Learning Melanocytic Proliferation Segmentation in Histopathology Images from Imperfect Annotations

Kechun Liu[†] [*] Mojgan Mokhtari[‡] Beibin Li[†] Shima Nofallah[†] Caitlin May[§]
Oliver Chang[††] Stevan Knezevich[|] Joann Elmore[‡‡] Linda Shapiro[†]

[†] University of Washington [‡] Isfahan University of Medical Sciences [§] Dermatopathology Northwest
[††] VA Puget Sound [|] Pathology Associates [‡‡] University of California, Los Angeles

Abstract

Melanoma is the third most common type of skin cancer and is responsible for the most skin cancer deaths. A diagnosis of melanoma is made by the visual interpretation of tissue sections by a pathologist, a challenging task given the complexity and breadth of melanocytic lesions and the subjective nature of biopsy interpretation. We leverage advances in computer vision to aid melanoma diagnosis by segmenting potential regions of lesions on digital images of whole slide skin biopsies. In this study, we demonstrate a Mask-R-CNN-based segmentation framework for such a purpose. To alleviate the cost of data annotation, we leverage a sparse annotation pipeline. Our model can be trained on sparse and noisy labels and achieves state-of-the-art performance in identifying melanocytic proliferations, producing a segmentation with Dice score 0.719, mIOU 0.740 and overall pixel accuracy 0.927.

1. Introduction

Melanoma is the third most common type of skin cancer and is responsible for most skin cancer deaths [18, 19]. In the United States, between 2007–2011, more than 63,000 people were diagnosed with melanoma, and nearly 9,000 people died from this disease each year [10, 14]. Although melanoma rates overall are highest among older adults, it is the third most common cancer in adolescents and young adults (aged 15–39 years) [38]. Recent analyses have found increases in incidence across all tumor thicknesses and stages [18]. According to a report for melanoma skin cancer in [14], the 5-year relative survival rates are 99%, 66%, and 27% in localized, regional and distant stages, respectively. This shows that although melanoma in advanced stages is difficult to treat, precursors of melanoma, (i.e. melanoma in situ) and thin melanomas (i.e., depth of invasion < 1 mm), are 99% likely to be cured. Thus, the early diagnosis of melanoma is important for reducing melanoma deaths.

The gold standard for melanoma diagnosis is microscopic examination of skin biopsies using routine hematoxylin and eosin (H&E)-stained tissue sections with supplemental immunohistochemistry as needed. Pathologists' diagnoses of melanocytic lesions have been noted to have both low accuracy and reproducibility [7]. The diagnosis of melanoma and melanoma precursors is predicated on an accurate assessment of architectural growth patterns. Assessing melanocytic lesions requires identifying where melanocytes are microanatomically situated in the skin (*e.g.* intraepidermal, dermal-epidermal junction, intradermal) and characterizing, in part, the architecture of the melanocytic population. Melanoma in situ, for example, exhibits confluent melanocytic growth of single cells and nests at the epidermal base and/or extension of melanocytes into the mid-to-upper levels of the epidermis (pagetoid spread). Invasive (malignant) melanoma contains atypical melanocytes within the dermis, often lacking features of maturation as they descend (e.g., smaller and more dispersed cells). While melanocytic proliferations exhibit numerous patterns of growth, our paper focuses on two fundamental patterns: single cell dispersion and nests. Figure 1 shows examples of singly dispersed intraepidermal melanocytes and nested melanocytes at the dermal-epidermal junction.

The diagnosis of melanoma and its precursors may have significant barriers on histopathology due to variable architectural growth patterns and cytomorphology. A critical first step in developing accurate machine algorithms is the recognition of melanocytic proliferations and how they are situated in cutaneous microanatomy. In view of this, we focus on the following question: can we design a computer-vision-aided system to automatically point out these growth patterns? Once the system can reliably detect single cell

[*] kechun@cs.washington.edu

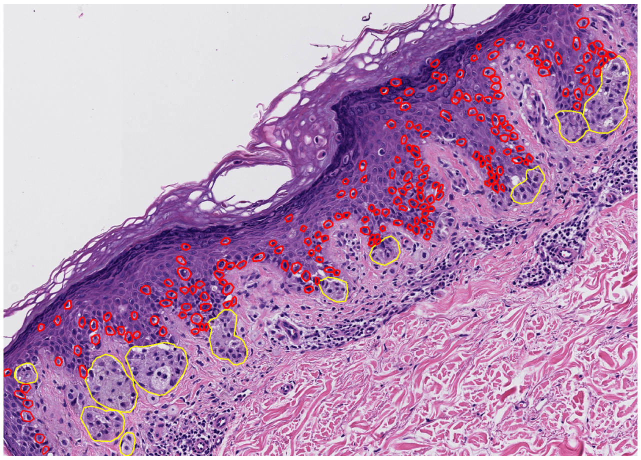


Figure 1: **Examples of melanocytic proliferations:** we use red polygons to mark the singly dispersed intraepidermal single melanocytes and yellow polygons to mark the melanocytic nests.

dispersion and nests of melanocytes, it can provide potential histological clues to aid pathologists. In the rest of the paper, we refer to singly dispersed intraepidermal and nested melanocytes as melanocytic proliferations for brevity.

Motivated by the recent advances in deep learning, an increasing number of researchers leverage neural networks for medical image segmentation, especially for histological image segmentation, in different disease domains. For instance, many studies use Convolutional Neural Networks (CNN) to find tumor regions or ducts for breast cancer [11, 22]. Similarly, researchers have utilized CNNs to segment prostate cancer grading to aid diagnosis [23, 16]. For skin cancer, Kucharski *et al.* [21] performed patch-level melanocytic nest segmentation using Autoencoders.

In our work, we developed a pipeline to identify image-level melanocytic proliferations with weak supervision. Our method leverages sparse and noisy annotations on skin biopsy images and uses weighted loss functions to account for the imperfect labels. Altogether, we achieve state-of-the-art performance on segmentation of melanocytic proliferations. Our work is validated by ground truth from experienced pathologists trained in dermatopathology.

In summary, our main contributions are three-fold:

(1) Our model provides image-level segmentation results that can assist in diagnosis by pathologists and aid in downstream computer vision analysis.

(2) Our approach achieves state-of-the-art accuracy on identification of melanocytic proliferations.

(3) Our framework only requires weakly-supervised training using sparse and noisy annotations, which greatly alleviates the annotation work by pathologists and, at the same time, achieves a solid performance.

2. Related Work

In this section, we first briefly review the previous work on melanoma diagnosis using computer vision techniques. Then, we discuss several recent representative medical image segmentation methods, including semantic segmentation models such as FCN and U-Net, and instance segmentation models like Mask R-CNN, upon which our approach is based.

2.1. Melanoma diagnosis work

A few previous papers have been published regarding melanoma diagnosis and melanocytic region segmentation using skin histological images. In terms of melanoma diagnosis, some notable examples include a feature-based diagnosis framework based on cytological and textural characteristics of the epidermis and dermis [41]; a method capable of diagnosing squamous cell carcinoma in situ by using an epidermis axis analysis [30]; a method for classifying nodular basal cell carcinomas (BCCs), dermal nevi, and seborrheic keratoses using a Fully Convolutional Network [32]; and methods for melanoma diagnosis based on tumor region segmentation [37, 33]. None of the above diagnosis methods considers melanocytic proliferations, which are essential clinical diagnostic criteria. A recent work developed a melanocytic nest segmentation method [21], which is currently the state-of-the-art model in this specific task. It first uses a convolutional autoencoder to train a reconstruction network using 128×128 patches from 70 WSIs. It then replaces the decoder part with a segmentation head to train a segmentation model with 39 annotated WSIs using previous layers in the encoder, which has learned the features of skin biopsy patches. While this method shows it is feasible to apply computer vision techniques for melanocytic nest segmentation, it is limited by the number of parameters in the convolutional autoencoder. In practice, it is hard to feed large patches to the network due to memory issues, which means a loss of contextual information that is essential for melanocytic proliferations segmentation. This leads to lower segmentation and detection accuracies. The autoencoder approach achieved promising results to identify melanocytic nests, and these aforementioned studies all inspired us to improve the segmentation performance for finding melanocytic proliferations.

2.2. Deep learning in medical image segmentation

Recent developments in medical image segmentation using deep learning, especially semantic segmentation and instance segmentation, provide our study's groundwork.

Semantic segmentation is a common task that classifies each pixel into a semantic category. For instance, the Fully

Convolutional Network (FCN) [26], which enables CNNs to take input images as arbitrary size and output its corresponding mask, has been widely used in biomedical image analysis [1, 17, 35]. Built upon the elegant structure of the FCN and a design of skip connections, U-Net overcomes the trade-off between localization and the use of context [34]. Many published works regarding medical image analysis benefit from U-Net [9, 39, 43]. Variations of U-Net, like 3D U-Net [5], attention U-Net [31] and V-Net [29], also help tackle many medical image analysis tasks [2, 4, 8].

While semantic segmentation becomes much more prevalent in medical image analysis, instance segmentation is rarely used. This is mainly because most medical image segmentation tasks don't need the target tissue to be separated as instances, and instance segmentation is more difficult than semantic segmentation. However, instance segmentation also shows promising results in many medical image analysis studies, such as segmenting glands in colon histology images [42, 3], ducts in breast biopsies [22], nuclei in any microscopy images [20] and different stages of prostate cancer [23]. Among many instance segmentation models, Mask R-CNN is a well-known and influential approach that leverages a two-stage structure to first roughly locate the target instances by a Region Proposal Network and second locate precisely, classify and segment the targets. In our study, we adopt Mask R-CNN to detect and segment melanocytic proliferations.

3. Methods

Figure 2 shows our proposed melanocytic proliferation segmentation pipeline, which consists of two main components: (1) data annotation and preprocessing procedure, (2) melanocytic proliferation segmentation model, and patch stitching. In this section, we first introduce the dataset and annotations used in our proposed melanocytic proliferation segmentation pipeline. Second, we describe the model used in both the segmentation and post-processing steps in detail. Finally, we provide evaluation metrics on which our model was assessed and compared with previous efforts.

3.1. Dataset

Our dataset consists of 227 region of interest (ROI) images extracted from hematoxylin and eosin (H&E) stained slides of skin biopsy images at 10x magnification and diagnosed by three expert pathologists, who all agreed on the consensus diagnosis and selected the ROIs [7]. Each ROI image indicates the diagnosis result to some extent, which ranges from class 1 (benign) with 29 samples, class 2 (moderately dysplastic nevi) with 49 samples, class 3 (melanoma in situ) with 67 samples, class 4 (invasive melanoma stage T1a) with 50 samples, and class 5 (invasive melanoma stage $\geq$ T1b) with 32 samples. All the images are compressed in tiff format.

3.1.1 Melanocytic proliferation annotations

Demarcating melanocytic proliferations in a biopsy image is difficult for three main reasons. First, melanocytic nests come in various sizes and shapes, which makes annotation difficult. Second, a typical whole slide image can include a few to several hundred melanocytic nests and hundreds of single melanocytes, depending on the cases. This causes annotating a single image to be extremely labor-intensive. Last but not least, annotating nests can only be done by an expert dermatopathologist, which makes it extremely costly in time and human resource to collect a sizable dataset for training a model. Altogether, these challenges render it difficult to create datasets of marked melanocytic proliferations suitable for deep learning techniques.

Training computer vision models require datasets of adequate sizes. Given the aforementioned difficulties in annotation, we designed the following annotation procedure. We ask an expert pathologist to partially mark the 227 ROI images, using the Sedeen Viewer [28]. Annotating the single melanocytes is challenging due to their small sizes and large quantities, as shown in red labels in Figure 1. To solve this, we drew polygons around many melanocytes instead of circling every single melanocyte. In addition, we had two other expert pathologists check the markings. Although this procedure leads to several sources of noise in our labels, which we will discuss in 3.1.2, it gives us significant savings in annotation time, in comparison to fully and exactly labeling every singly-dispersed and nested melanocyte.

3.1.2 Annotation Caveats

Sparse annotation, noisy annotation, and irreducible human errors are the main caveats of the efficient weakly annotation procedure, as shown in Figure 3. First, because of the sparse labeling strategy, not all the images are marked with melanocytic proliferations, and not all the melanocytic proliferations are marked in any one single image. Figure 3a shows an example of our sparse annotations. Second, when drawing polygons around many intraepidermal single melanocytes instead of circling every single melanocytes, this inevitably includes the ”false” background into the labels and leads to the noisy annotations as illustrated in Figure 3b. Third, since we are using a polygonal annotation tool, there are some misalignments between the annotated boundaries and the true boundaries. Fourth, there exist few false positive examples that are labeled by mistake. We refer to these as “human error” and show an example of them in Figure 3c. These limitations render our annotations “silver standard” rather than “gold standard”, even if a consensus was reached among pathologists in the markings. To address these annotation caveats, we incorporate different training strategies and loss functions to the computer vision model, which are shown in Section 3.2.

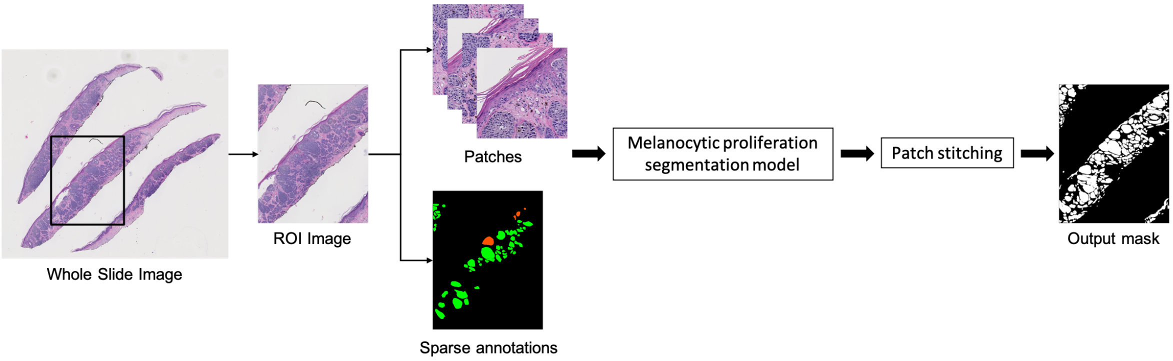
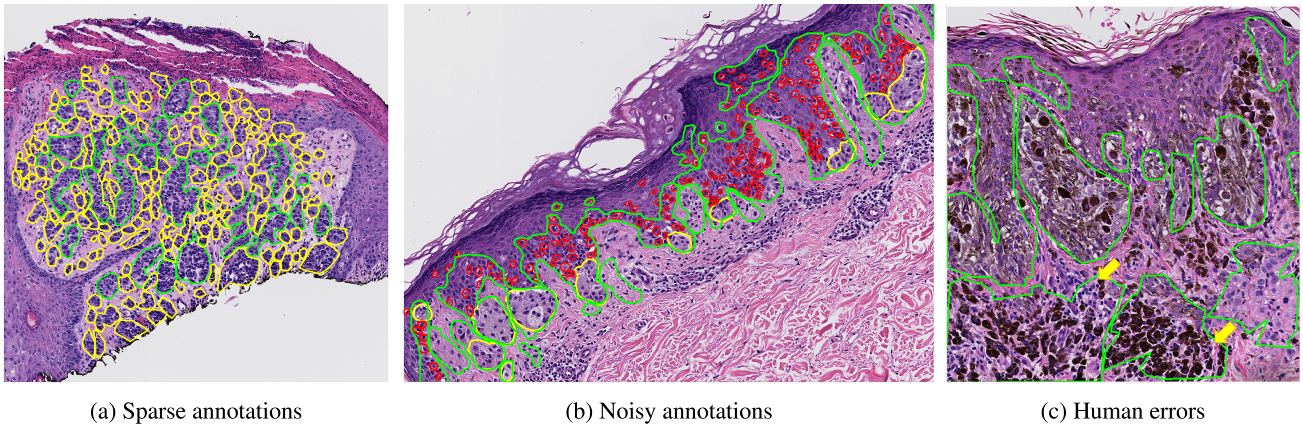


Figure 2: **Melanocytic proliferation segmentation pipeline:** this pipeline enables training from sparse annotations using the Mask R-CNN model with different loss functions, and aggregates results on patches to provide an image-level mask that can be used in further diagnosis. Note: only the middle slide is the important region of interest, even though parts of the other two slides fell into the box.


(a) Sparse annotations

(b) Noisy annotations

(c) Human errors

Figure 3: **Limitations in our annotations:** (a) sparse annotations: green markings belong to the sparse annotations, and yellow markings show the complementary annotations; (b) noisy annotations: red and yellow markings show the true intraepidermal single melanocytes and the melanocytic nests separately, while they are actually labeled as the green markings in our annotations; (c) human errors: the two green markings pointed to with yellow arrows are false positive examples which were inaccurately labeled.

3.1.3 Data split

We split our dataset randomly into training, validation, and testing subsets. The number of ROI images in each subset equals 174 (76%), 19 (8%), and 34 (15%), respectively. The validation set is used to choose the model and parameters. After the training is finished, the testing set is used to evaluate the model's performance. The dataset is split before further preprocessing steps.

3.1.4 Data preprocessing

Even if ROI images are cropped from the original whole slide images (WSIs), they are still too large to fit into memory, with the smallest size 428×381 to largest size 23691×22401 and median size 6221×3171 at the magnification 10x. A common strategy to deal with this memory issue is to extract patches [6, 13]. As we are using a pretrained Mask R-CNN model whose default anchor box sizes are 32, 64, 128, 256, and 512, we split the images into 1000×1000 patches; this avoids training new layers since the dataset is too small. The patches will be further resized with the shortest edge around 800, which is a default step in the model, so that the default anchor box sizes can cover most of the melanocytic proliferations. Besides, to reduce the boundary effect when stitching patches into images, we downscale the ROI images from 10x to 5x, *i.e.* down-sampling to half resolution, and extract the patches with 50% overlap.

3.2. Model

Among our processed data patches, most include only a few small-sized melanocytic proliferations, leaving the majority of the patches non-target tissues. This motivates us to adopt the Mask R-CNN [12], a widely-used instance segmentation model for our task. The advantage of Mask R-CNN is that it is a two-stage model, as described in Figure 4. The first stage is to roughly locate the target entities using a Region Proposal Network. The second stage is to further refine the anchor boxes and at the same time produce segmentation masks and classification results. This design helps us efficiently filter out the majority of the non-target tissues. We next introduce the loss function designed for our partially labeled dataset.

3.2.1 Loss Function

The original Mask R-CNN model was developed for instance segmentation on the Microsoft COCO: Common Objects in Context (MS COCO) dataset [25], a fully labeled dataset. In comparison to MS COCO, our dataset is only sparsely labeled, as described in section 3.1.1. We hereby describe our modification to the loss function in Mask R-CNN to better suit our dataset.

The loss function for Mask R-CNN consists of 5 parts. (1) $L_{\text{rpn\_cls}}$: Classification loss in the RPN. (2) $L_{\text{rpn\_loc}}$: Anchor box location loss in the RPN. (3) L_{cls}: Classification loss in the prediction head. (4) $L_{\text{box\_reg}}$: Bounding box regession loss in the prediction head. (5) L_{mask}: Segmentation loss in the prediction head.

In the training of the Mask R-CNN, $L_{\text{rpn\_loc}}$, $L_{\text{box\_reg}}$, L_{mask} only back-propagate the loss values on positive samples . However, $L_{\text{rpn\_cls}}$ and L_{cls} fully utilize the labeled and unlabeled areas to decide whether there is an instance in the anchor box. As we are using a partially-labeled dataset, treating those unlabeled areas as background, *i.e.* not nests, is unfair to our task. Thus, we changed the loss functions in these two parts to better train our data. The original forms of $L_{\text{rpn\_cls}}$ and L_{cls} are binary cross entropy, and categorical cross entropy. In our study, we tried two other loss functions, weighted cross entropy (WCE) and focal loss (FL).

Weighted Cross Entropy is a variation of cross entropy with weights given to different categories to address the dataset imbalance. This helps achieve higher recall and precision. The larger the weight of a specific category, the higher the recall is on that category. In our study, WCE is used to reduce punishment from unlabeled areas. We define WCE as:

$$L_{\text{WCE}} = -\sum_i (w * y_i * \log(\hat{p_i}) + (1 - y_i) * \log(1 - \hat{p_i})) \quad (1)$$

where $y_i \in \{0, 1\}$ is the ground truth label whether the object belongs to class i. $\hat{p_i} \in [0, 1]$ is the probability of the object being in class i, predicted by the model. w is the weight given to the categories.

Focal Loss was first introduced in [24], which adds adaptive weights on cross entropy to let the model focus on hard examples rather than treating hard and easy examples in the same way. This strategy helps to alleviate the imbalanced data problem. In our study, focal loss is used to reduce unfair punishment as well as let the model learn from hard examples and is given by

$$\begin{aligned} L_{\text{WFL}} = -\sum_i (w * y_i * (1 - \hat{p_i})^{\lambda} * \log(\hat{p_i}) \\ + (1 - y_i) * \hat{p_i}^{\lambda} \log(1 - \hat{p_i})) \end{aligned} \quad (2)$$

where λ is a hyper-parameter. The larger λ is, the more the model focuses on hard examples. We use $\lambda = 2$ in our experiment, following the same setting in [24]. The definitions of y_i, $\hat{p_i}$, and w remains the same as equation 1.

In both L_{WCE} and L_{WFL}, w is used to balance the labeled and unlabeled areas. The results of different values of w are shown in the ablation study 4.2.

3.2.2 Transfer Learning

Transfer Learning is an effective technique in computer vision tasks where the dataset is the main bottleneck. As with most medical image analysis domains, we are limited by a scarcity of accurately annotated training data due to the difficulty and cost of collecting and annotating data. The scarcity is even aggravated in our partially labeled dataset. In fact, we only have 130 images labeled with melanocytic proliferations in the training dataset. Thus, we use transfer learning via CNNs originally pretrained on natural images to compensate for this limitation.

Previous studies [15, 36] show that transfer learning in CNNs helps alleviate the need for large datasets. Despite the difference between natural images and medical images, neural nets can still learn some basic structures *e.g.* edges, blobs from natural images, and these parameters are shared in transfer learning. CNNs can be further fine-tuned in the limited medical datasets to learn a specific task like melanocytic nests. Thus, transfer learning can help to train a model on a limited dataset as well as allow us to take advantage of deep neural networks.

In our study, we used an off-the-shelf implementation of Mask R-CNN from detectron2 [40], pretrained on the MS COCO dataset, which has over 200,000 accurately labeled images and 80 categories. To better utilize transfer learning, we kept the pretrained model's parameters as much as possible, except we changed the prediction head since our task is for different categories and preprocessed the images to get their sizes close to MS COCO image sizes.

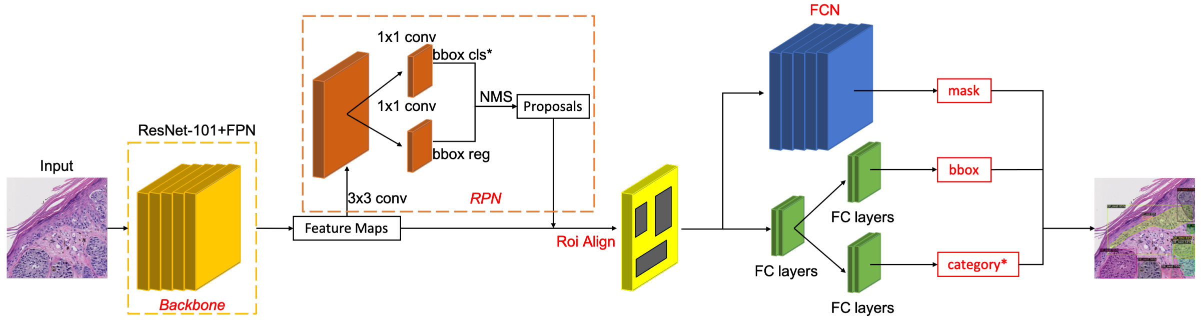


Figure 4: **Overview of Mask R-CNN model architecture:** We use ResNet-101+FPN as the backbone to extract feature maps from the input image. Combining the feature maps and the anchor box results from Region Proposal Network (RPN), fixed-size feature maps are fed into three prediction heads (classification, bounding box regression, and segmentation) to jointly generate instance segmentation results. Since our dataset is partially-labeled, we change the loss functions of bbox_cls in the RPN and classification head to reduce punishment from unlabeled data.

3.2.3 Post-processing

To provide a complete prediction on ROI images instead of patches, we stitched the patch results to image-level masks by only preserving instances with confidence scores over 0.5 and aggregating them together to generate masks. Although this step loses the information of separate instances, it is acceptable in our task as the delimitations on the melanocytic proliferations are also vague.

3.2.4 Implementation and Training

The Mask R-CNN model was fine-tuned on our dataset using the SGD optimizer for a total of 40 epochs with an initial learning rate of 0.001. We used learning rate warm-up in the first three epochs and decayed the learning rate by 0.5 after every 4 epochs. To achieve a stable measurement, we ran each model 10 times with different randomization (e.g., random mini-batch, random dropout, etc.). In the following sections, we report the mean and the standard deviation (STD) for all metrics in our experiments.

3.3. Evaluation Metrics

To make our model comparable with the state-of-the-art melanocytic nest segmentation method [21], we used the standard pixel-level metrics: Dice Score, mean Intersection Over Union (mIOU), accuracy, sensitivity and specificity to evaluate the model's segmentation performance. These metrics are calculated based on the pixel populations of true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

$$\text{mIOU} = \frac{1}{2} \times (\frac{TP}{TP + FP + FN} + \frac{TN}{TN + FP + FN})$$

4. Results and Discussion

Despite being trained with weak-supervision using only partially labeled datasets, our model was able to achieve good performance on the fully labeled test set. This is due to the architecture, loss function (WCE and FL), and transfer learning techniques, as discussed in Section 3.2. To have a fair evaluation, we asked our expert pathologist to thoroughly label the melanocytic nests in our test set, which consists of 34 ROI images. In this section, we provide experimental results on the fully labeled test set, ablation studies, as well as a detailed discussion of our results.

4.1. Experimental Results

We re-implemented the convolutional autoencoder model described in the previous state-of-the-art (SOTA) work [21], and trained it following all the detailed steps as described. Table 1 quantitatively compares this autoencoder with our method in different loss functions, including the default cross entropy loss, using the segmentation metrics described in Section 3.3. Although the SOTA autoencoder achieves higher sensitivity, it provides less accurate and more noisy segmentation results as shown in mIOU, accuracy, specificity and Figure 5. Overall, our method outperforms the SOTA autoencoder in Dice score, mIOU, accuracy and specificity.

Figure 5 qualitatively illustrates the goundtruth and the results of the autoencoder and our model overlaid on the H&E images. The first two rows in Figure 5 show two good examples of melanocytic proliferation segmentation results compared to both groundtruth and the autoencoder. The bottom two rows show two imperfect examples compared to the groundtruth. In Figure 5 (c), our model predicts a false positive proliferation in the middle layer of the epidermis, which consists of many keratinocytes with "halo" regions surrounding the nuclei. This is mainly because most

Method	**Dice**	**mIOU**	**Accuracy**	**Sensitivity**	**Specificity**
Autoencoder [21]	0.679	0.705	0.905	**0.814**	0.918
Mask R-CNN with CE loss	0.685	0.715	0.917	0.726	0.944
Mask R-CNN with WCE loss	0.705	0.726	0.917	0.792	0.935
Mask R-CNN with FL loss	**0.719**	**0.740**	**0.927**	0.751	**0.952**

Table 1: **Quantitative results:** Dice score, mIOU, Accuracy, Sensitivity and Specificity for all methods. The best performances are highlighted in bold font in this table.

Loss function	**Weight**	**Dice**	**mIOU**	**Accuracy**	**Sensitivity**	**Specificity**
Weighted Cross Entropy (WCE)	$w = 1$	0.685(0.013)	0.715(0.008)	0.917(0.002)	0.726(0.041)	0.944(0.006)
	$w = 2$	**0.705(0.003)**	**0.726(0.003)**	**0.917(0.003)**	**0.792(0.027)**	**0.935(0.007)**
	$w = 3$	0.701(0.009)	0.723(0.006)	0.915(0.003)	0.792(0.021)	0.933(0.005)
	$w = 5$	0.701(0.008)	0.722(0.006)	0.914(0.002)	0.813(0.028)	0.928(0.005)
	$w = 8$	0.700(0.007)	0.718(0.007)	0.909(0.005)	0.850(0.022)	0.918(0.008)
	$w = 12$	0.700(0.005)	0.716(0.003)	0.908(0.002)	0.847(0.021)	0.917(0.005)
Focal Loss (FL)	$w = 1$	0.717(0.018)	0.740(0.011)	0.928(0.002)	0.740(0.053)	0.954(0.007)
	$w = 2$	0.703(0.022)	0.731(0.014)	0.926(0.003)	0.710(0.053)	0.956(0.006)
	$w = 3$	0.702(0.021)	0.730(0.014)	0.926(0.003)	0.705(0.045)	0.957(0.004)
	$w = 5$	0.711(0.014)	0.735(0.008)	0.926(0.002)	0.730(0.044)	0.954(0.006)
	$w = 8$	**0.719(0.011)**	**0.740(0.007)**	**0.927(0.003)**	**0.751(0.027)**	**0.952(0.005)**
	$w = 12$	0.710(0.023)	0.734(0.015)	0.925(0.004)	0.742(0.056)	0.951(0.007)

Table 2: **Ablation experiments for weighted cross entropy (WCE) and focal loss (FL):** All the models with different weights were evaluated on our fully-labeled test set. The mean and standard deviation (in parenthesis) from 10 runs are reported.

intraepidermal melanocytes share the same characteristic of "halo" regions [27]. In Figure 5 (d), our model mispredicts some melanophages; this is mainly caused by the human errors in annotations described in 3.1.2.

4.2. Ablations

To understand the relationship between the weights in the loss function and segmentation performance, we tried several experiments with different weight values for WCE and FL. All the models with different loss function weights were evaluated in our fully-labeled test dataset. As shown in Table 2, we observe that the WCE loss performs the best when $w = 2$, and FL achieves the best performance when $w = 8$. The comparison between default cross entropy (w=1) and other weighted loss functions shows that adding weights helps improve performance in sparse annotation datasets (like ours). The larger standard deviations in focal loss compared to weighted cross entropy show that while focal loss enlarges the confidence scores in positive and negative samples, the noise in our dataset is also amplified, which leads to the uncertainty in the results.

4.3. Discussion

As shown in Table 1 and Figure 5, our proposed method achieved better results than the SOTA autoencoder [21] in all metrics except sensitivity, which can be explained by the autoencoder's tendency for overprediction. The identification of melanocytic proliferations could provide potential histological clues to help pathologists focus on important regions and reduce their workload. Moreover, inexperienced students could use our study to better understand the critical first step of the diagnosis process.

We chose Mask-RCNN as our model architecture because it is robust to noise. In comparison, the autoencoder mispredicts small and irrelevant entities as melanocytic proliferations, as shown in Figure 5. Mask-RCNN does not have the same drawback, because the anchor boxes help it focus on specific regions of interest and filter out the irrelevant background. The design of non-maximum suppression also reduces the noise around a target instance.

Diagnosis in practice requires many different features, such as confluent growth of melanocytes along the basal epidermis, mitotic figures, and melanocyte maturation on descent. Our model shows great potential serving as the first step of an automated diagnosis pipeline. Once we ac-

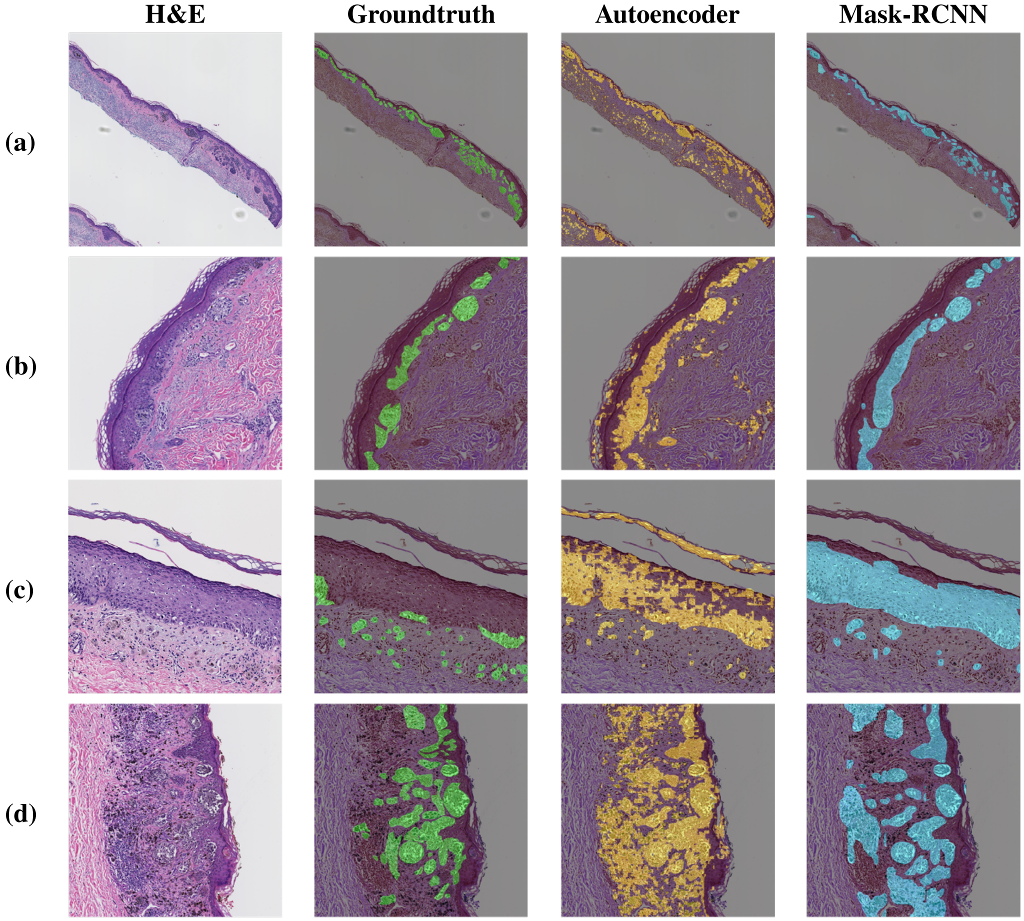


Figure 5: **Qualitative comparison between our model and SOTA autoencoder [21]:** From left to right, each column shows examples of ROI images stained by H&E, groundtruth annotated by expert pathologist, Autoencoder predictions, and Mask R-CNN predictions, separately.

curately recognize melanocytic proliferations and how they are situated in cutaneous microanatomy, we can incorporate other works to extract the aforementioned features. In the future, researchers can combine these features with classification techniques such as multi-instance learning and Transformers, to create an integrated diagnosis tool.

The quality of human annotation largely affects our model's performance. We find that irreducible human errors can lead to misclassification of melanophages as melanocytic proliferations in Figure 5 (d). One promising direction is to reduce human errors by leveraging our model's output. Noisy predictions from deep neural networks [22] can be used to assist data annotation in an interactive manner. We leave this direction to future work.

5. Conclusions

One important step in assessing melanocytic lesions is to identify melanocytic growth patterns such as single cell dispersion and nested melanocytes. In this study, we propose a weakly-supervised Mask-R-CNN-based model for melanocytic proliferations segmentation. By leveraging weak supervision, our model only requires partially labeled datasets, which vastly reduces the data annotation cost. We evaluated our method on ground truth labels provided by expert pathologists and found that it outperforms the previous state-of-the-art approach. Although, more comprehensive studies are needed to validate our approach in practice, we are excited about its potential to aid in melanoma diagnosis.

Acknowledgement: This work was supported by grants 622600 from the Melanoma Research Alliance, U01CA231782 from NIH/NCI and R01CA200690 from NIH/NCI. The opinions expressed in this paper are those of the authors and not of the funding agencies.

References

[1] Yaniv Bar, Idit Diamant, Lior Wolf, and Hayit Greenspan. Deep learning with non-medical training used for chest pathology identification. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 94140V. International Society for Optics and Photonics, 2015. 3

[2] Paul Blanc-Durand, Axel Van Der Gucht, Niklaus Schaefer, Emmanuel Itti, and John O Prior. Automatic lesion detection and segmentation of 18f-fet pet in gliomas: a full 3d u-net convolutional neural network study. *PLoS One*, 13(4):e0195798, 2018. 3

[3] Hao Chen, Xiaojuan Qi, Lequan Yu, Qi Dou, Jing Qin, and Pheng-Ann Heng. Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Medical image analysis*, 36:135–146, 2017. 3

[4] Xiaocong Chen, Lina Yao, and Yu Zhang. Residual attention u-net for automated multi-class segmentation of covid-19 chest ct images. *arXiv preprint arXiv:2004.05645*, 2020. 3

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 3

[6] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine*, 6:264, 2019. 4

[7] Joann G Elmore, Raymond L Barnhill, David E Elder, Gary M Longton, Margaret S Pepe, Lisa M Reisch, Patricia A Carney, Linda J Titus, Heidi D Nelson, Tracy Onega, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *Bmj*, 357, 2017. 1, 3

[8] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8):1822–1834, 2018. 3

[9] Yu Gordienko, Peng Gang, Jiang Hui, Wei Zeng, Yu Kochura, Oleg Alienin, Oleksandr Rokovyi, and Sergii Stirenko. Deep learning with lung segmentation and bone shadow exclusion techniques for chest x-ray analysis of lung cancer. In *International Conference on Computer Science, Engineering and Education Applications*, pages 638–647. Springer, 2018. 3

[10] US Cancer Statistics Working Group et al. United states cancer statistics: 1999–2010 incidence and mortality web-based report. *Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute*, 201, 2013. 1

[11] Zichao Guo, Hong Liu, Haomiao Ni, Xiangdong Wang, Mingming Su, Wei Guo, Kuansong Wang, Taijiao Jiang, and Yueliang Qian. A fast and refined cancer regions segmentation framework in whole-slide breast pathological images. *Scientific reports*, 9(1):1–10, 2019. 2

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 5

[13] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019. 4

[14] NNAM Howlader, AM Noone, M ea Krapcho, D Miller, A Brest, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, et al. Seer cancer statistics review, 1975-2016. *National Cancer Institute*, 2019. 1

[15] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, 2016. 5

[16] Nathan Ing, Zhaoxuan Ma, Jiayun Li, Hootan Salemi, Corey Arnold, Beatrice S Knudsen, and Arkadiusz Gertych. Semantic segmentation for prostate cancer grading by convolutional neural networks. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 105811B. International Society for Optics and Photonics, 2018. 2

[17] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016. 3

[18] Ahmedin Jemal, Mona Saraiya, Pragna Patel, Sai S Cherala, Jill Barnholtz-Sloan, Julian Kim, Charles L Wiggins, and Phyllis A Wingo. Recent trends in cutaneous melanoma incidence and death rates in the united states, 1992-2006. *Journal of the American Academy of Dermatology*, 65(5):S17–e1, 2011. 1

[19] Ahmedin Jemal, Edgar P Simard, Christina Dorell, Anne-Michelle Noone, Lauri E Markowitz, Betsy Kohler, Christie Eheman, Mona Saraiya, Priti Bandi, Debbie Saslow, et al. Annual report to the nation on the status of cancer, 1975–2009, featuring the burden and trends in human papillomavirus (hpv)–associated cancers and hpv vaccination coverage levels. *JNCI: Journal of the National Cancer Institute*, 105(3):175–201, 2013. 1

[20] Jeremiah W Johnson. Adapting mask-rcnn for automatic nucleus segmentation. *arXiv preprint arXiv:1805.00500*, 2018. 3

[21] Dariusz Kucharski, Pawel Kleczek, Joanna Jaworek-Korjakowska, Grzegorz Dyduch, and Marek Gorgon. Semi-supervised nests of melanocytes segmentation method using convolutional autoencoders. *Sensors*, 20(6):1546, 2020. 2, 6, 7, 8

[22] Beibin Li, Ezgi Mercan, Sachin Mehta, Stevan Knezevich, Corey W Arnold, Donald L Weaver, Joann G Elmore, and Linda G Shapiro. Classifying breast histopathology images with a ductal instance-oriented pipeline. *arXiv preprint arXiv:2012.06136*, 2020. 2, 3, 8

[23] Wenyuan Li, Jiayun Li, Karthik V Sarma, King Chung Ho, Shiwen Shen, Beatrice S Knudsen, Arkadiusz Gertych, and Corey W Arnold. Path r-cnn for prostate cancer diagnosis and gleason grading of histological images. *IEEE transactions on medical imaging*, 38(4):945–954, 2018. 2, 3

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[27] Cheng Lu, Muhammad Mahmood, Naresh Jha, and Mrinal Mandal. Detection of melanocytes in skin histopathological images using radial line scanning. *Pattern Recognition*, 46(2):509–518, 2013. 7

[28] Anne L Martel, Dan Hosseinzadeh, Caglar Senaras, Yu Zhou, Azadeh Yazdanpanah, Rushin Shojaii, Emily S Patterson, Anant Madabhushi, and Metin N Gurcan. An image analysis resource for cancer research: Piip—pathology image informatics platform for visualization, analysis, and management. *Cancer research*, 77(21):e83–e86, 2017. 3

[29] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 3

[30] Navid Noroozi and Ali Zakerolhosseini. Differential diagnosis of squamous cell carcinoma in situ using skin histopathological images. *Computers in biology and medicine*, 70:23–39, 2016. 2

[31] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 3

[32] Thomas George Olsen, B Hunter Jackson, Theresa Ann Feeser, Michael N Kent, John C Moad, Smita Krishnamurthy, Denise D Lunsford, and Rajath E Soans. Diagnostic performance of deep learning algorithms applied to three common diagnoses in dermatopathology. *Journal of pathology informatics*, 9, 2018. 2

[33] 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*, pages 0–0, 2019. 2

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[35] Holger R Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. An application of cascaded 3d fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99, 2018. 3

[36] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016. 5

[37] Mike Van Zon, Nikolas Stathonikos, Willeke AM Blokx, Selim Komina, Sybren LN Maas, Josien PW Pluim, Paul J Van Diest, and Mitko Veta. Segmentation and classification of melanoma and nevus in whole slide images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 263–266. IEEE, 2020. 2

[38] Hannah K Weir, Loraine D Marrett, Vilma Cokkinides, Jill Barnholtz-Sloan, Pragna Patel, Eric Tai, Ahmedin Jemal, Jun Li, Julian Kim, and Donatus U Ekwueme. Melanoma in adolescents and young adults (ages 15-39 years): United states, 1999-2006. *Journal of the American Academy of Dermatology*, 65(5):S38–e1, 2011. 1

[39] Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257, 2019. 3

[40] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 5

[41] Hongming Xu, Cheng Lu, Richard Berendt, Naresh Jha, and Mrinal Mandal. Automated analysis and classification of melanocytic tumor on skin whole slide images. *Computerized medical imaging and graphics*, 66:124–134, 2018. 2

[42] Yan Xu, Yang Li, Mingyuan Liu, Yipei Wang, Maode Lai, I Eric, and Chao Chang. Gland instance segmentation by deep multichannel side supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 496–504. Springer, 2016. 3

[43] Zizheng Yan, Xiaoguang Han, Changmiao Wang, Yuda Qiu, Zixiang Xiong, and Shuguang Cui. Learning mutually local-global u-nets for high-resolution retinal lesion segmentation in fundus images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 597–600. IEEE, 2019. 3