Journal of Digital Imaging
https://doi.org/10.1007/s10278-022-00641-8

ORIGINAL PAPER

Segmenting Skin Biopsy Images with Coarse and Sparse Annotations using U-Net

Shima Nofallah[1] · Mojgan Mokhtari[2] · Wenjun Wu[1] · Sachin Mehta[1] · Stevan Knezevich[3] · Caitlin J. May[4] · Oliver H. Chang[1] · Annie C. Lee[5] · Joann G. Elmore[5] · Linda G. Shapiro[1]

Received: 3 May 2021 / Revised: 11 February 2022 / Accepted: 15 April 2022

Abstract
The number of melanoma diagnoses has increased dramatically over the past three decades, outpacing almost all other cancers. Nearly 1 in 4 skin biopsies is of melanocytic lesions, highlighting the clinical and public health importance of correct diagnosis. Deep learning image analysis methods may improve and complement current diagnostic and prognostic capabilities. The histologic evaluation of melanocytic lesions, including melanoma and its precursors, involves determining whether the melanocytic population involves the epidermis, dermis, or both. Semantic segmentation of clinically important structures in skin biopsies is a crucial step towards an accurate diagnosis. While training a segmentation model requires ground-truth labels, annotation of large images is a labor-intensive task. This issue becomes especially pronounced in a medical image dataset in which expert annotation is the gold standard. In this paper, we propose a two-stage segmentation pipeline using coarse and sparse annotations on a small region of the whole slide image as the training set. Segmentation results on whole slide images show promising performance for the proposed pipeline.

Keywords Semantic segmentation · Dermatology · Whole slide imaging · Sparse annotation · Skin biopsy · Invasive melanoma

Introduction

The incidence of melanoma is rising faster than any other cancer [1–3]. The current gold standard for melanoma diagnosis is the microscopic examination of skin biopsies using hematoxylin and eosin (H&E) stained tissue sections; however, the histologic interpretation of melanocytic lesions is often inconsistent for pathologists. Our research team has highlighted these challenges by demonstrating that pathologists disagree on up to 60% of cases of melanoma in situ and T1a invasive melanoma, which can lead to both overtreatment and undertreatment [4]. Researchers have shown that automated diagnosis holds promise for improving accuracy and reproducibility in the diagnosis of histopathology [5–7]. The histologic evaluation of melanocytic lesions, including melanoma and its precursors, involves determining whether the melanocytic population involves the epidermis, dermis, or both. For example, the atypical melanocytes in melanoma in situ are contained within the epidermis, whereas an invasive melanoma shows atypical melanocytes which in the dermis. Semantic segmentation of various structures in skin biopsy images, including accurately distinguishing between the epidermis/dermis and identifying epidermal/dermal melanocytes, has the potential to improve the automated diagnosis systems or serve as a diagnostic aid in the decision-making process. The goal of semantic segmentation is to label each pixel of an image with the corresponding class of the objects being represented. Hence, semantic segmentation of clinically relevant structures in skin biopsy images can play a key role in an automated diagnosis system.

One key challenge in training a segmentation model is that it requires large-scale and fine annotations. However,

✉ Shima Nofallah
shimz@uw.edu

1 University of Washington, Seattle, WA 98195, USA

2 Pathology Department, Isfahan University of Medical Sciences, Isfahan, Iran

3 Pathology Associates, Clovis, CA 983611, USA

4 Dermatopathology Northwest, Bellevue, WA 98005, USA

5 David Geffen School of Medicine, UCLA, Los Angeles, CA 90024, USA

Published online: 02 May 2022

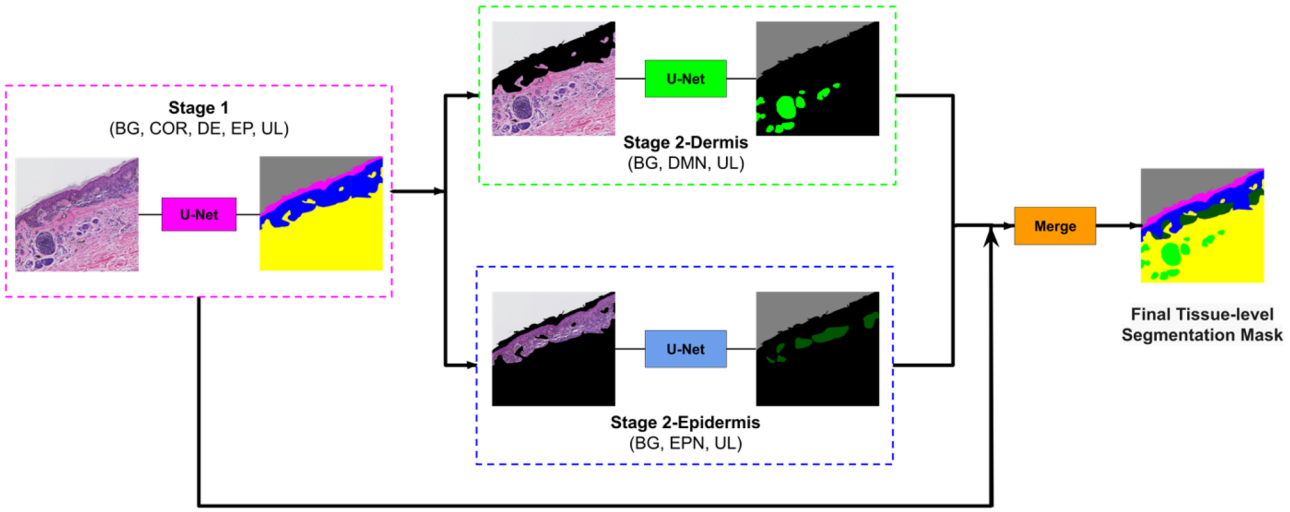


Fig. 1 Overview of our approach. The image first goes to stage 1, and the segmentation mask of entities (COR, stratum corneum; EP, epidermis; DE, dermis; BG, background; and UL, unlabeled) in stage 1 is generated. Then this mask is used to remove the epidermis from stage 2-Dermis input and remove the dermis from stage 2-Epidermis input. The modified images are fed to their corresponding trained model. Stage 2-Dermis generates the segmentation masks of entities present in the dermis (DMN, dermal nests), and stage 2-Epidermis generates the entities in the epidermis (EPN, epidermal nests). In the end, stage 2-Dermis and stage 2-Epidermis segmentation masks are overlaid on the stage 1 mask, and the final tissue-level segmentation mask is generated

collecting fine tissue-level annotations for biopsy images is an onerous, exhaustive, and expensive task because of the sheer size of biopsy images and the fact that domain experts are required for annotations. As a result, full annotation of the whole slide image (WSI) for large datasets is the leading limitation of medical imaging research. This work introduces a simple two-step approach for learning representations with coarse and sparse labels. The overview of our approach is shown in Fig. 1. The core principle is to segment larger and smaller entities separately, allowing us to segment images with good accuracy.

Related Work

Various approaches have been developed to overcome imperfect and limited data annotation and vary with the specific challenges posed by the specific dataset on which they were developed.

When a small portion of an image is fully annotated, different methods of augmentation have proven to be helpful. [8] showed that data augmentation by adjusting image quality produces performance gain in magnetic resonance imaging (MRI), especially image sharpening through the application of unsharp masking, which has the largest improvement. In another study, [9] proposed asymmetric mixup that turns soft labels generated by mixup into hard labels, which improves the segmentation of brain tumors according to their experiments.

Active learning is another popular method in the case of limited annotation. [10] proposed a probabilistic active learning pipeline where the probability of an unlabeled sample that is queried in the next round of annotation is estimated based on its Fisher information. [11] used a Bayesian neural network for active learning: using a combined metric based on noise in the data and uncertainty over their convolutional neural network (CNN) parameters, they selected the most informative samples. [12] proposed a one-shot active learning method, which eliminates the need for iterative sample selection and annotation. However, active learning generally requires a base segmentation model with careful annotation; hence, a dataset with only coarse annotations may not benefit from active learning, unless a pretrained model from a similar domain is available [13].

In some studies, modification of a loss function solved the sparse annotation challenge to some extent. [14] used class-balancing methods to improve the segmentation performance given sparse annotations without trying to fill in the missing mask pixels. In this proposed method, only the labeled pixels contribute to a weighted segmentation loss. The dataset used in this work contains some densely annotated WSIs and some sparsely annotated WSIs. However, segmenting whole slides images using coarse and sparse annotations is challenging and remains understudied in the literature.

Utilizing domain adaptation and leveraging external data have generated promising segmentation results. However, to the best of the authors' knowledge, no carefully labeled public dataset is available on skin biopsy images, and datasets

from other domains have significantly different morphological features compared to those of skin biopsy images.

There are limited studies on skin biopsy image segmentation. [15] presented a robust technique for epidermis segmentation in whole slide skin histopathological images, using thresholding and shape analysis. [16] produced a model for segmenting psoriasis-affected human skin biopsy images into the dermis, epidermis, and non-tissue regions. [17] developed a fully automated technique for lymph node segmentation that is robust to stains such as H&E, MART-1, S-100, and KI-67. However, semantic segmentation of clinically important structures in skin biopsy images is one of the most understudied areas in the literature. This is especially true for the datasets with imperfect and limited ground-truth annotations.

In this paper, we describe a carefully designed segmentation pipeline that can train a CNN on images with coarse and sparse annotation to accurately segment clinically important tissue structures in WSIs. This approach can be extremely helpful for medical image researchers because both data and annotations are expensive to acquire.

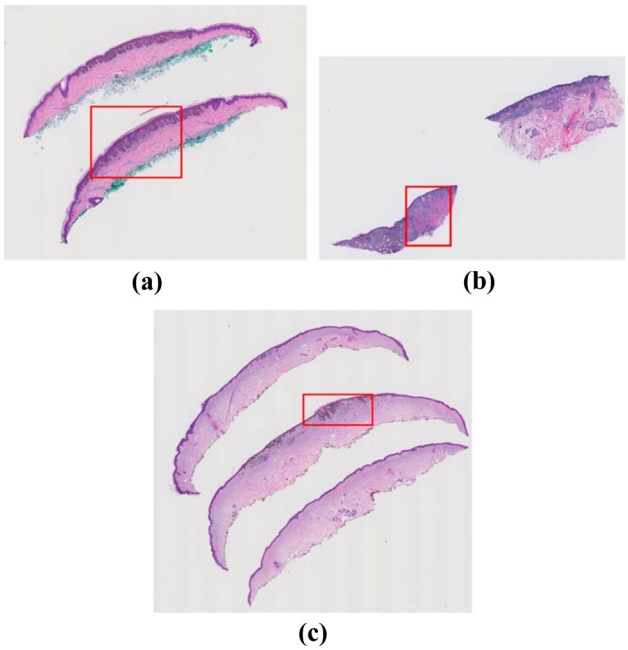


Fig. 2 Examples of variably sized whole slide images are shown. Region of interests (ROIs) that helped pathologists in diagnosis are shown in red boxes. **a**) a moderately dysplastic nevus case **b**) an invasive melanoma stage ≥ T1b case, **c**) an invasive melanoma stage T1a case

Dataset

Our dataset includes 240 hematoxylin and eosin (H&E) stained slides of digitized skin biopsy images, acquired by a Bellevue, Washington, dermatopathology laboratory for the MPATH study [4]. This dataset contains melanocytic skin lesions from shave, punch, and excisional specimens. The cases can be classified into five different Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis (MPATH-Dx) simplified categories based on presumed risk of the lesion and suggested treatment recommendations [18]. Example diagnostic terms for each MPATH-Dx class are as follows: (I) mildly dysplastic nevus, (II) moderately dysplastic nevus, (III) melanoma in situ, (IV) invasive melanoma stage T1a, and (V) invasive melanoma stage ≥ T1b.

A consensus panel of three dermatopathologists with internationally recognized expertise met over several days to reach consensus diagnoses for all cases, using the aforementioned MPATH-Dx classification tool [19]. Following these consensus meetings, the consensus panel members, as well as an additional dermatopathologist on the MPATH research team (S. Knezevich), utilized digitized images of all cases to identify one rectangular area as a region of interest (ROI) per case. These regions represent an important area of the WSI for the diagnosis. The size of these ROIs is not fixed and varies from one case to another (Fig. 2). We can extract the ROIs using their coordinates and perform various analyses on them to improve the overall diagnosis (Fig. 3).

To train a segmentation model, labels of different tissues as the ground-truth are required. However, since the annotation task is a very labor-intensive task, we obtained coarse and sparse annotations only on the ROI images by an expert pathologist (M. Mokhtari). Not only are the annotations not on the full WSI (Fig. 4a), but they are also sparse within the annotated ROI (Fig. 4b). Moreover, the annotations are coarse, i.e., they are not pixel-level accurate, as shown in Fig. 4c.

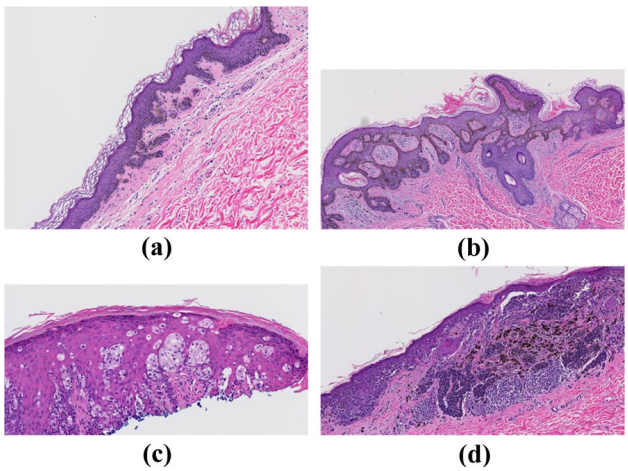


Fig. 3 Examples of four variably sized ROIs with different diagnoses in the MPATH dataset. **a** MPATH-Dx class I, mildly dysplastic nevus. **b** MPATH-Dx class II, moderately dysplastic nevus. **c** MPATH-Dx class III, melanoma in situ. **d** MPATH-Dx class IV, invasive melanoma stage T1a

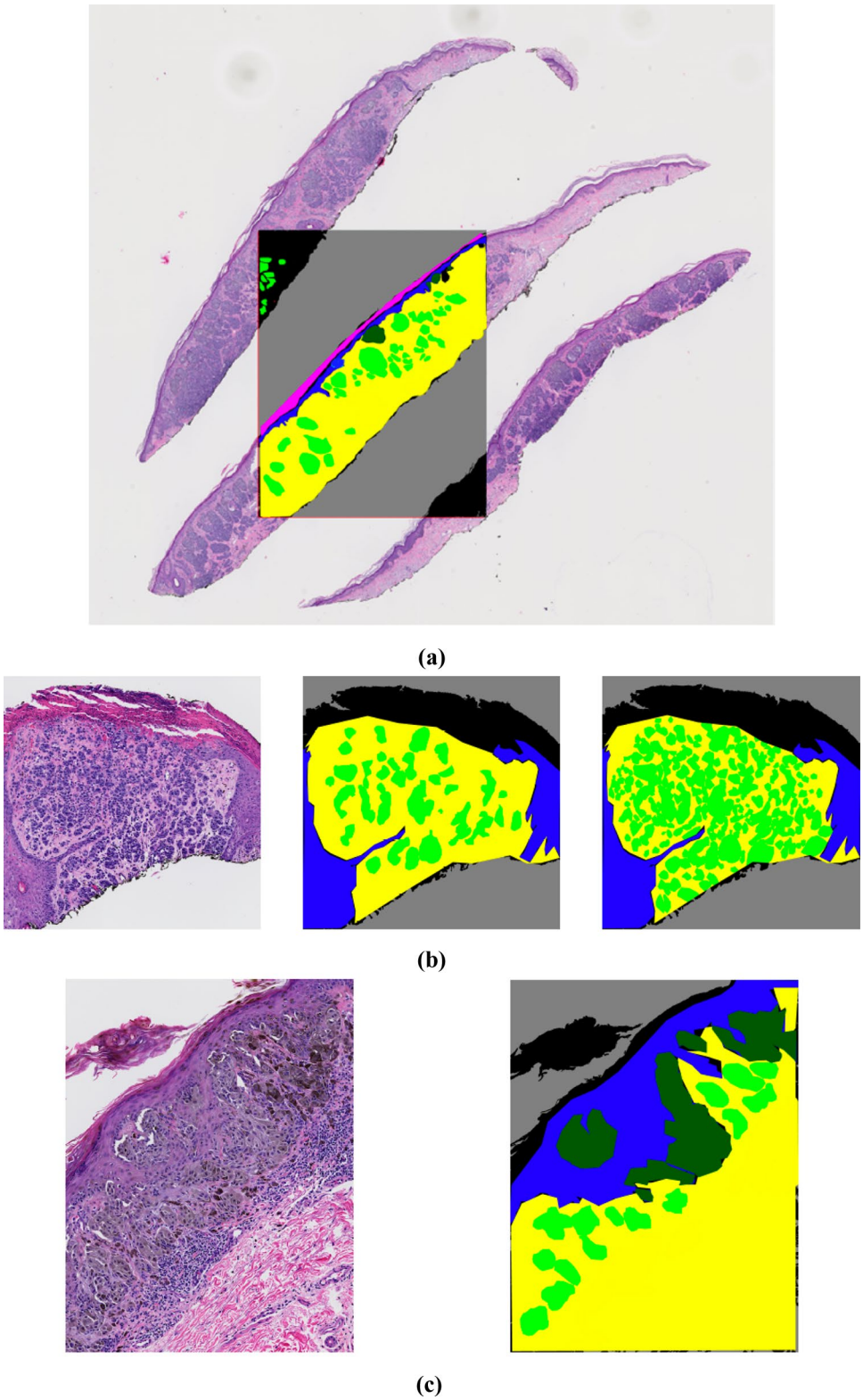


Fig. 4 Examples of sparse and coarse annotation in our ground-truth. **a** Sparse annotation of an ROI overlaid on the corresponding WSI. **b** Example of an ROI (left) with its corresponding sparse annotation of dermal nests (DMN) (middle) and full annotation of dermal nests (DMN) (right); this full annotation was acquired for the sake of comparison and is not available in the training set. **c** An example of coarse annotations (right) of different tissues in an ROI (left)

For pixel-level annotations, Sedeen,[1] a pathology image viewer, was used. Various structures were labeled with different names and corresponding colors as follows: epidermis (EP) in blue, dermis (DE) in yellow, stratum corneum (COR) in pink, epidermal nests (EPN, corresponding to epidermal melanocytic nests) in dark green, and dermal nests (DMN, corresponding to dermal melanocytic nests) in light green. Using the threshold-based segmentation method of [20], the background (BG) was detected and added to the labels in gray. We followed existing segmentation dataset annotation protocols (e.g., Cityscapes) and marked the pixels that do not correspond to any informative entity as unlabeled (UL) in black.

Method and Model

Medical imaging literature has witnessed great progress in the design and performance of deep convolutional models for medical image segmentation [13]. Thus, we utilized a CNN for our task of semantic segmentation.

Preprocessing

Since the labeling was done on ROIs, we started the process of training and evaluation on ROIs. As the preprocessing step, cropping, resizing, and augmentation were performed on these images.

Cropping and Resizing

Since the ROI sizes vary from ~ 480×360 to ~ $23{,}500 \times 22{,}400$, we chose the smallest size of 480×360 as the model input. However, resizing the biggest crop of $23{,}500 \times 22{,}400$ to such a small size (480×360) can significantly impact the information that can be acquired from such images. Instead, we follow a standard approach wherein bigger ROIs are divided into patches, and then patch-level segmentation masks are generated and combined to produce a ROI-level segmentation mask [21, 22]. In particular, for bigger ROIs, we extract patches of size 1440×1080 and then resize them to 480×360 before feeding them to the model. The segmentation output is then upsampled using nearest-neighbor interpolation to produce the segmentation mask that is of the same size as the patch before resizing.

Augmentation

We used various augmentation techniques such as horizontal flipping, affine transformation, perspective transformation, brightness/contrast/color manipulations, image blurring, sharpening, Gaussian noise, and random cropping to improve the robustness of our model. We used the fast augmentation library for these augmentation techniques [23].

Data Split

For a fair evaluation of the model, we divided the ROI dataset into two subsets, the training and testing sets, with a ratio of 80/20, respectively. The testing set was kept unseen from the model until the last step of the final evaluation. The training set is further split into train and validation sets, with a ratio of 80/20. We use the validation set for monitoring the training process and model selection and the testing set for the evaluation. Training, validation, and testing samples are all from different patients. While splitting the dataset, we were careful not to include any patient data from the training set in the testing or validation sets.

U-Net

The U-Net architecture of [22] is a well-known segmentation network and has shown good performance across different biomedical segmentation applications, such as MRI images [24], COVID-19 [25], skin lesion images [26], and lung, heart, and clavicle X-ray images [27].

We extended the U-Net encoder–decoder model for segmenting skin biopsy images. We used the implementation of U-Net by [28] in our work. We used ResNet-34 [29] pretrained on the ImageNet dataset [30] as the encoder and a standard U-Net decoder [28].[2]

Two-Stage Pipeline

Skin biopsy images have entities of variable size. Entities like the dermis and epidermis are large and easy to segment [15], while entities like dermal and epidermal nests are small and more difficult to segment. This issue becomes especially more troublesome when the smaller entities have sparser labeling compared to the larger entities. Hence, if the segmentation model is trained in a single-stage with all the labels at once, the model will perform better on larger entities and not as well on the smaller ones.

To overcome this problem, we developed a two-stage segmentation pipeline: first, a segmentation U-Net model is trained with labels of large entities in the histopathology image (background, stratum corneum, epidermis, dermis). Then, in the second stage, there are two sub-stages: (1) stage 2-Dermis is trained on the dermis portion of the images and uses the ground truth for the smaller entities that are present in the dermis (i.e., DMN). (2) Stage 2-Epidermis is trained

[1] https://pathcore.com/sedeen

[2] We did not use attention in the U-Net decoder block.

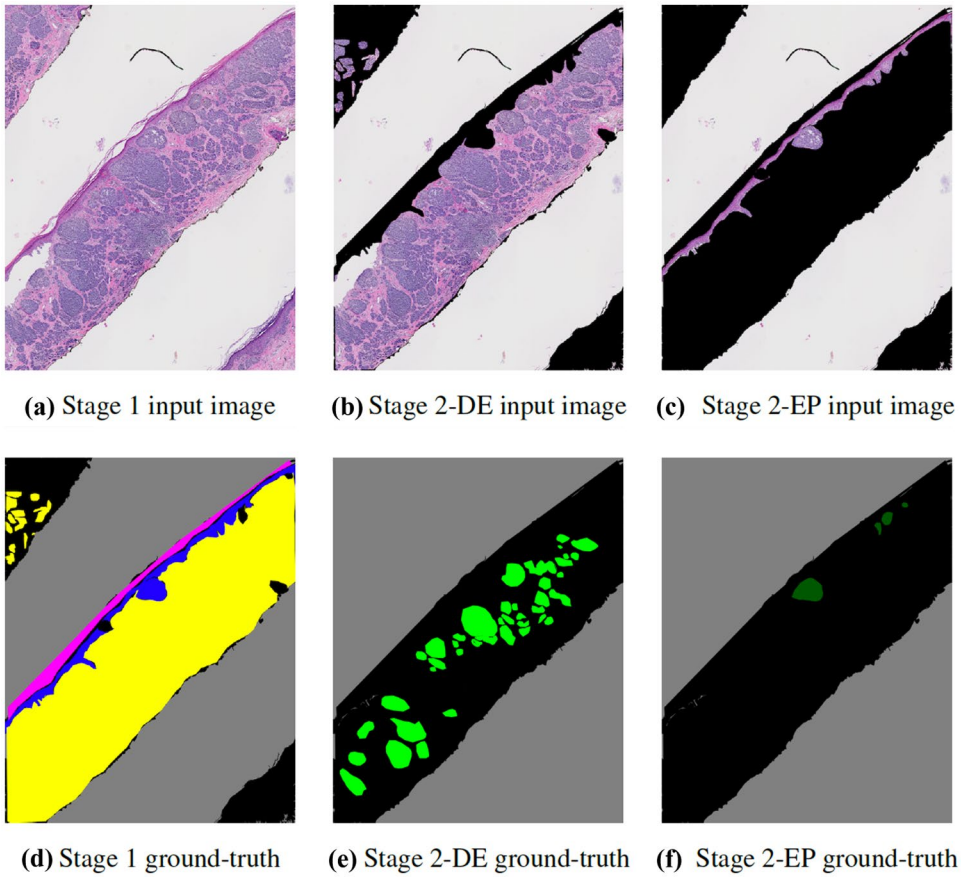


Fig. 5 Examples of input images and their corresponding ground-truth for the proposed two-stage pipeline. **a** and **d** show the input image and ground-truth to stage 1, containing the dermis (DE-yellow), epidermis (EP-blue), corneum (COR-pink), and background (BG-gray). **b** and **e** show the input image and ground-truth to stage 2-Dermis, containing dermal nests (DMN-light green) and background (BG-gray). **c** and **f** show the input image and ground-truth to stage 2-Epidermis, containing epidermal nests (EPN-dark green) and background (BG-gray)

on the epidermis portion of the images and uses the ground truth for the smaller entities that are present in Epidermis (i.e., EPN).[3] Figure 5 shows an example of one ROI and its corresponding mask, which is modified for different stages of our proposed pipeline.

As previously mentioned, the whole segmentation pipeline is trained in two stages: the first stage for big entities, such as the dermis and epidermis, and the second stage for smaller entities within the dermis and epidermis, such as dermal nests and epidermal nests. All the training stages used the stochastic gradient decent (SGD) optimizer with a momentum of 0.9 and a learning rate of 0.0001. The stage 1 encoder is trained for 1000 epochs. For the second stage weight initialization, U-Net in both the dermis and epidermis branches was initialized with the stage 1 model and fine-tuned for 100 epochs. This helps the model to converge faster. All the experiments were performed on an Intel(R) Xeon(R) Silver 4110 CPU 2.10 GHz with NVIDIA GeForce GTX 1080 GPU.

[3] While there are other small structures, such as hair follicles and blood vessels present in skin biopsy images, they are not clinically important for the diagnosis, so we do not try to segment them in this work.

Evaluation Metrics

To evaluate our models, we used the mean intersection over union (IoU). The IoU is a number from 0 to 1 that specifies the amount of overlap with the ground-truth (Eq. 1). An IoU of 0 means that there is no overlap between the prediction and ground-truth, and an IoU of 1 means the prediction and ground-truth completely overlap. Thus, a higher value of IoU means better performance:

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

For the final evaluation, we calculated another metric, Dice coefficient, which is 2 × the area of overlap divided by the total number of pixels in both images (Eq. 2):

$$Dice = \frac{2 \times TP}{(TP + FP) + (TP + FN)} \tag{2}$$

where true positive (TP) is the number of pixels that are correctly predicted as nest, true negative (TN) is the number of pixels that are correctly predicted as not-nest, false negative (FN) is the number of pixels that are incorrectly predicted as not-nest, and false positive (FP) is the number of pixels that are incorrectly predicted as nest.

Acquiring pathologists' annotations was a challenge. While we did not have full annotations for the whole dataset, we acquired fine-grained nest annotations on ROIs in the testing set for quantitative evaluation.

Results

In the training set, labels of dermis and epidermis are present in the ground-truth labels, which are used for the extraction of epidermis in stage 2-Epidermis and extraction of dermis from stage 2-Epidermis. However, for the testing set, the generated segmentation mask of stage 1 must be used to extract dermis and epidermis in their corresponding stage 2 branches. Since the important tissues that we aim to segment in stage 2 are DMN in dermis and EPN in epidermis, those entities are extracted from stage 2 and are overlaid on the stage 1 segmentation mask to generate the final segmentation mask. Figure 1 shows the application of the trained model on the testing set. As the final post-processing step, the separate crops of the ROIs are merged back to the original shape of the ROI.

Figure 6 shows some examples of the original ROI in the testing set, the corresponding coarse and sparse annotations provided, initially, the corresponding new full annotations (available only on DMN and EPN) and the segmentation mask generated by our model, which was trained on the coarse and sparse annotations. Quantitative results are shown in Table 1.

Generating WSI Segmentation Masks

The final goal of this work is to train a segmentation model on ROI images with sparse and coarse labels and produce segmentation masks for WSIs. To this end, we used the validation pipeline in Fig. 1 on WSI to generate a segmentation mask of the stratum corneum, dermis, epidermis, dermal nests, and epidermal nests. To feed the images to the segmentation model, first, a threshold-based method was applied on each WSI to extract individual slices as explained in the Extraction of Individual Slices section; then the same preprocessing as on the ROI images was applied on individual slices of the WSI. After the preprocessing, the crops were fed to the model, and after acquiring the segmentation masks, they were merged to create a WSI segmentation mask.

Extraction of Individual Slices

Prior to generating the segmentation masks, each whole slide biopsy image was split into individual slices using a slice extraction method. There are two benefits in performing

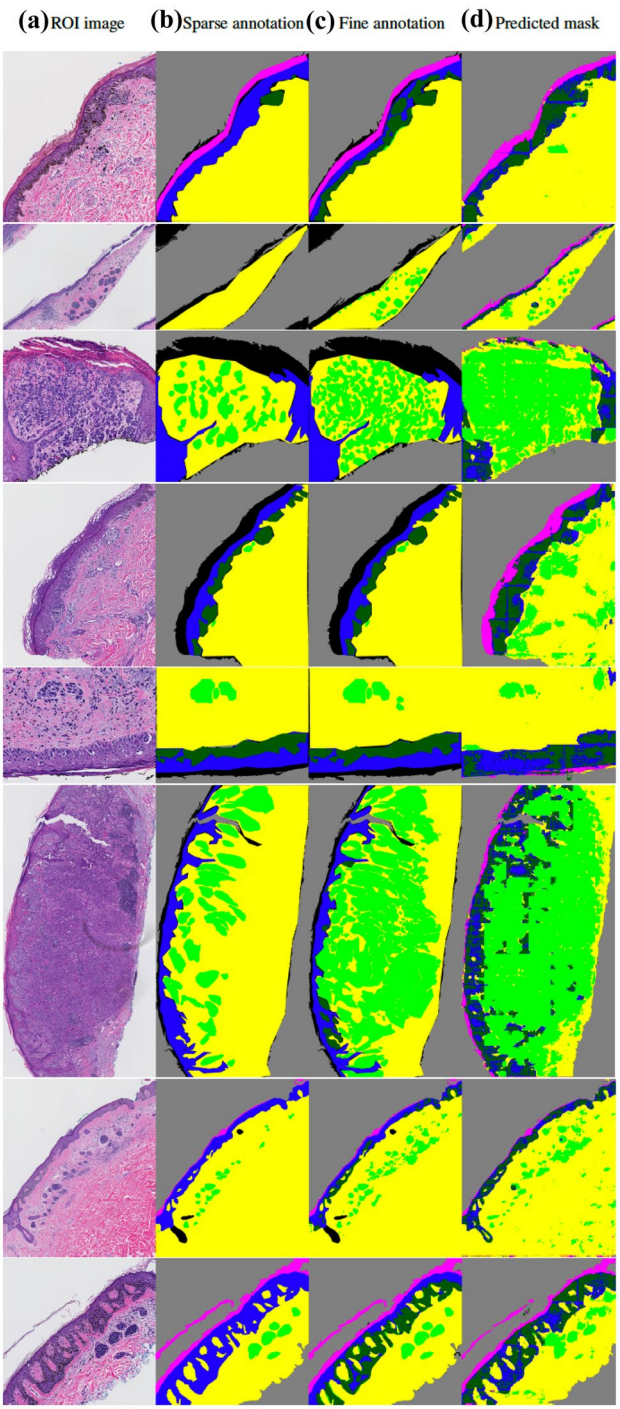


Fig. 6 Examples of original ROI, sparse and coarse annotation, fine pixel-level nest annotation, and segmentation mask by our pipeline. The annotation and segmentation images contain the dermis (DE-yellow), epidermis (EP-blue), stratum corneum (COR-pink), background (BG-gray), dermal nests (DMN-light green), and epidermal nests (EPN-dark green). The model has been trained on sparse and coarse annotations similar to column (**b**) and can generate results of column (**d**) which are comparable to the fine pixel-level annotation of column (**c**). *Full annotations on nests are only available for the testing set*

Table 1 Evaluation of the segmentation model on ROI testing set

Segmentation stage	Dice score	IoU
Stage 1 (all tissues)	0.942	0.906
Stage 2-Dermis (DMN)	0.558	0.638
Stage 2-Epidermis (EPN)	0.332	0.558

the slide segmentation: (1) we reduced the size of the input images, and (2) we can eliminate the effect of the slides' orientations since this information does not aid in the model's prediction of the diagnosis. Figure 7 shows an example of a WSI containing three individual slices, which are extracted before feeding to the segmentation model.

Subjective Assessment with Pathologists

Since full annotations for the entire WSIs are not available for our dataset, to evaluate the WSI segmentation results qualitatively, three of our expert dermatopathologists were asked to review the segmentation masks on the WSI validation set containing 111 WSIs and grade the model's performance on several areas and tissue structures using discrete scoring. These dermatopathologists (C. May, O. Chang, S. Knezevich) are different from the original dermatopathologist (M. Mokhtari) who provided sparse annotations on the dataset and full nest annotations on a set of test ROIs. Their task was to evaluate the segmentation of the whole slide images.

To create the surveys and distribute the work, the validation set was divided into three subsets of 37 images without any overlap for each dermatopathologist to review, preserving the distribution of diagnosis class over each subset. For each dermatopathologist, an individual survey in Google Forms was provided with their corresponding subset. Each WSI was evaluated regarding four segmentation tasks: epidermis (EP), dermis (DE), epidermal nest (EPN), and dermal nest (DMN), chosen as being most important for diagnosis. For each segmented structure label, the dermatopathologists were asked to answer two questions with an objective to see if model is oversegmenting or under-segmenting:

Q1: How much of the tissue/area that is present in the corresponding WSI has been correctly identified by the model? Rate low, medium, or high.
Q2: How much of the label identified by the model is the correct tissue/area? Rate low, medium, or high.

The results from these three surveys were analyzed, both individually and in combination. To translate the qualitative grading into a subjective assessment that can be used to plot visual bar charts, we provided a numerical conversion as follows: if the grade of a label is low, the numerical equivalent is 1, medium is 2, and high is 3. The numerical equivalents of these ratings for each label in all the images were used to generate opinion score (OS) which is the arithmetic mean of each label rated by the dermatopathologists (Eq. 3), where R_n are the individual ratings for a given tissue structure, and N is the number of cases in the corresponding survey.

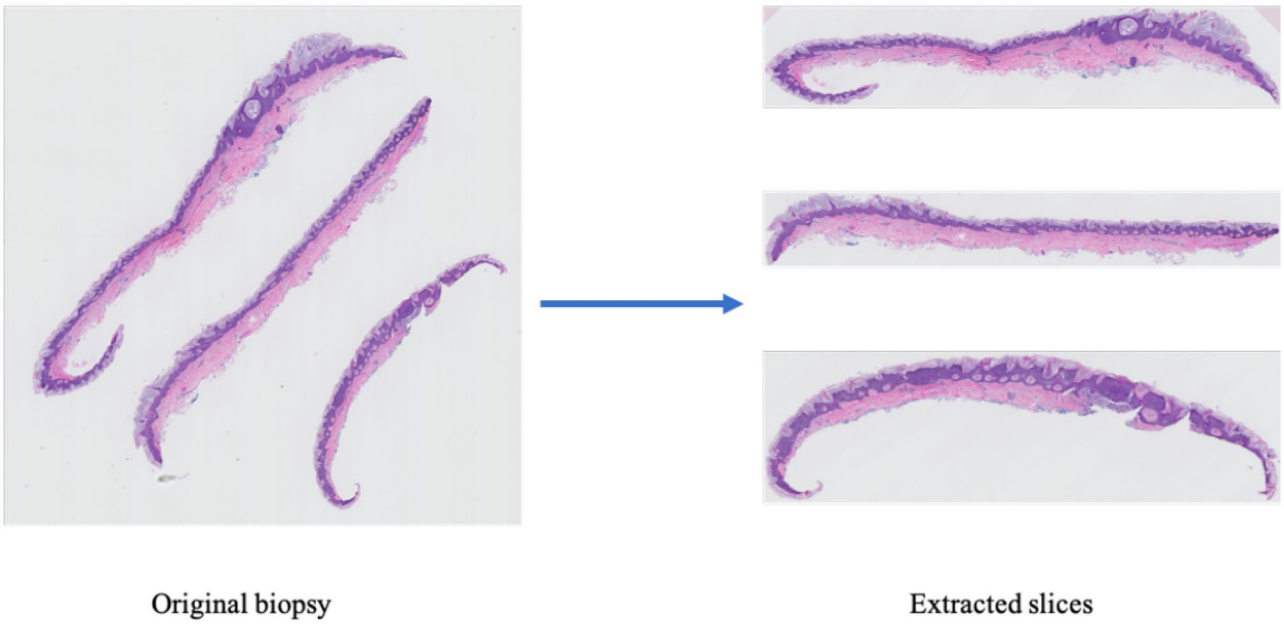
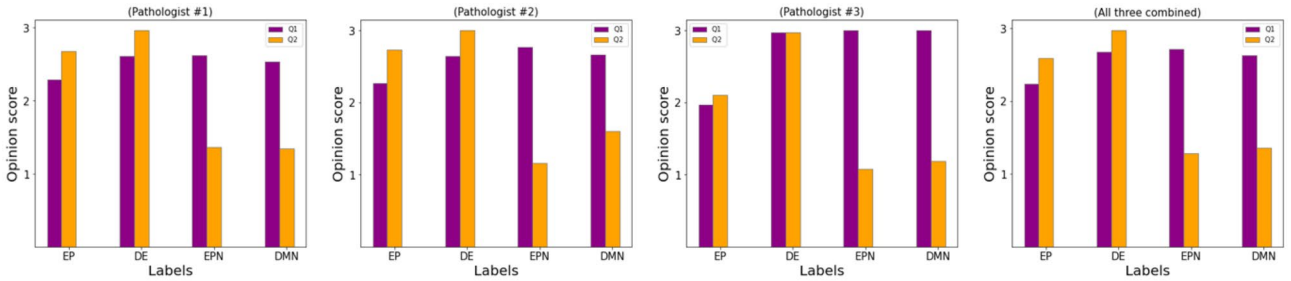
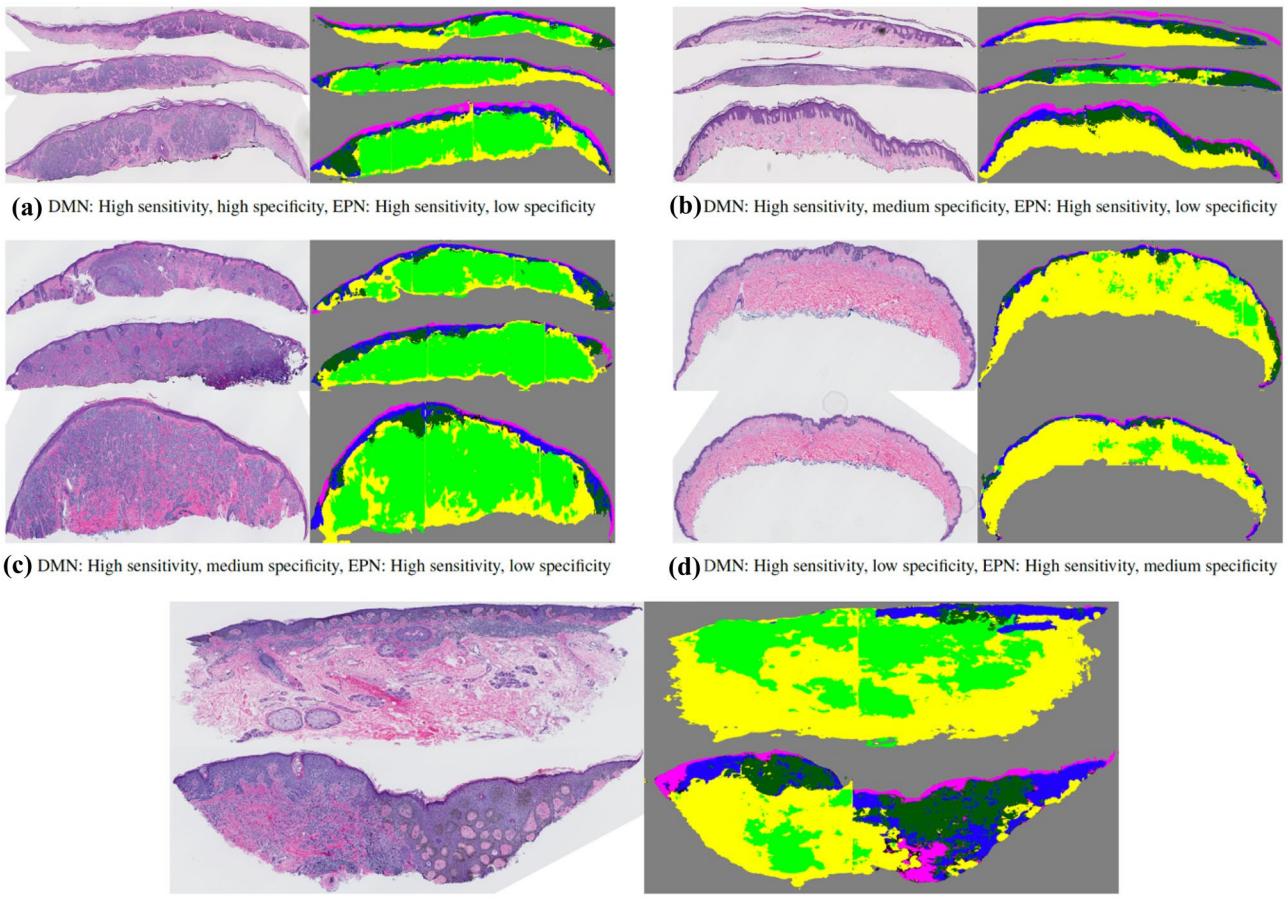


Fig. 7 An example of a WSI (left) and its corresponding slice extraction (right)



(a) Individual survey #1 **(b)** Individual survey #2 **(c)** Individual survey #3 **(d)** Combination of all three surveys

Fig. 8 Opinion score (OS) as subjective assessment for each label, epidermis (EP), dermis (DE), dermal nest (DMN), and epidermal nest (EPN), in terms of Q1 and Q2 for that tissue structure. The qualitative ratings by dermatopathologists are converted to their numerical equivalent as explained in Sect. 5.2. Each pathologist reviewed 37 different cases; **a**, **b**, and **c** are the individual surveys, and **d** is the combination of all three surveys



(a) DMN: High sensitivity, high specificity, EPN: High sensitivity, low specificity

(b) DMN: High sensitivity, medium specificity, EPN: High sensitivity, low specificity

(c) DMN: High sensitivity, medium specificity, EPN: High sensitivity, low specificity

(d) DMN: High sensitivity, low specificity, EPN: High sensitivity, medium specificity

(e) DMN: Medium sensitivity, low specificity, EPN: medium sensitivity, low specificity

Fig. 9 Examples of original WSI (left) and its corresponding segmentation mask (right). Slices of each WSI are extracted and concatenated vertically. The segmentation images contain the dermis (DE-yellow), epidermis (EP-blue), stratum corneum (COR-pink), background (BG-gray), dermal nests (DMN-light green), and epidermal nests (EPN-dark green). The model has been trained on coarse and sparse annotations. The captions show the dermatopathologists' qualitative grading on each WSI segmentation mask for dermal nests (DMN) and epidermal nests (EPN) **a**) high sensitivity and high specificity on DMN, high sensitivity and low specificity on EPN **b**) high sensitivity and medium specificity on DMN, high sensitivity and low specificity on EPN **c**) high sensitivity and medium specificity on DMN, high sensitivity and low specificity on EPN **d**) high sensitivity and low specificity on DMN, high sensitivity and medium specificity on EPN **e**) medium sensitivity and low specificity on DMN, medium sensitivity and low specificity on EPN

Figure 8 shows the OS for each label in terms of Q1 and Q2 for individual pathologists and their combination:

$$OS = \frac{\sum_{n=1}^{N} R_n}{N} \tag{3}$$

A close examination of the WSI segmentation masks (Fig. 9) shows that the sparse and coarse annotations provide the possibility of segmenting the tissue structures with high-quality performance on whole slide images. However, the presence of different types of noise due to coarse labeling in the training set, such as inaccurate borders and unintentional human error in labeling, plus the lack of labels on entities that are similar to dermal nests, such as inflammatory cells and eccrine ducts, results in over-labeling of nests overall. The over-labeling of the epidermal nests is higher than that of the dermal nests, which follows the pattern of our training ground-truth, in which epidermal nest annotations are noisier than dermal nest annotations. While having high sensitivity (i.e., finding all the nests) is critical in medical dataset analysis, having high specificity (i.e., reducing the false positives) is also required for accurate diagnosis. Hence, reducing noise from even sparse annotations is an important step before training a segmentation model. This can be done by having the ground truth checked by a separate pathologist from the one who created it.

Discussion

As the number of melanoma cases continues to increase, the accurate diagnosis of melanocytic lesions in skin biopsies is becoming more critical for patient care and treatment. For the pathologist, a crucial step in interpreting a melanocytic proliferation involves assessing the microanatomic location of the melanocytic population, including whether the process involves the epidermis, dermis, or both. The semantic segmentation of these tissues (e.g., epidermis and dermis) and melanocyte position (e.g., epidermal nests and dermal nests) in skin biopsies is a required initial step in creating an automated diagnostic tool that has the potential to assist pathologists in their evaluation of melanocytic lesions, including melanoma and its precursors. Automated diagnosis tools have the potential to assist pathologists in their diagnoses.

While segmentation is a significant element in the diagnosis pipeline, training a segmentation model generally requires a large, high-quality annotated ground-truth. However, most medical datasets require expert-level annotation as ground-truth, and such a requirement is a challenging, time-consuming, and expensive task, leading to a scarcity of sufficiently sized and carefully annotated datasets for training; overcoming this challenge is a necessity in medical image research to produce computer-aided diagnosis systems. Hence, a segmentation pipeline that can use coarse and sparse annotation to produce a segmentation model is likely to be quite beneficial.

In this work, we proposed a two-stage pipeline for the segmentation of important tissue structures in skin biopsy images using coarse and sparse annotations on small regions of WSIs. In this pipeline, larger entities were trained in the first stage, and smaller entities were trained in two subbranches. The testing segmentation results, both on the ROIs and the WSIs, show the potential of this pipeline. Dermal nests (DMN) and epidermal nests (EPN), alongside the dermis and epidermis, are important tissues/areas in the histopathology of skin biopsy images that play a crucial role in the diagnosis. Our system was able to generate segmentation masks for both epidermis/dermis and nests with high-quality performance, indicating that having sparse annotation on important tissues has the potential for producing a useful segmentation model. On the other hand, our results suggest that both the DMN and EPN can be over-labeled by the model, highlighting the problems that coarse annotation can cause for the system, especially on a small dataset in which the ground-truth did not clearly distinguish between nests and other similar structures. These two findings suggest that having sparse, but fine, annotation on a small region of the WSI may be enough for training a better segmentation model. It is important to note that having a disease such as melanoma that might severely disrupt the dermal–epidermal junction causes complications that increase the necessity of having fine annotation on epidermal nests.

The primary purpose of generating a semantic segmentation model using sparse and coarse annotation is to provide valuable information for a future automated diagnosis pipeline. Such information has been shown to be helpful in our breast pathology analysis work [7]. Furthermore, a semantic segmentation model can be beneficial for human pathologists. While segmenting the dermis and epidermis is not a challenging task for a trained pathologist, providing information about melanocyte position (e.g., epidermal nests and dermal nests) can assist pathologists in their decision-making, especially in challenging cases.

The limitations of our work are as follows: since stage 2 relies on stage 1 as a preprocessing step, an incorrect segmentation of the epidermis and dermis in stage 1 renders inaccuracy in the second stage. In future work, we will update the training labels to correct all noisy annotations while still training on a sparse subset of possible annotations.

Author Contribution Shima Nofallah, conceptualization, methodology, software, validation, investigation, and writing — original draft. Mojgan Mokhtari, resources. Wenjun Wu, software. Sachin Mehta, methodology and writing — review and editing. Stevan Knezevich, resources. Caitlin J. May, resources. Oliver H. Chang, resources. Annie C. Lee,

writing — review and editing. Joann G. Elmore, funding acquisition, supervision, and resources. Linda Shapiro, supervision and writing — review and editing.

Funding The research reported in this study was supported by grants R01CA200690 and U01CA231782 from the National Cancer Institute of the National Institutes of Health, 622600 from the Melanoma Research Alliance, and W81XWH-20–1-0798 from the US Department of Defense. The funders had no role in the design and conduct of the study, collection, management, analysis, and interpretation of the data, preparation, review, or approval of the manuscript nor decision to submit the manuscript for publication.

Data Availability The dataset that has been used in this research is a private dataset.

Code Availability Our code is open-source and available at https://github.com/shimaxy/Sparse_Annotation_Segmentation.

Declarations

Ethics Approval Not applicable.

Consent to Participate. Not applicable.

Consent for Publication. Not applicable.

Conflict of Interest The authors declare no competing interests.

References

1. Rigel, D.S. and J.A. Carucci, *Malignant melanoma: prevention, early detection, and treatment in the 21st century.* CA: a cancer journal for clinicians, 2000. **50**(4): p. 215–236.
2. Kosary, C.L., et al., *Clinical and prognostic factors for melanoma of the skin using SEER registries: collaborative stage data collection system, version 1 and version 2.* Cancer, 2014. **120**: p. 3807-3814.
3. Guy Jr, G.P., et al., *Vital signs: melanoma incidence and mortality trends and projections—United States, 1982–2030.* MMWR. Morbidity and mortality weekly report, 2015. **64**(21): p. 591.
4. Elmore, J.G., et al., *Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study.* Bmj, 2017. **357**: p. j2813.
5. Sirinukunwattana, K., et al., *Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images.* IEEE transactions on medical imaging, 2016. **35**(5): p. 1196-1206.
6. Esteva, A., et al., *Dermatologist-level classification of skin cancer with deep neural networks.* nature, 2017. **542**(7639): p. 115–118.
7. Mercan, E., et al., *Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions.* JAMA network open, 2019. **2**(8): p. e198777-e198777.
8. Zhang, L., et al., *When unseen domain generalization is unnecessary? rethinking data augmentation.* arXiv preprint arXiv:1906.03347, 2019.
9. Li, Z., K. Kamnitsas, and B. Glocker. *Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019. Springer.
10. Sourati, J., et al., *Active deep learning with fisher information for patch-wise semantic segmentation*, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. 2018, Springer. p. 83-91.
11. Mahapatra, D., et al. *Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2018. Springer.
12. Zheng, H., et al. *Biomedical image segmentation via representative annotation*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019.
13. Tajbakhsh, N., et al., *Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation.* Medical Image Analysis, 2020. **63**: p. 101693.
14. Bokhorst, J.-M., et al. *Learning from sparsely annotated data for semantic segmentation in histopathology images*. in *International Conference on Medical Imaging with Deep Learning--Full Paper Track*. 2018.
15. Xu, H. and M. Mandal, *Epidermis segmentation in skin histopathological images based on thickness measurement and k-means algorithm.* EURASIP Journal on Image and Video Processing, 2015. **2015**(1): p. 1-14.
16. Pal, A., et al., *Psoriasis skin biopsy image segmentation using Deep Convolutional Neural Network.* Computer methods and programs in biomedicine, 2018. **159**: p. 59-69.
17. Alheejawi, S., et al., *Novel lymph node segmentation and proliferation index measurement for skin melanoma biopsy images.* Computerized Medical Imaging and Graphics, 2019. **73**: p. 19-29.
18. Piepkorn, M.W., et al., *The MPATH-Dx reporting schema for melanocytic proliferations and melanoma.* Journal of the American Academy of Dermatology, 2014. **70**(1): p. 131-141.
19. Carney, P.A., et al., *Achieving consensus for the histopathologic diagnosis of melanocytic lesions: use of the modified Delphi method.* Journal of cutaneous pathology, 2016. **43**(10): p. 830-837.
20. Otsu, N., *A threshold selection method from gray-level histograms.* IEEE transactions on systems, man, and cybernetics, 1979. **9**(1): p. 62-66.
21. Mehta, S., et al. *Learning to segment breast biopsy whole slide images*. in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018. IEEE.
22. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.
23. Buslaev, A., et al., *Albumentations: fast and flexible image augmentations.* Information, 2020. **11**(2): p. 125.
24. Çiçek, Ö., et al. *3D U-Net: learning dense volumetric segmentation from sparse annotation*. in *International conference on medical image computing and computer-assisted intervention*. 2016. Springer.
25. Saood, A. and I. Hatem, *COVID-19 lung CT image segmentation using deep learning methods: U-Net versus SegNet.* BMC Medical Imaging, 2021. **21**(1): p. 1-10.
26. Mirikharaji, Z. and G. Hamarneh. *Star shape prior in fully convolutional networks for skin lesion segmentation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2018. Springer.
27. Frid-Adar, M., et al., *Improving the segmentation of anatomical structures in chest radiographs using u-net with an imagenet pre-trained encoder*, in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. 2018, Springer. p. 159-168.
28. Yakubovskiy, P. *Segmentation Models Pytorch*. 2020; Available from: https://github.com/qubvel/segmentation_models.pytorch.
29. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

30. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *2009 IEEE conference on computer vision and pattern recognition*. 2009. Ieee.