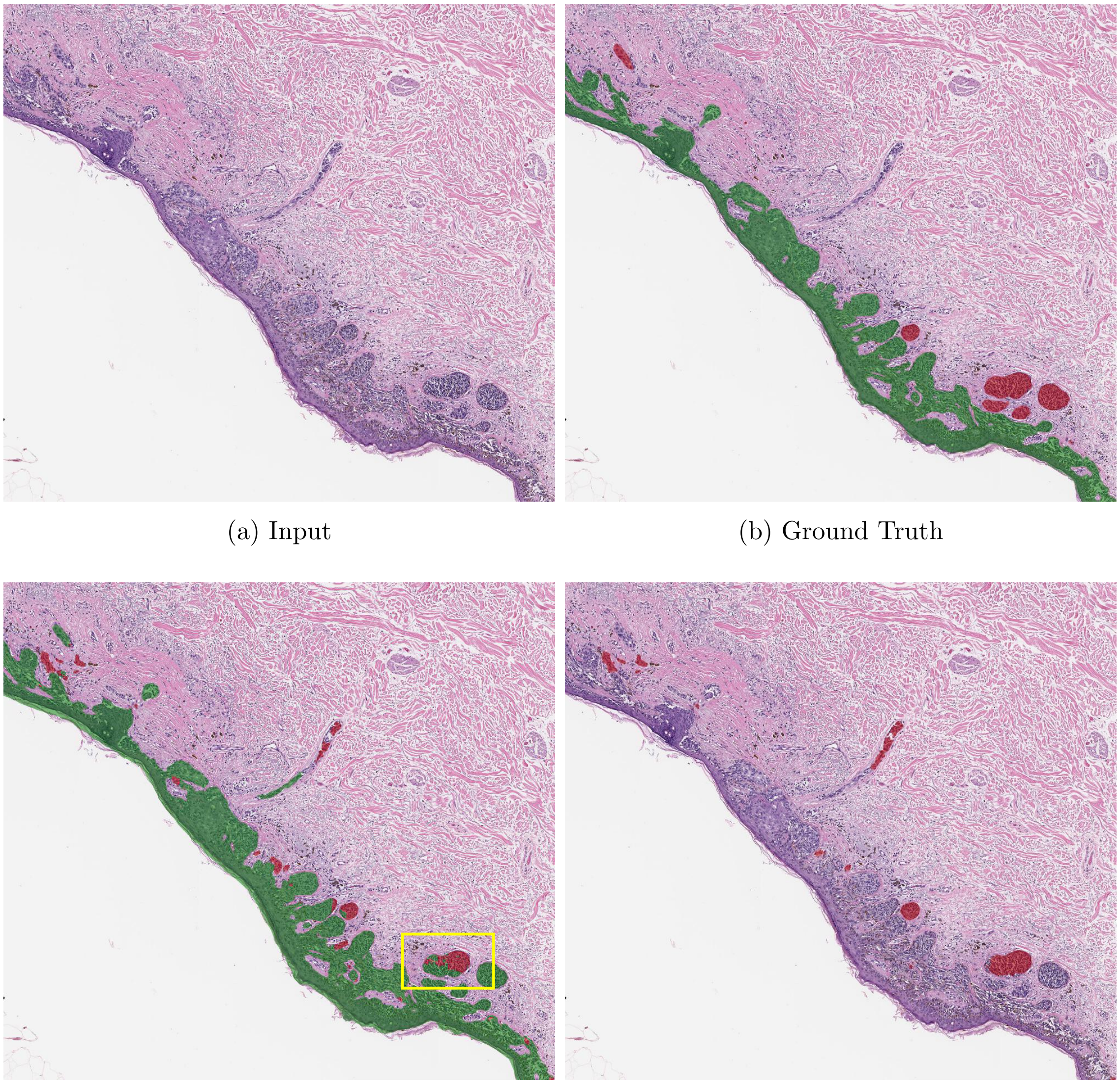
(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.1: Our method significantly refines the initial segmentation mask and reduce incorrect predictions made by Segformer.



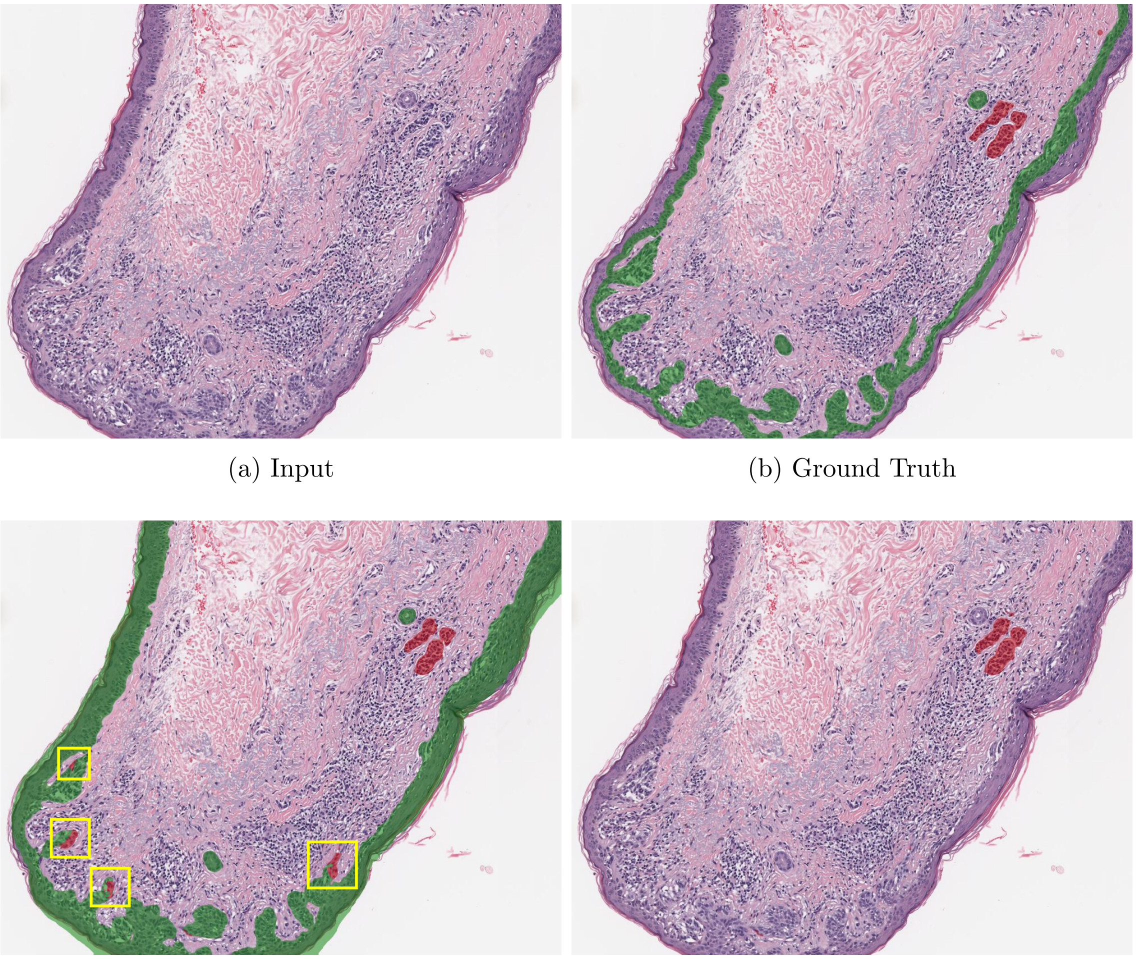
(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.2: Our method effectively reduce errors of predictions in in-situ melanoma regions.



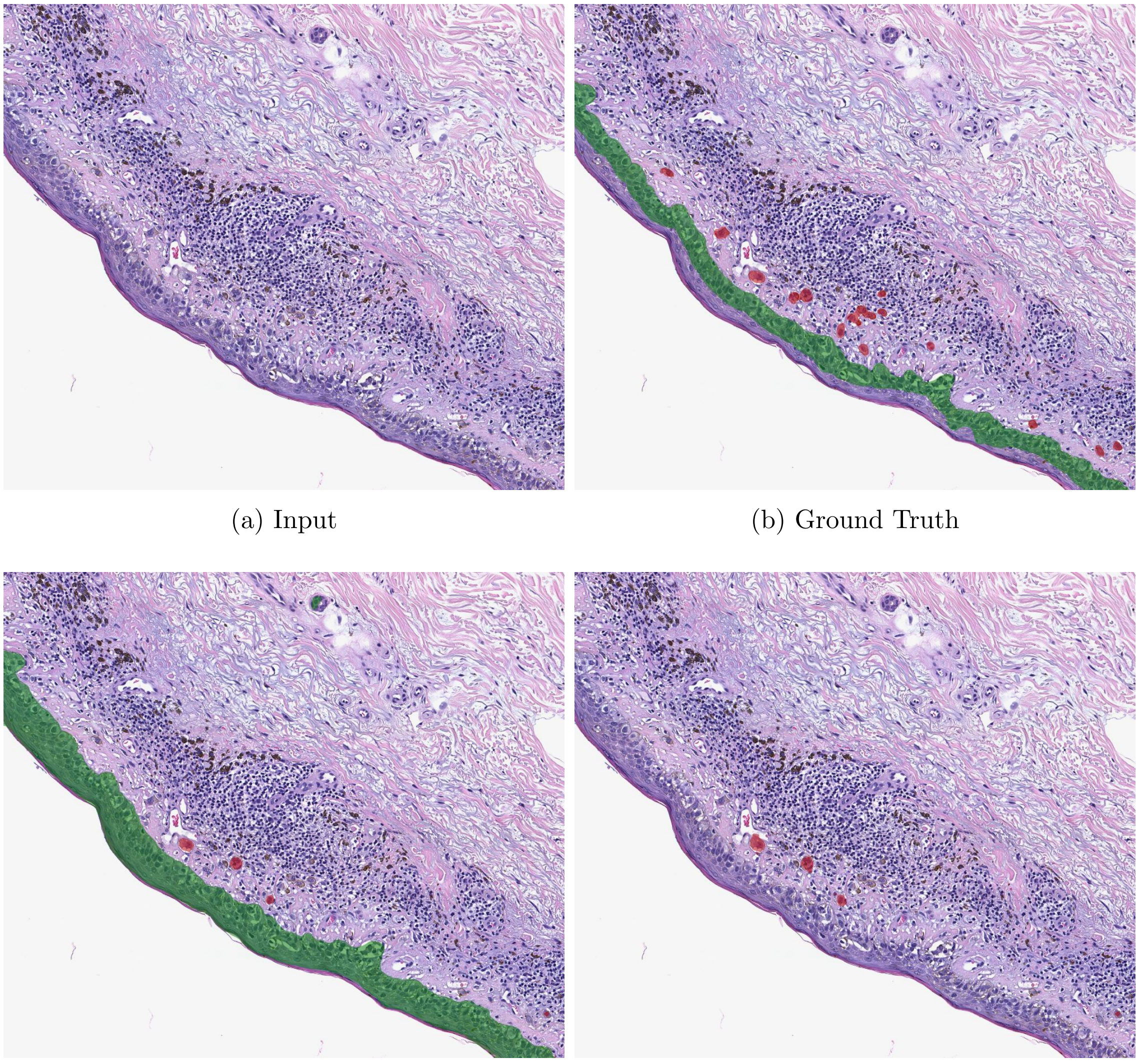
(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.3: Our method effectively avoids incorrect predictions in the epidermis.



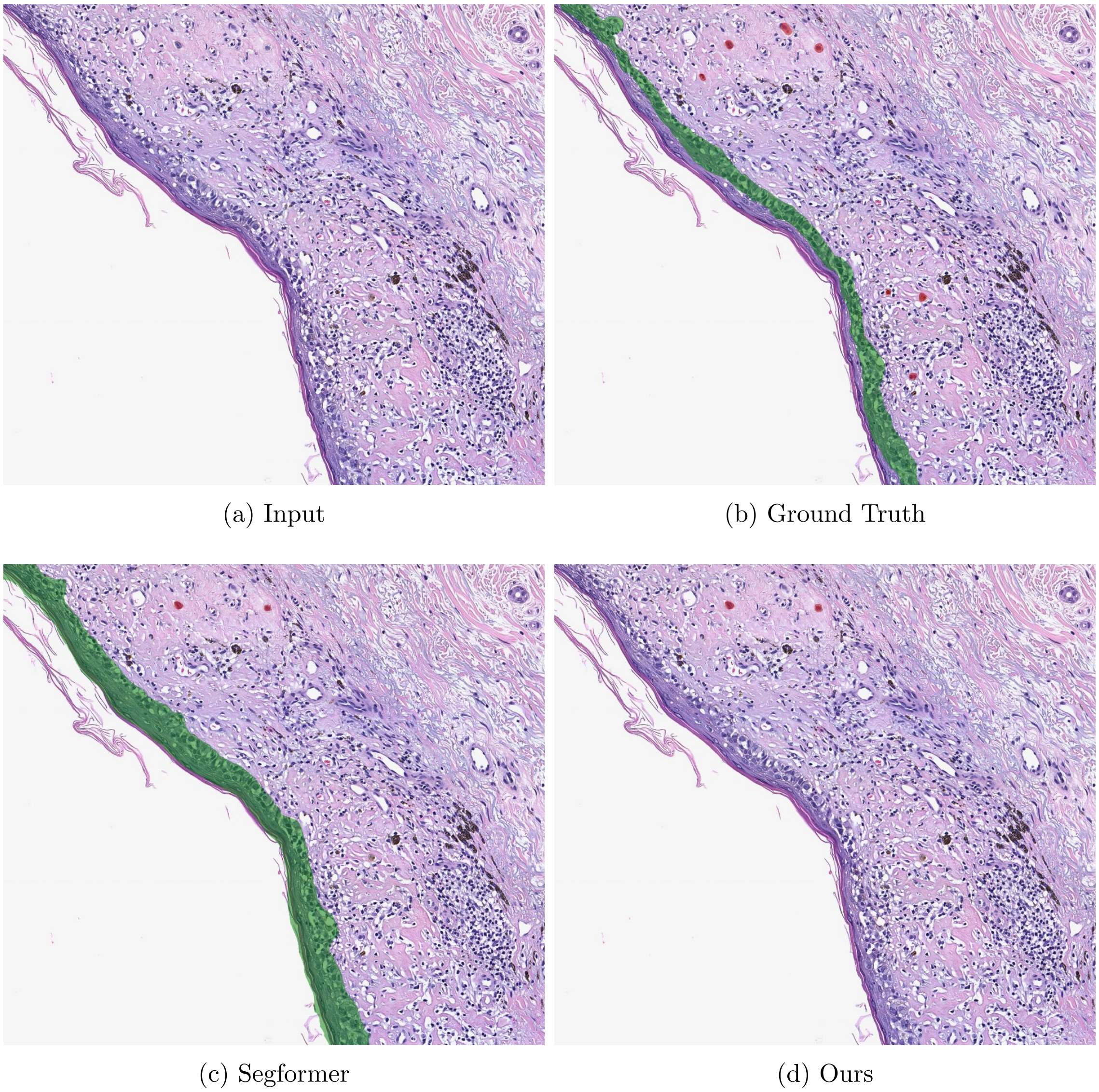
(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.4: Our method makes predictions similar to Segformer.



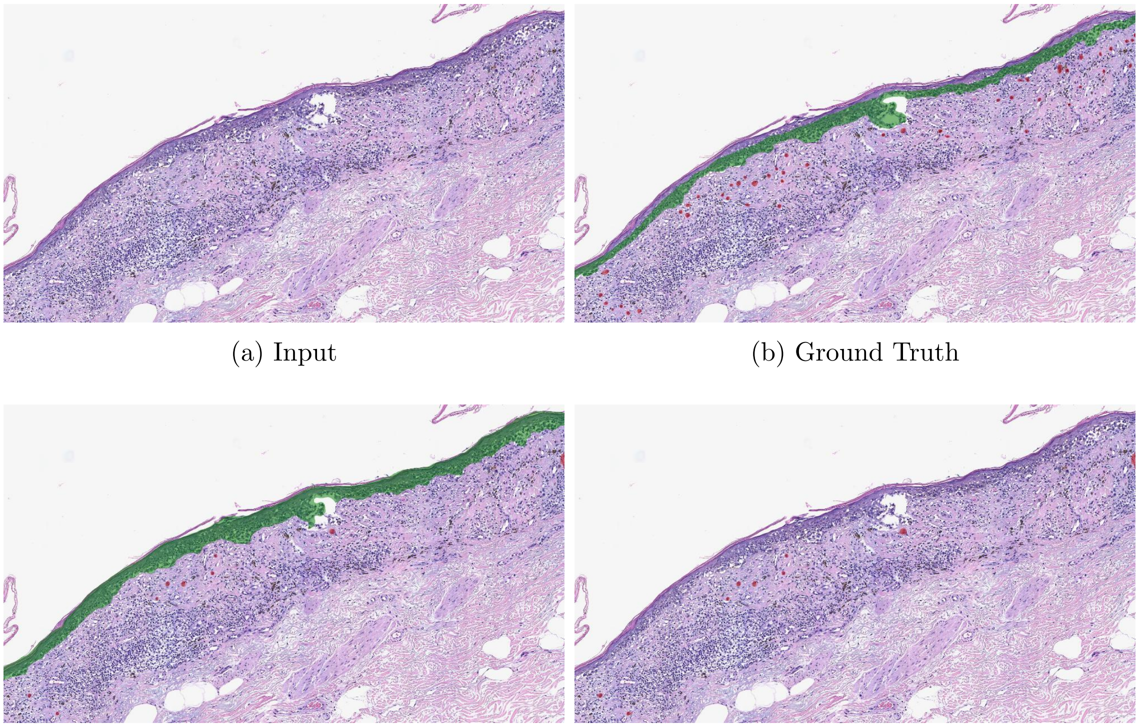
(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.5: Our method makes predictions similar to Segformer.



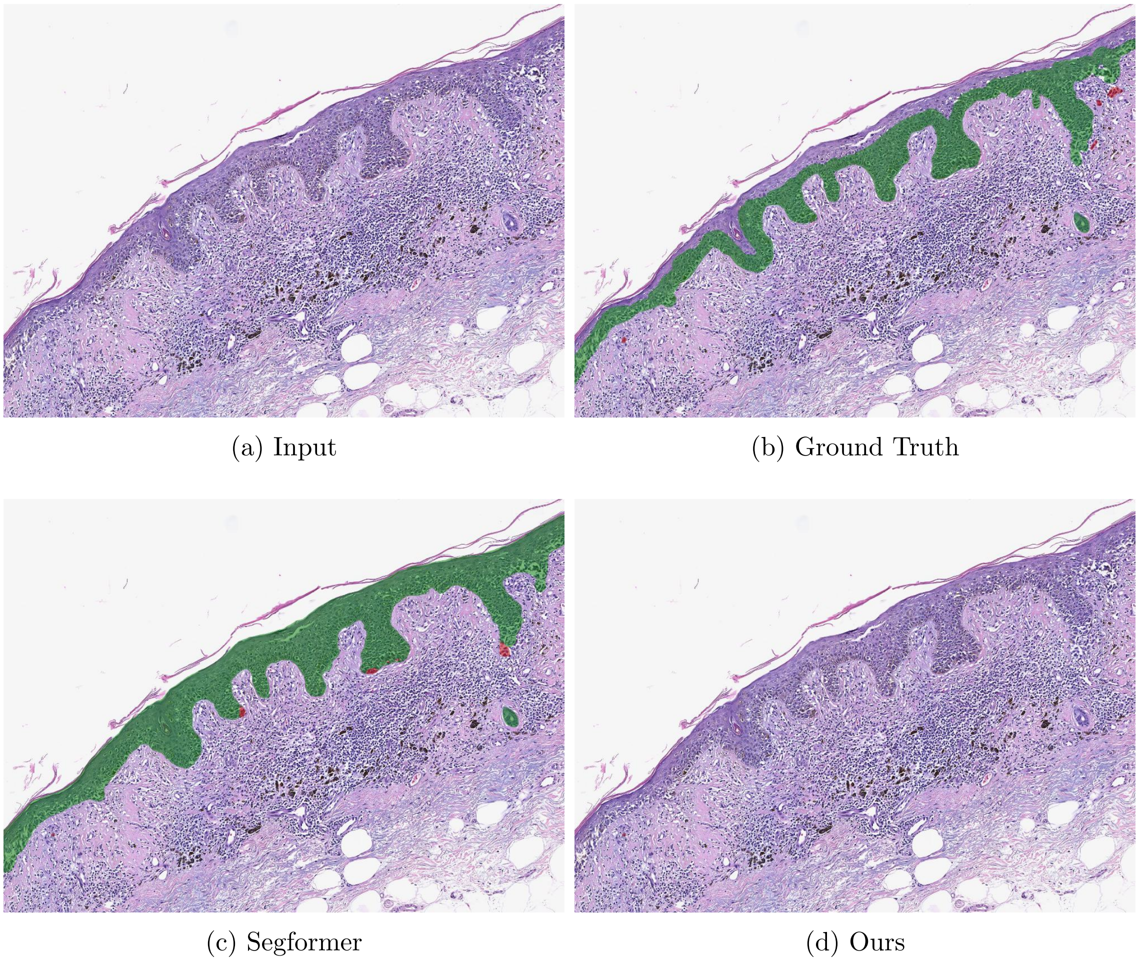
(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.6: Our method makes predictions similar to Segformer.

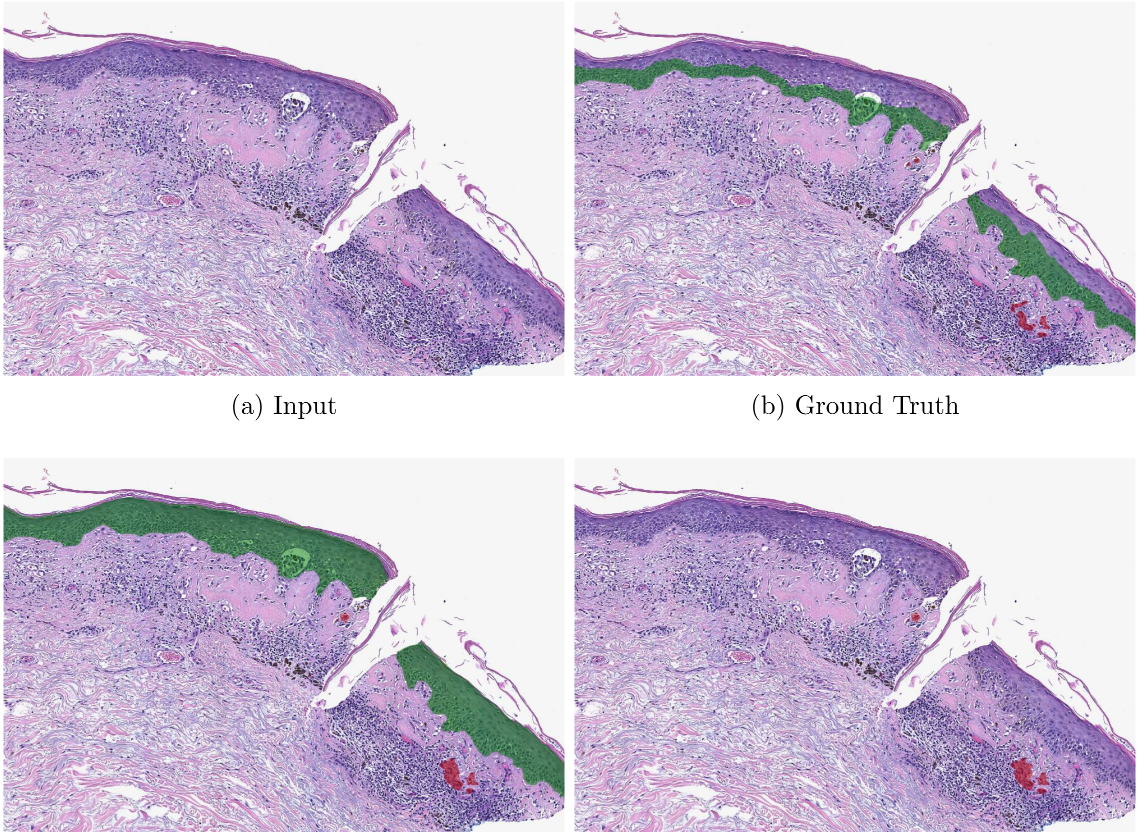

(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.7: Our method effectively ignores low-confidence invasive melanoma regions predicted by Segformer.



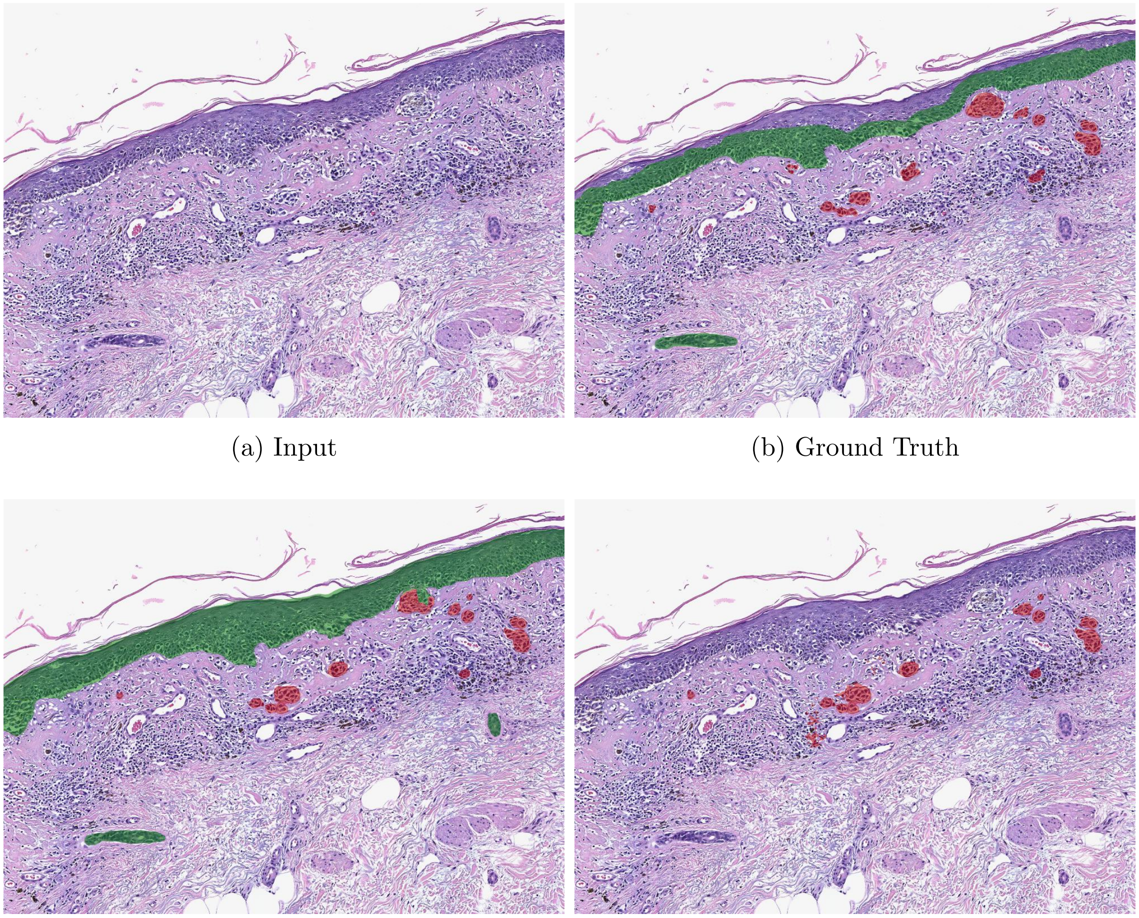
(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.8: Our method makes predictions similar to Segformer.



(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.9: There is one invasive melanoma component missed in our prediction. It is filtered out due to the low confidence level predicted by Segformer.

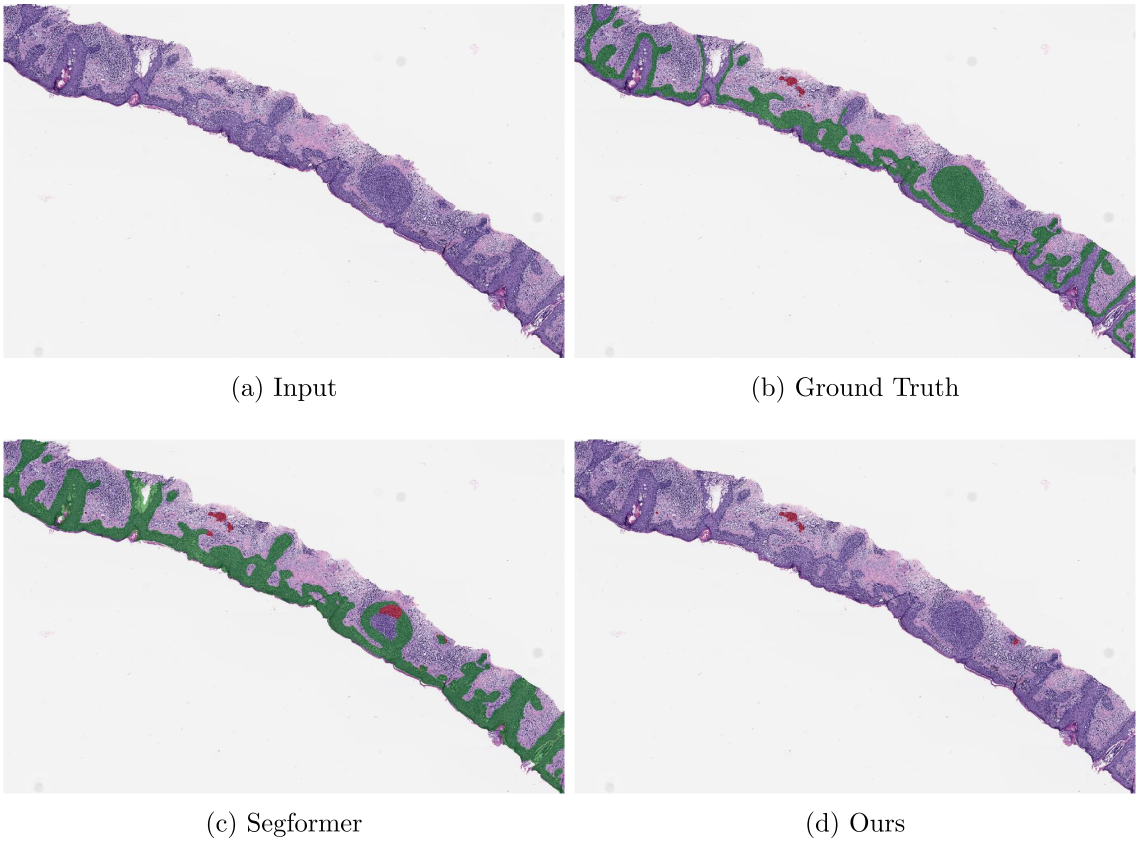
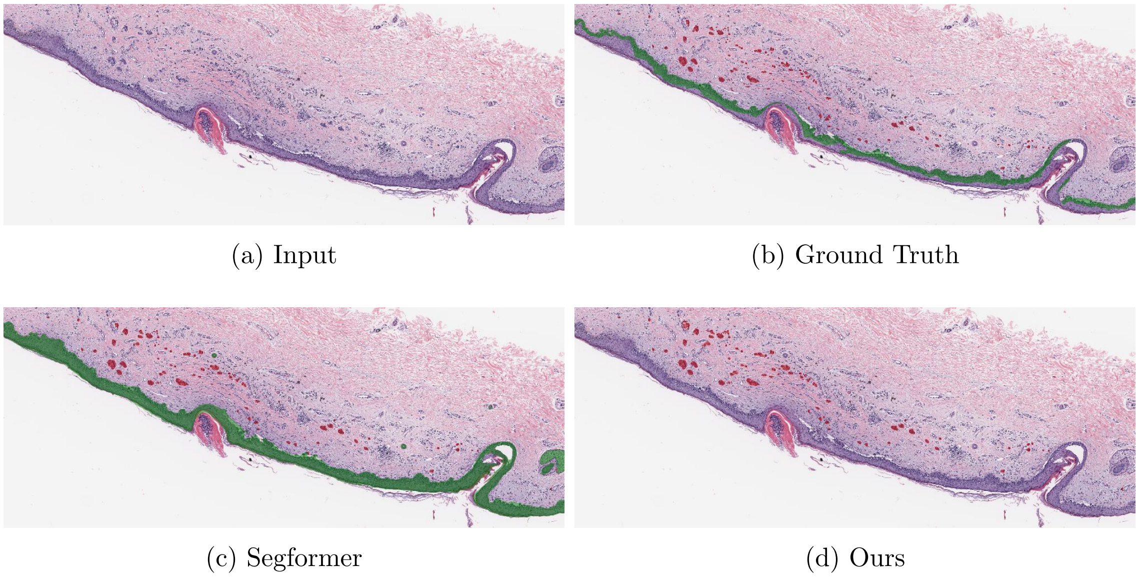


Figure A.10: Our method effectively reduces incorrect predictions for invasive melanoma in the epidermis.



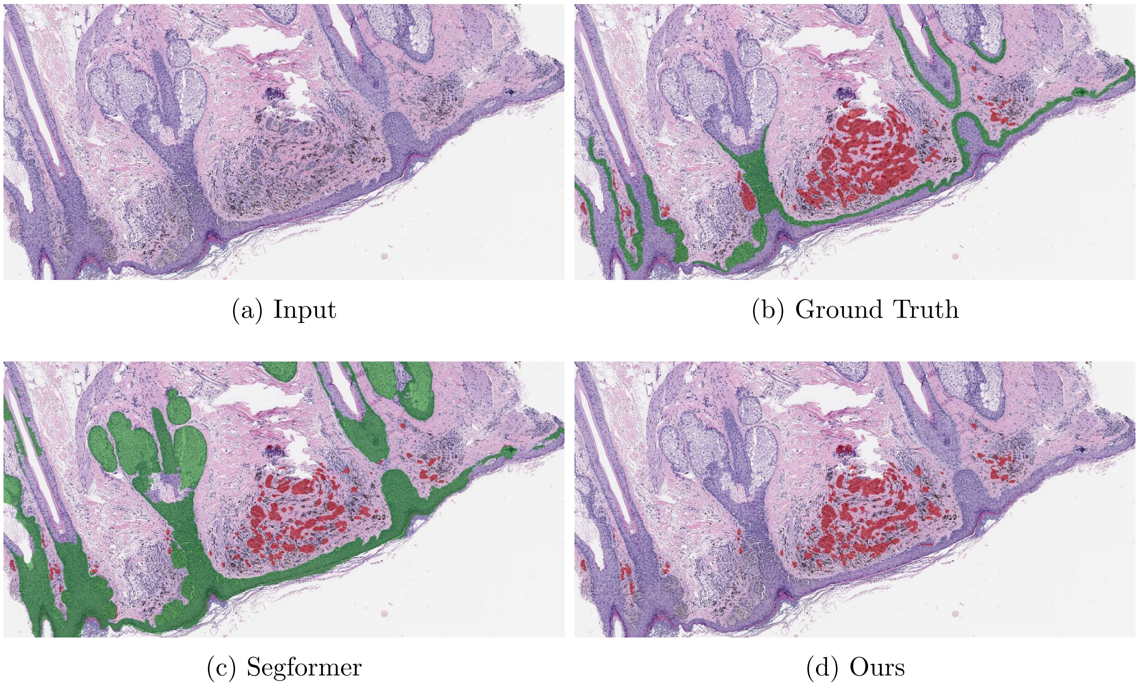
(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.11: Our method predicts more accurate boundaries for small regular melanoma than Segformer.



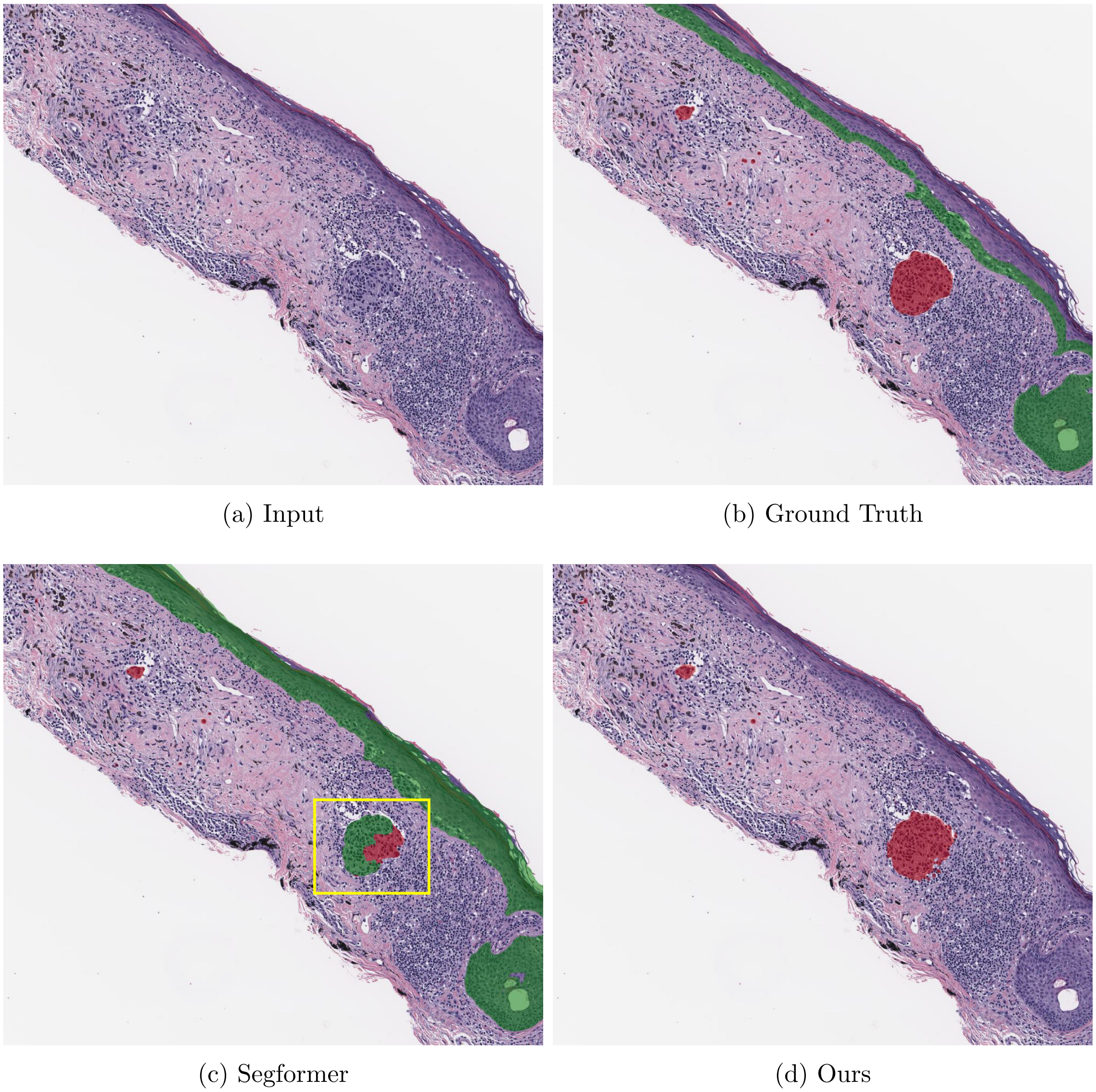
(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.12: Our method performs slightly better than Segformer for large melanoma clusters.



(a) Input

(b) Ground Truth

(c) Segformer

(d) Ours

Figure A.13: Our method significantly improves upon predictions made by Segformer.