Interpretable Analysis of Melanoma in Whole Slide Imaging: Detection, Virtual Staining, and Diagnostic Insights

Kechun Liu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:
Linda Shapiro, Chair
Sheng Wang
Steve Tanimoto

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Interpretable Analysis of Melanoma in Whole Slide Imaging: Detection, Virtual Staining, and Diagnostic Insights

Kechun Liu

Chair of the Supervisory Committee:
Linda Shapiro
Computer Science & Engineering

Whole slide imaging (WSI) has transformed digital pathology, offering extensive details in skin biopsies used for melanoma diagnosis. However, clinical assessments remain challenging, with diagnostic accuracy and efficiency limited by the inherent complexity and variability of these images. While computer-aided diagnosis (CAD) systems can analyze WSIs using deep learning approaches, they often treat the images as pure data inputs, lacking the clinical understanding essential for nuanced assessment. Developing an accurate and reliable CAD model therefore requires not only detecting diagnostically relevant structures but also capturing the clinical context in which these structures are assessed. This dissertation aims to address these needs by introducing a novel diagnosis model that integrates both the key diagnostic structures and the interpretive processes pathologists use to evaluate WSIs.

The initial focus of the work is on detecting and segmenting diagnostically relevant structures within WSIs, beginning with a method for identifying melanocytic proliferations using sparse and noisy annotations to highlight suspicious regions that guide diagnostic reasoning. To further investigate cellular entities, **VSGD-Net** was developed to accurately detect melanocytes in H&E-stained slides, a crucial step for analyzing melanocyte distribution and growth patterns in melanoma. Additionally, **VSGD-Net** enables virtual synthesis of IHC-stained images from standard H&E WSIs, facilitating further insights without the need

for additional staining procedures. This method is extended by **CC-WSI-Net**, which enables seamless synthesis across entire slides rather than isolated patches, enhancing contextual coherence at the whole-slide level.

To support pathologists' diagnostic workflow, the Semantics-Aware Attention Guidance (**SAG**) framework is introduced, integrating semantic information to guide model's attention toward regions with high diagnostic relevance. Finally, a Multi-level Region-of-Interest Attending Network (**MiRA**) is developed to emulate how pathologists diagnose WSIs by integrating information from both low-resolution whole slides and high-resolution regions of interest. This dual-level approach improves diagnostic efficiency and aligns the model's behavior with clinical workflows, making it both effective and interpretable for pathologists.

In summary, this dissertation presents deep learning methods for interpretable melanoma diagnosis, integrating key diagnostic structures with clinical reasoning. These advancements aim to improve the reliability and consistency of melanoma diagnoses, supporting more efficient clinical workflows and better patient outcomes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support, guidance, and encouragement of many individuals, to whom I am deeply grateful.

I would like to express my deepest gratitude to my advisor, Dr. Linda Shapiro, a woman with immense courage, ambition, and confidence. She has not only been a mentor but also a true role model, inspiring me both personally and professionally. Her encouragement and insightful advice have played a crucial role in shaping my research and professional growth.

I am also incredibly grateful to Dr. Joann Elmore, co-PI of my advisor, for her guidance and collaboration. Additionally, I would like to extend my sincere appreciation to the talented pathologist team - Dr. Stevan Knezevich, Dr. Caitlin May, Dr. Oliver Chang, and Dr. Mojgan Mokhtari - for their expertise and contributions, which have been instrumental in advancing my research.

A special thanks to my co-authors, Dr. Wenjun Wu, Dr. Beibin Li, Sitong Liu, and Dr. Shima Nofallah, for their collaboration, dedication, and insightful discussions. It has been a privilege to work alongside such brilliant researchers.

To my labmates - Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Wisdom Ikezogwo, Zixuan Liu, Mahtab Bigverdi, Rustin Soraki, Kalyani Marathe, Yuguang Lee, and Nishat Anjum Khan - thank you for fostering a supportive and inspiring research environment. The shared challenges and memorable moments made this journey much more enjoyable.

Beyond the academic sphere, I am deeply thankful for my friends from different parts of the world who have provided endless encouragement and emotional support, especially during the pandemic. Your presence in my life has been a constant source of motivation and strength.

A special acknowledgment goes to my furry friends—Didi the Cat, who kept me company during the lockdown, Milky the Cat, and Chickpea the Dog. Their unconditional love and adorable antics brought warmth and comfort to my life during this intense academic journey.

Most importantly, I want to express my heartfelt gratitude to my family. To my husband, Dr. Zhiyang He, for being my confidant, supporting me through everything from research to daily life, and providing me with unwavering encouragement and confidence. To my parents, for their continuous love, concern, and support, which have been a constant source of strength, even from afar.

Thank you all for being a part of this journey.

DEDICATION

to my loving parents, Hansen Liu and Pu Tang,
and my supportive husband, Zhiyang He.

Chapter 1

INTRODUCTION

Melanoma is a highly aggressive form of skin cancer originating in melanocytes, the cells responsible for producing melanin, the skin's pigment. Despite accounting for only about 1% of all skin cancers, melanoma causes the majority of skin cancer-related deaths due to its rapid growth and high potential for metastasis [4]. Early detection is essential, as timely intervention significantly improves patient survival rates, with a 5-year survival rate exceeding 99% if diagnosed at an early stage but dropping substantially for advanced cases [98]. However, the complex and variable nature of melanoma at the cellular level makes accurate diagnosis challenging, underscoring the importance of developing advanced diagnostic approaches to assist pathologists and improve outcomes [31].

The gold standard for diagnosing melanoma remains the visual assessment of hematoxylin and eosin (H&E) stained skin biopsies through microscopic examination. Pathologists perform diagnosis of melanoma by carefully analyzing the histopathological features of the sample, focusing on the presence of atypical melanocytes and their distribution, which provide critical clues for diagnosis. Like many other cancers, the diagnosis of melanoma has been significantly impacted by the advancement of Whole Slide Imaging (WSI). WSI offers high-resolution digital images of tissue sections, allowing pathologists to review samples in greater detail and facilitating the detection of subtle diagnostic features, which is especially crucial for complex diseases like cancer [28]. Despite these advantages, the process still heavily relies on the expertise and experience of the pathologist to identify relevant features, which can be labor-intensive and prone to human error. Diagnostic errors are common, with substantial inter- and intra-observer variability in the interpretation of melanocytic lesions [19, 2, 26, 24]. For instance, pathologists disagree in up to 60% of cases involving *melanoma in situ* and

stage T1a invasive cases [25]. The diagnostic variability poses a serious concern, as it can negatively impact treatment outcomes. Furthermore, the shortage of pathologists aggravates these challenges, undermining diagnostic accuracy and hindering medical progress. Therefore, a computer-aided diagnosis (CAD) system could play a pivotal role in reducing diagnostic uncertainties, easing pathologists' workloads, and improving the overall clinical process.

Deep learning methods have significantly advanced the automated analysis of WSI. Specifically, CNNs have been widely adopted for their ability to learn hierarchical patterns from image data. Nevertheless, due to the large size of gigapixel WSIs, CNNs are typically applied to smaller image patches extracted from the slides, and the features of these patches are aggregated for image-level predictions through Multiple Instance Learning (MIL) approaches [5, 40, 116, 16, 50]. While these methods provide valuable local information, they often fail to capture inter-patch correlations and overlook the broader context within the WSI. Transformer-based models have made strides in this direction by capturing both local and long-range global features [84, 13, 12, 121, 108]. However, they still operate on all the patches in a WSI indiscriminately, resulting in a relatively blind search for image patterns across the entire WSI.

Although both MIL and transformer-based methods offer improvements in WSI analysis, they still fall short in interpretability and clinical understanding, which limits their trustworthiness among pathologists. These methods typically attempt to analyze features across the entire WSI, rather than focusing on diagnostically relevant areas that a pathologist would naturally prioritize. Some methods incorporate region of interest (ROI) annotations to train ROI retrieval networks for a subsequent computer-aided diagnosis on specific ROIs [46, 122]. But this approach requires expert annotations for every data sample, and the ROI demarcation is often ambiguous given the irregular shapes of these regions. Alternatively, incorporating prior semantic knowledge from other models as weak supervision can potentially guide the model towards diagnostically important regions. These challenges underscore two key problems: 1) detecting diagnostically relevant entities within skin biopsies, and 2) enhancing the analysis and diagnosis of WSIs by incorporating both high-resolution, localized ROIs and the broader,

low-resolution context of the entire WSI. Addressing these challenges is essential to develop models that align more closely with clinical decision-making and improve diagnostic reliability in real-world applications.

This dissertation addresses the unique challenges in automated WSI analysis for melanoma diagnosis, with a focus on enhancing interpretability and diagnostic accuracy. During the exploration of these challenges, the need for efficient learning—arising from limited high-quality annotations and the large size of WSIs—consistently emerged as a critical obstacle. These constraints complicate the development of models that can effectively emulate the clinical expertise of pathologists. To overcome these challenges, this dissertation introduces several innovative approaches, including melanocyte detection on H&E stained biopsy images, virtual staining from H&E to immunohistochemistry (IHC), semantic attention guidance for attention-based classification models, and a multi-level region-attending network for WSI analysis. Together, these contributions have the potential to promote a more accurate, interpretable, and clinically applicable approach to WSI diagnosis.

In this dissertation, a series of projects are presented to demonstrate how advanced models tackle the specific challenges of WSI analysis in skin cancer research. The findings in this dissertation lay the groundwork for more reliable, interpretable, and efficient computer-aided tools, ultimately aiding in the clinical decision-making process for melanoma and potentially other forms of cancer.

Chapter 2 provides an overview of the three main skin biopsy datasets used in the projects: 1) an H&E WSI dataset for melanoma classification, 2) a ROI dataset with sparse and noisy annotation on melanocytic proliferation, and 3) an image dataset with H&E and SOX10 (an IHC staining type) paired WSIs for melanocyte detection.

Chapter 3 presents a project, *Identifying Melanocytic Proliferation* [69], which focuses on detecting and segmenting diagnostically relevant structures in skin cancer biopsy images. To address the challenge posed by extremely limited and low-quality annotations, this project

employs a weakly-supervised learning approach, effectively balancing annotation scarcity with model performance.

Chapter 4 introduces **VSGD-Net** [68], a novel detection model that learns melanocyte identification through the virtual staining from H&E-stained images into SOX10-stained images. **VSGD-Net** takes full advantage of SOX10's unique ability to highlight melanocytes in distinct colors, enabling the model to simultaneously learn both detection and virtual staining tasks. Building upon **VSGD-Net**, we present **CC-WSI-Net** [72], which learns to seamlessly synthesize IHC-stained WSIs. The synthetic WSIs from **CC-WSI-Net** are free from the patch stitching artifacts, offering a more realistic appearance that could potentially enhance the accuracy and efficiency of WSI analysis.

Chapter 5 presents **SAG** [70], an innovative flexible framework that integrates semantics-aware attention guidance into any attention-based classification models, including transformer and MIL methods. In this work, we propose a heuristic attention-generation method to convert diagnostically relevant entities to heuristic guidance signals. These semantic guidance signals are employed to supervise the attention learning process through a novel attention guiding loss. The application of **SAG** effectively enhances the classification performance on two cancer datasets and represents a crucial step towards the interpretability of WSI diagnosis models.

Chapter 6 outlines **MiRA**, a novel multi-level region-of-interest attended network designed for WSI analysis. **MiRA** employs dynamic ROI retrievals using weak supervision, allowing it to focus on diagnostically relevant areas within the WSI while also capturing broader, global image patterns. This dual attention mechanism mimics the clinical pathological decision-making process, where pathologists not only examine the high-level structure of the entire slide but also zoom in on specific regions for detailed analysis. By incorporating both global WSI features and localized ROI characteristics, **MiRA** leverages a multi-level design that

effectively integrates information from multiple scales. This architecture significantly enhances diagnostic performance, making it a promising tool for improving melanoma diagnosis in a clinical setting.

Chapter 7 summarizes all the projects and discusses the possible future work.

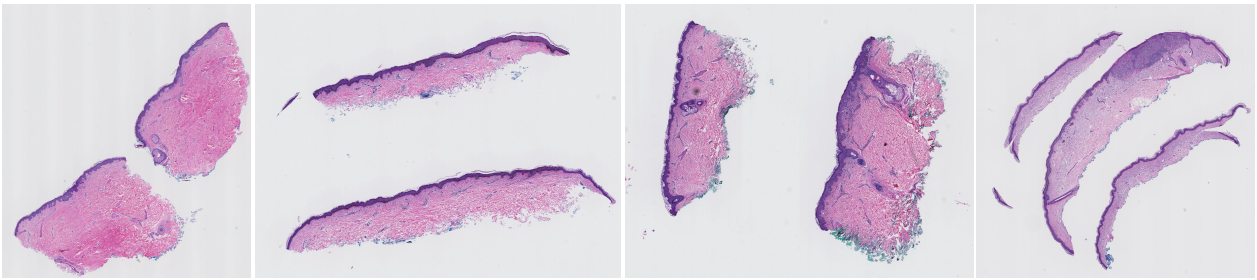
Chapter 2

DATASETS

This chapter introduces the main datasets used in the following chapters for melanoma WSI analysis. To develop interpretable and effective models, it's important to understand the inherent properties and challenges associated with the skin cancer. Here we provide a description on the M-PATH dataset, the skin cancer ROI dataset, and the melanocyte detection dataset. We also include the statistics and detailed preprocessing of these datasets.

2.1 M-PATH Dataset

The dataset consists of 222 H&E stained WSIs of skin biopsy images collected for the M-PATH study (R01CA151306) [25]. This study was approved by the Institutional Review Board at the University of Washington with protocol number STUDY00008506. The 222 cases were interpreted by a consensus panel of three experienced dermatopathologists using the MPATH-Dx (Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis) classification



(a) MMD (b) MIS (c) pT1a (d) pT1b

Figure 2.1: Example H&E stained WSIs from the M-PATH dataset.

schema. The consensus assessments were categorized into five simplified classes based on the progression risk, which spans from low-risk dysplastic nevi to higher-risk invasive melanoma. Example diagnostic terms for each MPATH-Dx class are as follows: (I)mildly dysplastic nevi, (II) moderately dysplastic nevi, (III) melanoma in situ and severely dysplastic nevi, plastic nevi, (IV) invasive melanoma stage T1a, and (V) invasive melanoma stage $\geq$ T1b.

In our diagnosis research, the five classes were grouped into four diagnostic classes due to limited sample size in Classes I and II and the low clinical progression risk of both Class I and Class II. The diagnostic terms we use for each class are as follows: 1) Class I-II: mild and moderate dysplastic nevi (MMD), 2) Class III: melanoma in situ (MIS), 3) Class IV: invasive melanoma stage pT1a (pT1a), and 4) Class V: invasive melanoma stage $\geq$ pT1b (pT1b). We randomly split the dataset into training, validation and testing subsets with 89, 22 and 111 WSI samples. The distribution and sample image of the diagnostic classes are provided in Table 2.1 and Figure 2.1.

Table 2.1: Statistics of the MPATH dataset.

Diagnostic	**Number of WSIs**			
Class	**Training**	**Validation**	**Testing**	**Total**
MMD	23	6	29	58
MIS	24	5	30	59
pT1a	26	6	30	62
pT1b	16	5	22	43
Total	89	22	111	222

2.2 Skin Cancer ROI Dataset

This dataset consists of 227 ROIs that best represent the diagnostic classifications within the M-PATH dataset. These ROIs were selected by the aforementioned consensus panel of pathologists and were extracted from the H&E stained WSIs at 10x magnification. The distribution of ROIs across MPATH-Dx categories is as follows: 29 cases in class I (benign), 49 cases in class II (moderate dysplastic nevi), 67 cases in class III (melanoma in situ), 50 cases in class IV (stage pT1a invasive melanoma), and 32 cases in class V (stage pT1b or higher invasive melanoma). For experimental use, we randomly split the dataset into training, validation, and testing with 174, 19, and 34 samples, respectively.

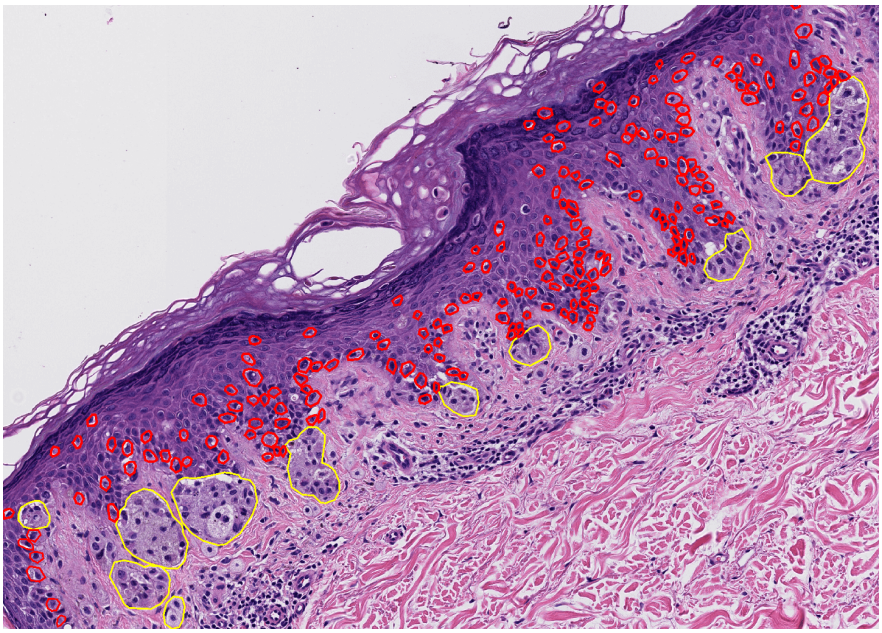
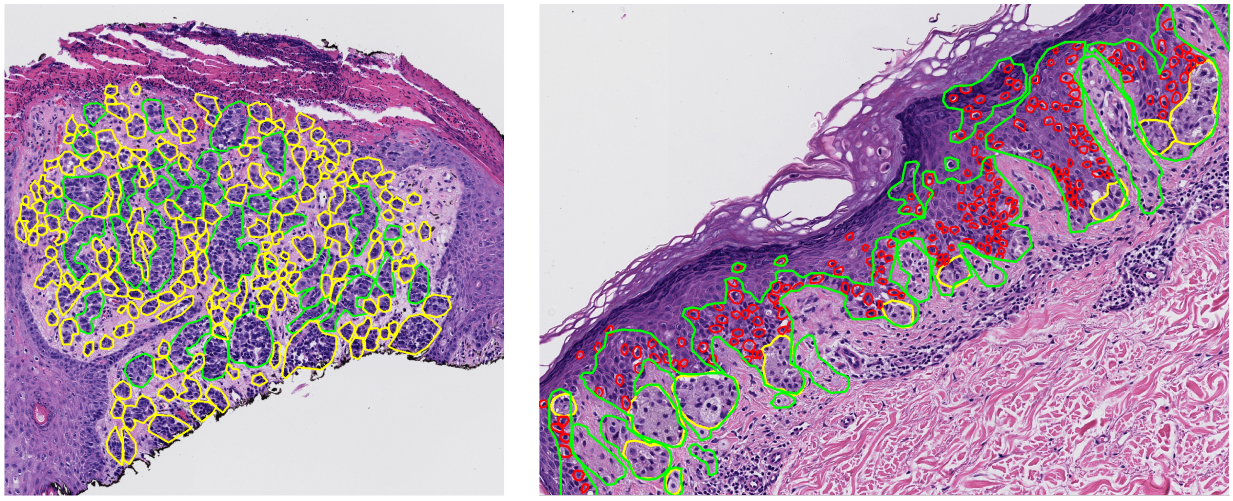


Figure 2.2: **Examples of melanocytic proliferations:** we use red polygons to mark the singly dispersed melanocytes and yellow polygons to mark the melanocytic nests, which represent proliferations of melanocytes.

We enlisted an additional expert pathologist (Dr. Mojgan Mokhtari) to label the melanocytic proliferations within the ROIs for training the segmentation model described

in Chapter 3. Accurate identification of melanocytic proliferations, including both single melanocytes and melanocytic nests, as shown in Figure 2.2, is critical for diagnosis. However, the annotation is challenging due to the varied shapes and sizes of nests, the high density of melanocytes in each slide, and the need for costly, expert annotation, making it difficult to gather sufficient labels for training deep learning models. Given the difficulties, the pathologist only partially marked the ROIs (i.e. not every ROI is marked and not every melanocytic proliferation is marked), and drew polygons around many melanocytes instead of each individual cell. Example markings are provided in Figure 2.3. To improve annotation reliability, two additional pathologists reviewed these markings. While this approach yields a sparse and somewhat noisy annotation, it greatly reduces the time required for expert involvement.



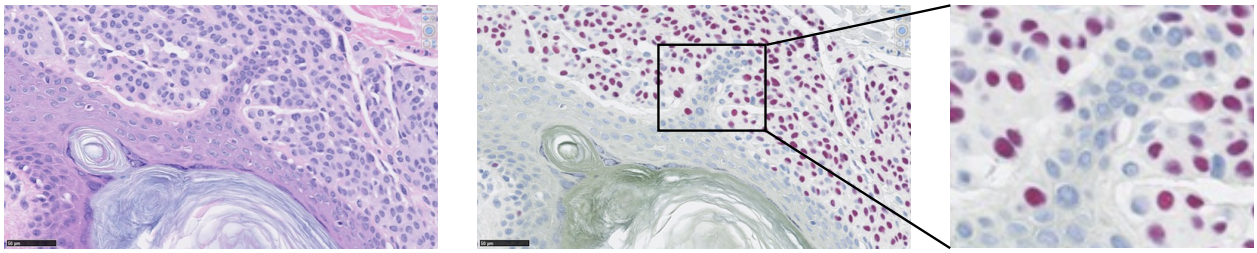
(a) Sparse annotations

(b) Noisy annotations

Figure 2.3: **Melanocytic proliferation 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 single melanocytes and the melanocytic nests separately, while they are actually labeled as the green markings in our annotations.

2.3 Melanocyte Detection Dataset

A skin biopsy dataset was curated to facilitate learning melanocyte detection on H&E stained WSIs, as detailed in Chapter 4. This dataset consists of skin tissue from paraffin-embedded blocks of 15 cases, randomly chosen from historical cases at Dermatopathology Northwest laboratory. These cases represent three samples from each MPATH-Dx diagnostic category [25, 87]. Each skin biopsy sample was sectioned into 4–6 thin slices for microscopic examination, resulting in a total of 75 slices at 20x magnification. To obtain accurate melanocyte ground truth labels without incurring high annotation costs, we utilized Sox10, an immunohistochemistry (IHC) stain specific to melanocytic nuclei. Initially, each WSI was stained with H&E (Figure 2.4a). The same tissue sections were then carefully destained and re-stained with Sox10, which highlights melanocyte nuclei in red, while nuclei of other cells appear blue, providing a clear distinction between melanocytes and non-melanocytes (Figure 2.4b).



(a) H&E Staining (b) Sox10 Staining – melanocytes are red (c) Crop from Sox10

Figure 2.4: Sample H&E stained image and Sox10 stained image. The Sox10 stain highlights the nuclei of melanocytes in red, while the nuclei of other cells appear in blue.

To generate groundtruth labels for melanocyte detection, we introduce a pseudo-automatic procedure. First, we carefully registered raw Sox10 stained images to their corresponding H&E stained images using Histokat software[1][75]. We then trained a Random Forest classifier on a set of 100 manually labeled melanocytes in Sox10 images to generate the preliminary

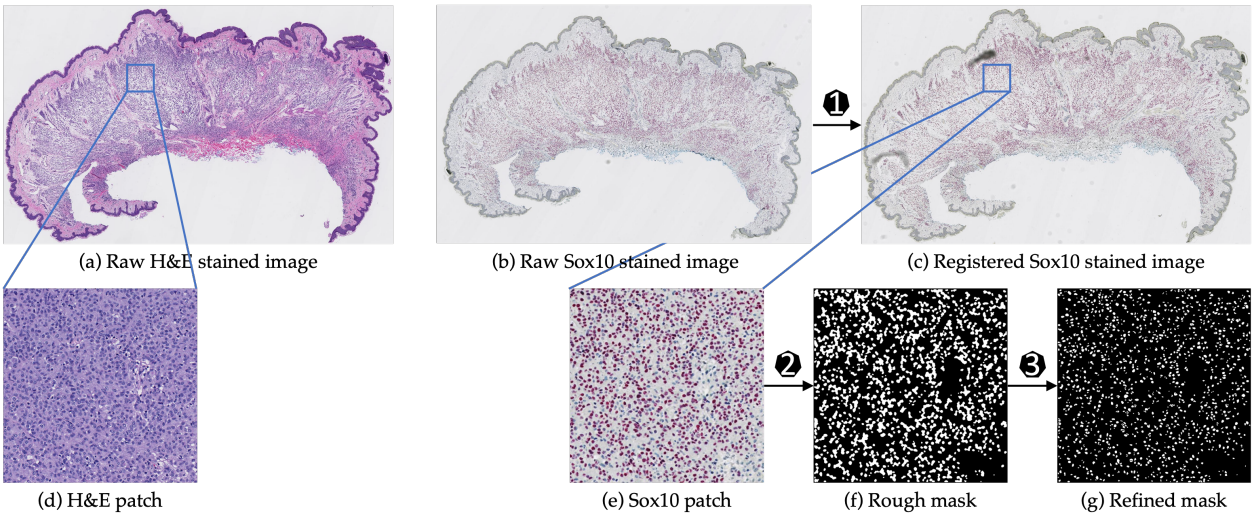


Figure 2.5: **Melanocyte groundtruth generation:** First, we register raw Sox10 images (b) into aligned Sox10 images (c) using template H&E images (a). Then, we apply a Random Forest classifier to classify pixels into melanocyte or non-melanocyte. At last, the pretrained NuSeT [115] separates touching nuclei and refine the masks.

melanocyte segmentation masks. To refine these masks, we applied NuSeT [115], a pretrained nuclei detection model, to accurately separate touching nuclei. This procedure yielded precise melanocyte masks, which serve as ground truth labels for the study (see Figure 2.6).

To fit images into memory while retaining sufficient details, we cropped the registered paired images into 256x256 patches at 10x magnification. Background patches were excluded, resulting in a total of 25,314 patches to use. We allocated 9,652 paired patches from five patients as the testing set, ensuring that patient data in the testing set was completely absent from the training and validation sets. Both the training and testing sets included samples representing the full range of MPATH-Dx diagnostic classes to ensure a comprehensive and unbiased evaluation.

[1] `https://histoapp.mevis.fraunhofer.de/`

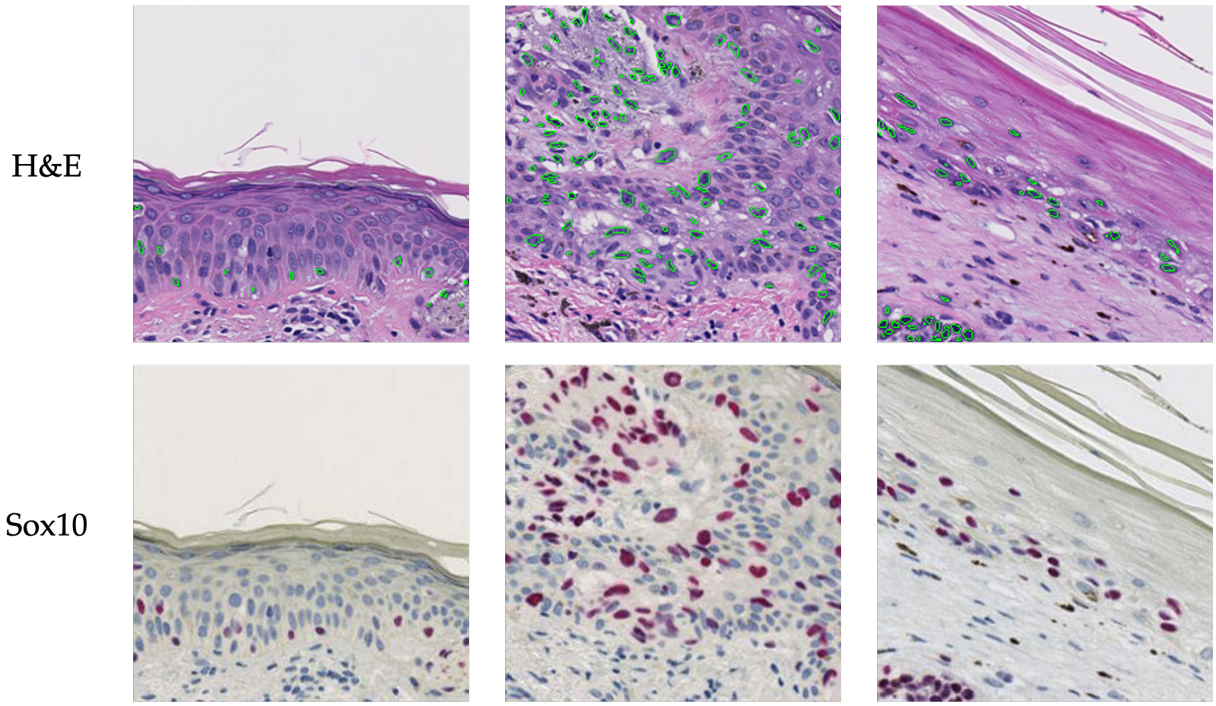


Figure 2.6: **Nuclei groundtruth:** The boundaries of the melanocyte masks, created from the Sox10 stained patches (bottom row) using our pseudo groundtruth generation method, are overlayed onto the H&E stained patches (top row) in green color. The precise alignment of the cell boundaries across these images shows that the melanocyte masks serve as reliable groundtruth labels for this dataset.

Chapter 3

SEGMENTING MELANOCYTIC PROLIFERATION USING IMPERFECT ANNOTATIONS

3.1 Introduction

The accurate diagnosis of melanoma and its precursors relies on the detailed assessment of architectural growth patterns of melanocytes. This process involves identifying the microanatomical localization of melanocytes within the skin (e.g., intraepidermal, dermal-epidermal junction, intradermal) and characterizing their architectural arrangements. For example, melanoma in situ exhibits confluent melanocytic growth of single cells and nests along the epidermal base, often extending to mid-to-upper epidermal layers (pagetoid spread). Invasive 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 can exhibit numerous patterns of growth, this study focuses on two fundamental patterns: single cell dispersion and nests.

The complexity of diagnosing melanocytic lesions on histopathology arises from variable architectural patterns and cytomorphological features. 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? Specifically, we aim to detect singly dispersed melanocytes and nests, providing histological insights to support pathologists. For simplicity, we refer to these patterns collectively as melanocytic proliferations throughout this project.

Recent advancements in deep learning have enabled researchers to apply neural networks for histological image segmentation across various medical domains. For instance, CNNs have

been used to identify tumor regions and ducts in breast cancer studies [37, 62]. Similarly, researchers have leveraged CNNs to segment prostate cancer grading to aid diagnosis [65, 51]. In dermatopathology, Kucharski *et al.*, [60] utilized Autoencoders for patch-level segmentation of melanocytic nests in skin cancer.

Building on these advances, we developed a weakly supervised pipeline to identify image-level melanocytic proliferations. Given the annotation challenges discussed in Chapter 2.2, our approach emphasizes training with sparse and noisy annotations from skin biopsy images. To address these limitations, we employed weighted loss functions that effectively accommodate imperfect labels. The proposed method achieves state-of-the-art performance in segmenting melanocytic proliferations, validated against ground truth annotations provided by experienced dermatopathologists.

3.2 Related Work

Semantic segmentation is a common task that assigns a semantic label to every pixel in an image. For instance, the Fully Convolutional Network (FCN) [74] adapts CNNs to take input images of arbitrary size and output its corresponding mask. FCNs have been widely used in biomedical image analysis [6, 53, 90]. Building on the structure of FCNs, U-Net [89], with its encoder-decoder structure and skip connections, became a widely adopted model for segmentation in histopathology and other domains [106, 113]. Variants like 3D U-Net [17] and attention U-Net [86] have further extended these capabilities, addressing tasks that require 3D contextual information or region-specific focus [14, 33].

Unlike semantic segmentation, instance segmentation distinguishes individual objects, which is less frequently used in medical image analysis. However, it has proven valuable in specific tasks like gland segmentation in colon histology [111, 10], duct segmentation in breast biopsies [62], and nuclei segmentation in microscopy images [55]. Mask R-CNN stands out in this domain, employing a two-stage approach with a Region Proposal Network (RPN) for coarse localization followed by fine-grained segmentation and classification. For our study, we adopt Mask R-CNN to identify and segment melanocytic proliferations.

A notable work by Kucharski et al. [60] addressed melanocytic nest segmentation using a convolutional autoencoder, representing the state-of-the-art in this area. This approach combines a convolutional autoencoder for feature extraction with a segmentation head trained on annotated whole slide images (WSIs). While effective, the model's reliance on small patch sizes (128×128) due to memory limitations leads to a loss of contextual information critical for segmenting melanocytic proliferations. This limitation reduces the segmentation and detection accuracy. Inspired by this work, we aim to enhance the performance by focusing on segmenting melanocytic proliferations with improved contextual awareness.

3.3 Methodology

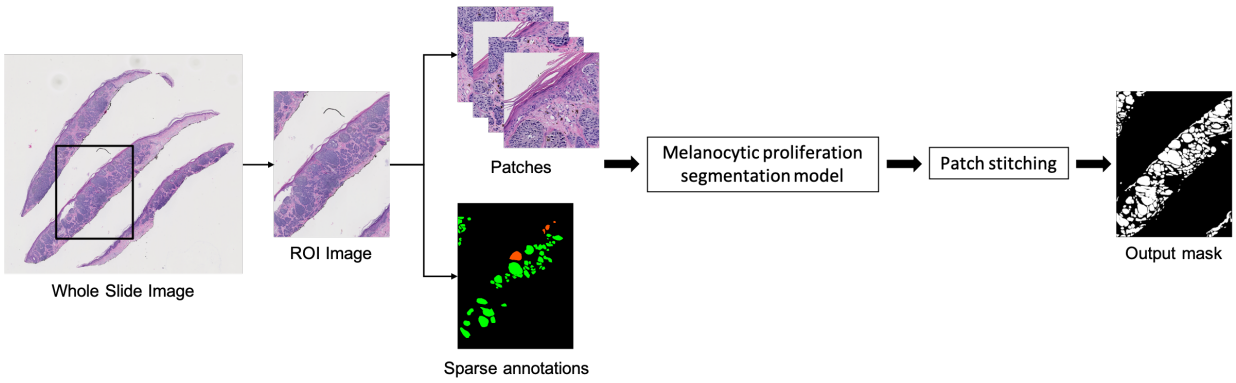


Figure 3.1: **Melanocytic proliferation segmentation pipeline:** this pipeline enables training from sparse annotations using the Mask R-CNN model with different loss functions, and aggregates results on patches to provide an image-level mask that can be used in further diagnosis. Note: only the middle tissue slice in the WSI is the important region of interest, even though parts of the other two slides fell into the box.

Figure 3.1 shows our proposed pipeline for melanocytic proliferation segmentation, 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 preprocessing procedure. 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.

3.3.1 Data preprocessing

In this study, we leverage the ROI dataset described in Chapter 2.2 for learning melanocytic proliferation. Even if ROI images are cropped from the original 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 [21, 43]. In Mask R-CNN, the region proposal network predicts the candidate anchor boxes which are likely to contain an object and feed the anchor boxes for the downstream heads to classify and segment. Since the default anchor box sizes in the pretrained Mask R-CNN are 32, 64, 128, 256, and 512, we split the images into 1000×1000 patches, so that the anchor boxes can cover most of the labeled objects in the patches; this avoids training new layers as the dataset is too small. The patches are 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 artifacts 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.

3.3.2 Model

Among the processed data patches, most contain only a few small-sized melanocytic proliferations, leaving the majority comprising non-target tissues. This imbalance motivates us to adopt the Mask R-CNN [41], a widely-used instance segmentation model. Mask R-CNN offers a significant advantage as a two-stage model, illustrated in Figure 3.2. In the first stage, a Region Proposal Network (RPN) identifies candidate regions that likely contain target entities. In the second stage, these regions are refined to adjust anchor boxes, generate segmentation masks, and produce classification results. This architecture enables efficient filtering of non-target tissues, making it particularly suited for our task.

To address the limitation of data scarcity, *i.e.*, we only have 130 images partially labeled in the training dataset, we leverage transfer learning via CNNs originally pretrained on natural images. We used an off-the-shelf implementation of Mask R-CNN from detectron2 [109], 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.

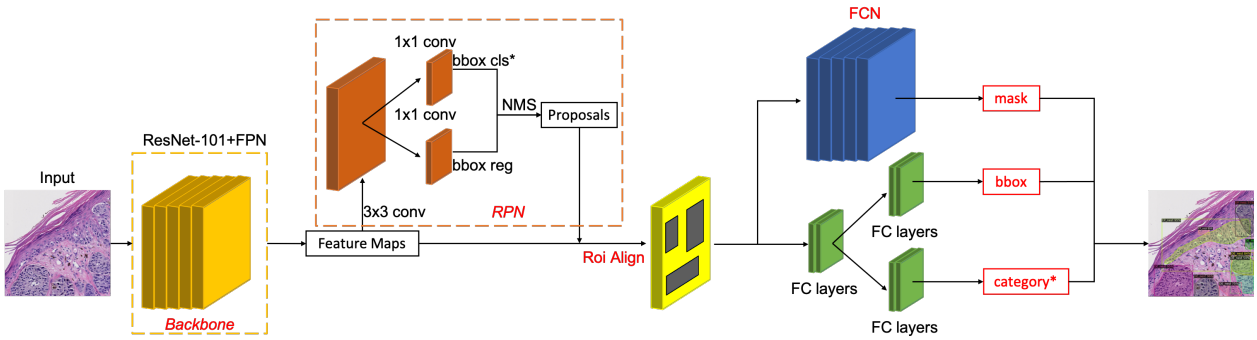


Figure 3.2: **Overview of Mask R-CNN model architecture:** We use ResNet-101+FPN as the backbone to extract feature maps from the input image. Combining the feature maps and the anchor box results from Region Proposal Network (RPN), fixed-size feature maps are fed into three prediction heads (classification, 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.

3.3.3 Loss function

The original Mask R-CNN model was developed for instance segmentation on a fully labeled dataset. We hereby describe our modification to the loss function 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 regression loss in the prediction head. (5) L_{mask}: Segmentation loss in the prediction head.

During Mask R-CNN training, $L_{\text{rpn\_loc}}$, $L_{\text{box\_reg}}$, L_{mask} back-propagate values only from positive samples. In contrast, $L_{\text{rpn\_cls}}$ and L_{cls} leverage both labeled and unlabeled regions to determine the presence of an instance in the anchor box. Given the partially labeled nature of our dataset, treating unlabeled regions as background (i.e., not nests) introduces bias to the 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 experimented with two alternative 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:

$$L_{\text{WCE}} = -\sum_{i}(w \cdot y_i \cdot \log(\hat{p_i}) + (1 - y_i) \cdot \log(1 - \hat{p_i})) \tag{3.1}$$

where $y_i \in \{0, 1\}$ is the ground truth label, indicating 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 positive category.

Focal Loss was first introduced in [66], 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

$$L_{\text{WFL}} = -\sum_{i}(w \cdot y_i \cdot (1 - \hat{p_i})^{\lambda} \cdot \log(\hat{p_i}) + (1 - y_i) \cdot \hat{p_i}^{\lambda} \log(1 - \hat{p_i})) \tag{3.2}$$

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 [66]. The definitions of y_i, $\hat{p_i}$,

and w remains the same as equation 3.1.

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 (Chapter 3.4).

3.3.4 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.

3.3.5 Evaluation metrics

To make our model comparable with the state-of-the-art melanocytic nest segmentation method [60], 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). The definitions of Dice coefficient and mIoU are given in equation 3.3 and equation 3.4.

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3.3}$$

$$\text{mIoU} = \frac{1}{2} \times \left(\frac{TP}{TP + FP + FN} + \frac{TN}{TN + FP + FN}\right) \tag{3.4}$$

3.4 Results and Ablation Study

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. To have a fair evaluation, we asked our expert pathologist to thoroughly label the melanocytic nests in our test set, as shown in Figure 2.2, 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.

3.4.1 Main results

We re-implemented the autoencoder model [60], and trained it following all the detailed steps as described. Table 3.1 provides a quantitative comparison between the autoencoder approach and our method across various loss functions, including the default cross-entropy loss. Due to the limited contextual information fed into the model, the autoencoder produces less accurate and noisier segmentation results. Additionally, Figure 3.3 presents qualitative comparisons, where ground truth annotations and segmentation results are overlaid on H&E-stained images. The first two rows illustrate examples where our method achieves superior segmentation performance compared to the autoencoder and closely aligns with the ground truth. In Figure 3.3 (c), our model predicts a false positive proliferation in the middle layer of the epidermis, which contains keratinocytes with "halo" regions surrounding the nuclei. This misclassification arises because intraepidermal melanocytes often share similar features with "halo" regions [76]. Overall, our method outperforms the previous SOTA autoencoder across all key metrics.

Method	**Dice**	**mIOU**	**Acc**	**Sensitivity**	**Specificity**
Autoencoder[60]	0.655	0.693	0.907	0.795	0.921
Mask R-CNN w. CE loss	0.698	0.728	0.928	0.749	0.950
Mask R-CNN w. WCE loss	0.709	0.733	0.925	**0.826**	0.937
Mask R-CNN w. FL loss	**0.732**	**0.756**	**0.939**	0.746	**0.964**

Table 3.1: **Quantitative results:** Dice score, mIOU, Accuracy (Acc), Sensitivity and Specificity for all methods. The best performances are highlighted in bold font in this table.

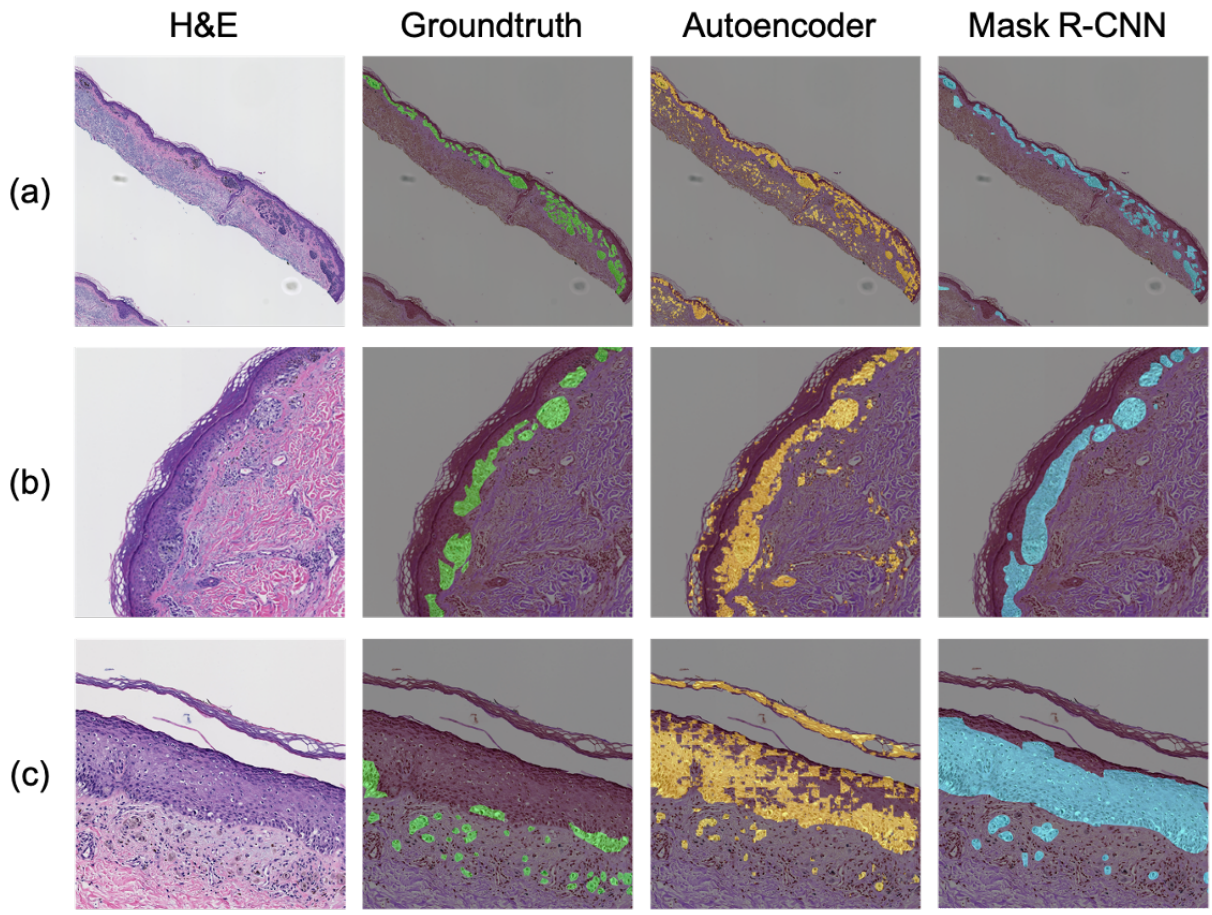


Figure 3.3: **Qualitative comparison between our model and SOTA autoencoder [60]:** From left to right, each column shows examples of H&E stained ROI images, groundtruth annotation, Autoencoder results, and Mask R-CNN results.

3.4.2 Ablation study

To investigate the impact of the weight in the loss functions on segmentation performance, we conducted experiments using various weight values for WCE and FL. As summarized in Table 3.2, the WCE loss achieved optimal performance when $w = 3$, and FL performs the best when $w = 2$. The comparison between default cross entropy (w=1) and other weighted loss functions demonstrates that incorporating weights significantly improves performance,

particularly in sparsely annotated datasets like ours.

Loss function	Weight	Dice	mIOU	Acc	Sensitivity	Specificity
Weighted Cross Entropy (WCE)	$w = 1$	0.698	0.728	0.928	0.749	0.950
	$w = 2$	0.708	0.735	0.928	0.782	0.947
	$w = 3$	**0.709**	**0.733**	**0.925**	**0.826**	**0.937**
	$w = 5$	0.697	0.723	0.920	0.831	0.931
	$w = 8$	0.693	0.718	0.915	0.856	0.923
	$w = 12$	0.657	0.686	0.895	0.899	0.894
Focal Loss (FL)	$w = 1$	0.723	0.749	0.938	0.727	0.964
	$w = 2$	**0.732**	**0.756**	**0.939**	**0.746**	**0.964**
	$w = 3$	0.729	0.753	0.938	0.748	0.962
	$w = 5$	0.730	0.754	0.938	0.755	0.961
	$w = 8$	0.722	0.748	0.937	0.736	0.962
	$w = 12$	0.732	0.755	0.937	0.773	0.958

Table 3.2: **Ablation experiments for weighted cross entropy (WCE) and focal loss (FL):** All models were evaluated on our fully-labeled test dataset. The average performance from 10 runs are reported.

3.4.3 Discussion

As shown in Table 3.1 and Figure 3.3, our proposed method achieved better results than the autoencoder [60] in all metrics. The ability to accurately identify melanocytic proliferations offers valuable histological insights, helping pathologists focus on critical regions and reducing their workload. Additionally, this approach can aid students in gaining a clearer understanding of the foundational steps in the diagnostic process.

We selected Mask R-CNN as our model architecture due to its robustness to noise. Unlike the autoencoder, which often misclassifies small or irrelevant entities as melanocytic proliferations (Figure 3.3), Mask R-CNN leverages anchor boxes to prioritize regions of interest and filter out irrelevant background. Its use of non-maximum suppression further minimizes noise around target instances, enhancing segmentation accuracy.

3.5 Summary

Identifying melanocytic growth patterns, such as single-cell dispersion and nests, is a critical step in assessing melanocytic lesions. In this study, we proposed 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. Our model holds promise as an initial step in an automated diagnostic pipeline. Once we accurately recognize melanocytic proliferations and how they are situated in cutaneous microanatomy, we can incorporate other works to extract the aforementioned features. Future research could extend this work on a larger dataset, incorporate additional diagnostic features, and integrate advanced classification methods, such as multi-instance learning and Transformer architectures, to develop a comprehensive diagnostic tool.

Chapter 4

VIRTUAL STAINING GUIDED MELANOCYTE DETECTION

4.1 Introduction

In biomedical image analysis, the automatic detection of specific cell types in microscopy images is crucial to a broad spectrum of biological research and clinical practices. Accurate identification of particular cell types supports the interpretation of biopsies and the diagnosis of various diseases. For example, diagnosing melanoma requires the assessment of the distribution disorder of melanocytes[1] under the microscopic examination of H&E-stained skin biopsies by pathologists. Although Chapter 3 introduced a promising approach for segmenting melanocytic proliferations, it lacks the capability to identify individual melanocytes. This limitation hinders its ability to provide critical diagnostic information, such as melanocyte maturation and distribution patterns.

Identifying melanocytic populations is challenging on routine H&E-stained slides due to their visual similarity to other cells. Pathologists may rely on special additional immunohistochemistry (IHC) stains, such as Sox10 – a specific immunomarker for melanocytes (Figure 2.4c) – to address this issue. However, Sox10 staining is not routinely obtained in clinical practice due to its high cost. Therefore, building computer-aided detection methods would support the melanoma diagnosis workload and improve diagnostic accuracy.

Deep learning is widely adopted in various computer vision tasks, with architectures such as CNNs [36, 49], U-Net [11], and R-CNN [103] being applied to localize nuclei in H&E images. However, these methods face limitations when detecting specific cell types, as they cannot fully utilize information from multiple stainings/modalities and struggle with inter-class visual

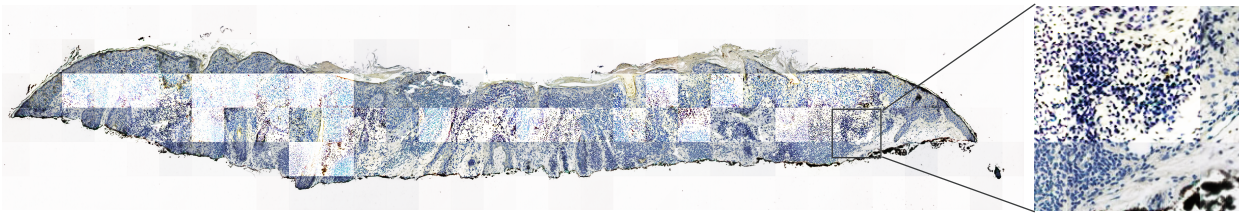
[1]For example, melanoma in situ exhibits confluent growth of single and nested melanocytes at the epidermal base and/or extension into the mid-to-upper levels of the epidermis.

similarities. Generative Adversarial Networks (GANs) have shown promise in tasks such as virtual staining and image synthesis. For example, the unsupervised CycleGAN [125] and the supervised conditional GAN [52] architectures have been leveraged to synthesize medical images across modalities, such as MR to CT [114, 44], and H&E to IHC [112, 80]. Despite these advancements, current GAN approaches often fail to incorporate feedback from other tasks to guide the image synthesis. While some studies [34, 120] integrate a segmentation network after the generator, they overlook the intermediate features generated during the image synthesis process, which are empirically critical for improving downstream performance.

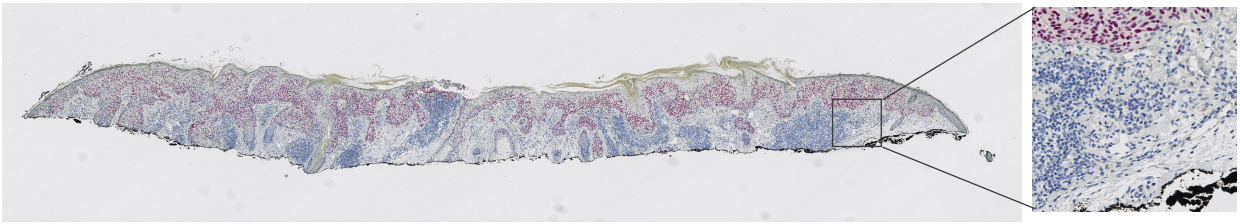
To address these challenges, we propose **VSGD-Net**, a novel virtual-staining-guided detection network that simultaneously tackles detection and virtual staining tasks. By exploiting the hidden correlations between two image modalities, **VSGD-Net** enhances both detection accuracy and image synthesis quality. We evaluate our method using a curated dataset (Chapter 2.3) containing H&E and Sox10-stained biopsy images and demonstrate the importance of intermediate features through extensive experiments.

Building upon **VSGD-Net**, we further introduce **CC-WSI-Net**, designed to overcome the limitations in WSI synthesis. Due to the gigapixel size limit, most virtual staining approaches [68, 112, 18] stitch patches to obtain WSIs, which suffer from content, texture, and color inconsistencies (shown in Figure 4.1). These inconsistencies often hinder the applications of the synthesized WSIs. Alternatively, slide-level synthesis techniques often fail to ensure cell-level accuracy required for clinical use [39]. **CC-WSI-Net** addresses these gaps by incorporating a content and color consistency module, enabling seamless WSI generation while preserving diagnostic fidelity. Extensive image quality analysis and a subjective survey conducted with three pathologists confirm the clinical quality of the synthetic WSIs.

In this chapter, we present both **VSGD-Net** and **CC-WSI-Net**. **VSGD-Net** is the first to investigate the cell detection problem using image synthesis features between two stainings, achieving state-of-the-art performance in melanocyte detection and virtual staining. Extending its capabilities, **CC-WSI-Net** facilitates the synthesis of diagnostically accurate, seamless WSIs, paving the way for future advancements in computer-aided diagnosis. Together,



(a) Synthetic Sox10 WSI



(b) Ground Truth Sox10 WSI

Figure 4.1: **Color and Content Inconsistencies**: (a) shows a Sox10 WSI generated by VSGD-Net [68], while (b) displays the corresponding ground truth Sox10 WSI.

these frameworks significantly improve the accuracy, efficiency, and clinical relevance of virtual staining and cell detection in histopathology.

4.2 Related Work

4.2.1 Nuclei Detection

Deep learning-based nuclei detection methods have been widely studied. As a variant of the fully convolutional network (FCN) [74], U-Net [89] made a huge impact on the medical image research community. Many researchers extended the U-Net structure [89] into more efficient variants to identify nuclei in histopathological images, for example, R2U-Net [3], U-Net++[124], Micro-Net [88], and Triple U-Net [119]. To incorporate nuclei contour-aware modules, Zhou *et al.*, presented CIA-Net [123] which contains two task-specific decoders to learn either the nuclei or the contours. Similarly, Schmidt *et al.*, proposed StarDist [92] to localize nuclei via star-convex polygons. In the task of detecting nuclei of specific cells, Graham *et al.*, proposed Hover-Net [36] by utilizing three downstream branches, namely

segmentation, classification, and a novel Hover branch, which used the horizontal and vertical distance maps to segment attached nuclei. For better distance-map generation, Gao *et al.*, presented the two-stage CHR-Net [30], which leveraged the W-Net structure [110] and high-resolution feature extractors, and achieved the new state-of-the-art performance.

Another line of approaches, e.g. Mask RCNN [41], have also achieved promising results in nuclei instance segmentation [73, 103, 104]. The feature pyramid network (FPN) backbone allows the model to extract features in multiple scales and feed into the region proposal network (RPN) to generate reasonable instance candidates in varying sizes for downstream tasks like segmentation and classification. Our proposed model, **VSGD-Net**, also takes advantage of the FPN and RPN modules to better exploit the intermediate features for nuclei detection.

4.2.2 Image-to-Image Translation

Generative Adversarial Networks (GANs), first introduced by Goodfellow *et al.*, [35], leverage adversarial learning to train a generator and a discriminator in a minimax zero-sum game. This framework has inspired numerous GAN variants tailored for diverse image synthesis tasks. For instance, conditional GANs (cGANs) incorporate additional constraints by conditioning the generator and discriminator on auxiliary information [83]. Notable cGAN derivatives include Pix2Pix [52] and Pix2PixHD [105], which have become benchmarks for paired image-to-image translation, as well as StyleGAN [57] and StyleGAN2 [58], which allow control over image style and quality. These GAN-based approaches have been widely adopted in histopathological image analysis for applications like stain normalization, modality conversion, and virtual staining. For example, Stain-GAN [94] addresses biopsy stain normalization, while CycleGAN [125] enables unpaired domain translation and has been adapted for virtual staining tasks, such as mapping H&E to IHC [112].

However, many GAN-based methods face limitations when bridging domains without paired data or enforcing domain-specific constraints. For instance, while CycleGAN employs a cycle-consistency loss to ensure that translations between domains preserve the original

content, it lacks pathological constraints necessary for reliable virtual staining in biomedical contexts. Extensions such as pathology-consistent CycleGAN [71] have addressed this by incorporating pathological property alignment across domains. Despite these advances, traditional GANs often overlook the latent features generated during synthesis, limiting their utility for downstream tasks such as cell detection and segmentation. Recent studies have begun integrating task-specific networks, such as R-CNN-based detectors, into GAN frameworks to enhance downstream task performance [67, 34, 120]. Nevertheless, these approaches fail to leverage intermediate features effectively, leaving room for improvement in joint task optimization.

Diffusion models, a rapidly emerging class of generative models, offer an alternative approach to image synthesis by iteratively refining random noise into structured data. Unlike GANs, which rely on adversarial learning, diffusion models use a denoising framework for generation. Ho *et al.*, introduced the Denoising Diffusion Probabilistic Model (DDPM) [45], where Gaussian noise is incrementally added to an image and later removed to reconstruct it. This process offers flexibility and stability, avoiding issues like GAN mode collapse. Building upon the DDPM, recent works have adapted diffusion models for tasks like image-to-image translation. For example, Li *et al.*, proposed the Brownian Bridge Diffusion Model (BBDM) [64], which leverages a latent-space diffusion process to transform a source image into a target image. However, the accuracy and reliability of synthetic images remain uncertain in tasks such as virtual staining.

Building upon prior studies and addressing existing challenges, our proposed **VSGD-Net** jointly optimizes image synthesis and cell-type detection by leveraging shared intermediate features. By integrating supervision from both tasks, it significantly enhances the diagnostic accuracy of synthetic images and the precise localization of melanocytes.

4.2.3 WSI Synthesis

While virtual staining methods have shown success in translating image modalities, they typically operate at the patch level, limiting their utility for pathologists who rely on WSIs

for comprehensive diagnosis. To address this problem, researchers have explored various strategies to generate WSIs using advanced architectures. For example, Harb *et al.* proposed a diffusion-based method that starts by generating low-resolution WSIs from noise, progressively enhancing resolution through a coarse-to-fine sampling scheme. Lahiani *et al.* introduced a Perceptual Embedding Consistency (PEC) loss to improve CycleGAN, achieving seamless color, contrast, and brightness consistency across patches [61]. Similarly, Sun *et al.* developed the Bi-directional Feature Fusion GAN (BFF-GAN), which combines global and local features via a dual-branch architecture, addressing inconsistencies in color and brightness between adjacent patches [99]. Although these methods achieve promising results in generating WSIs with consistent staining, they often neglect clinical quality and diagnostic effectiveness in their evaluations—an essential consideration for their application in real-world clinical practice. Hence, to address both the inconsistency issues and the clinical quality of the synthetic WSIs, we extend **VSGD-Net** to **CC-WSI-Net**, which synthesize diagnostically equivalent seamless WSIs.

4.3 VSGD-Net

Figure 4.2 illustrates the architecture of **VSGD-Net**. We built the generator G based on an adapted UNet [89] structure with ResNet-50 [42] as the encoder. The encoder extracts multi-scale, high-dimensional features from input H&E images, while the decoder, comprising five deconvolution layers, translates these features into Sox10-stained target images. To emphasize melanocyte regions without increasing model complexity, attention blocks inspired by CBAM [107] are integrated into the skip connections. These blocks combine channel attention using a 3-layer MLP and spatial attention using convolution to generate dimension-specific attention maps (see Figure 4.3).

While the generator G learns the virtual staining process, the discriminator D attempts to differentiate real and synthesized Sox10 images. Following Pix2PixHD [105], D adopts a multi-scale architecture with two identical CNN discriminators operating at coarse and fine levels. The coarse-level discriminator processes images downsampled by a factor of 2

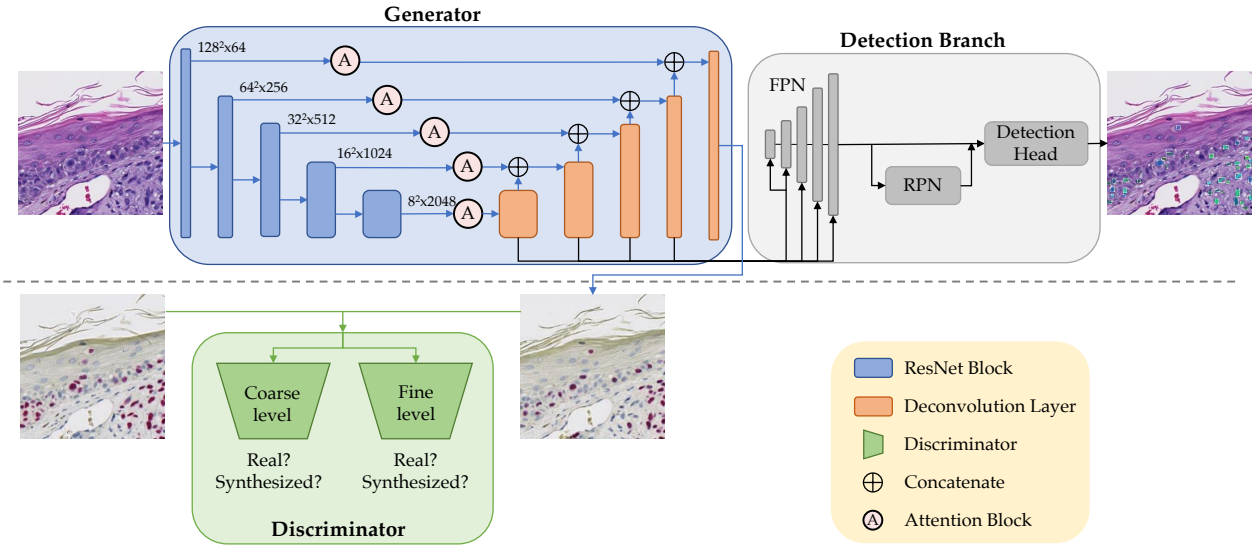


Figure 4.2: **VSGD-Net** framework: H&E images are virtually stained to Sox10. The jointly trained detection branch utilizes the intermediate features in the generator to detect melanocytes and provides feedback to the generator to enhance synthesis quality. The inference phase only uses the upper part of the architecture.

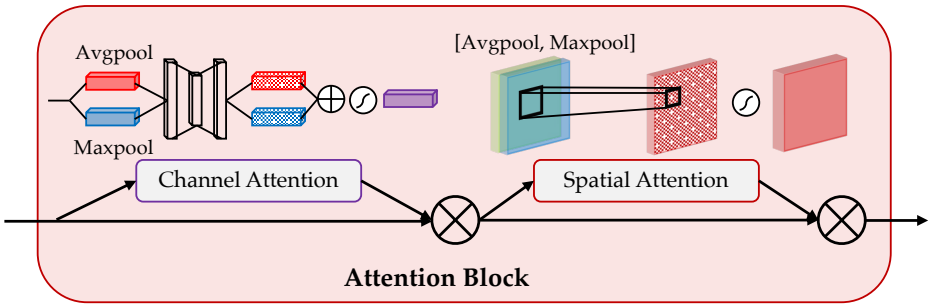


Figure 4.3: Attention block: Channel attention and spatial attention are consecutively computed to refine the features.

compared to the fine-level input. Similar to PatchGAN [52], each discriminator evaluates the realism of every fixed-sized patch in the image instead of directly evaluating the realism of the whole image. With the minimax loss introduced in [35], this multi-scale design guides G

to synthesize images with globally consistent patterns as well as finer details.

For detection, inspired by Mask R-CNN [41], we incorporate a feature pyramid network (FPN), a region proposal network (RPN), and the downstream heads. Learning to generate Sox10 images, the decoder layers correlate more closely with the Sox10 images than the encoder layers; moreover, Sox10 staining can highlight melanocytes in a red chromogenic color, which is consistent with the detection goal. In light of this, we place the detection branch within the decoder instead of the encoder, which is proven to be effective in the ablation study.

4.3.1 Training Process

In our end-to-end model, the virtually stained images and the detected instances are predicted from the shared intermediate features. To incorporate the feedback from both the image synthesis and the instance detection, we train G, D, and the detection branch jointly to learn from both the GAN loss L_{GAN} and the detection loss L_{DET}.

GAN Loss The generator G and the multi-scale discrimator D are optimized following the minimax loss [35]:

$$\min_G \max_D \sum_{i=1,2} \big(\log(D_i(X_s)) + \log(1 - D_i(G(X_h))) \big) \tag{4.1}$$

where D_1 and D_2 are the coarse- and fine-level discriminators, and X_s and X_h are the Sox10 and H&E images.

Besides the minimax loss, we add a feature similarity loss L_{feat} to improve the similarity between the generated and the real images. The calculation of L_{feat} involves multiple layers in D and a pretrained VGG19 model, and is given by the following equation:

$$L_{feat} = \sum_{i=1}^{N} ||D_i(X_s) - D_i(G(X_h))||_1 + \sum_{j=1}^{M} ||VGG_j(X_s) - VGG_j(G(X_h))||_1 \tag{4.2}$$

where N and M denote the layers to extract features. The feature similarity loss calculates the L1 term of the features of real and fake data given by the discriminator and the pretrained

VGG19. The features of the pretrained VGG19 model are the outputs of layers 1, 6, 11, 20, 29.

Detection Loss The detection loss L_{DET} is separated into L_{rpn}, L_{box_c}, L_{box_r}, and L_{seg}. L_{rpn} is the total loss of the candidate classification and the coarse bounding box regression in the RPN, given by the summation of binary cross entropy of the candidate classification and L1 loss on the coarse bounding box regression in the RPN. It forces the RPN to learn the location of anchor boxes and whether the anchor boxes contain objects. L_{box_c}, L_{box_r}, and L_{seg} are the losses for the instance classification, the final bounding box regression, and the segmentation in the downstream heads, which are given by the binary cross entropy of the instance classification, the binary cross entropy of the mask prediction, and the L1 loss of bounding box coordinates. The total loss is defined as:

$$L_{DET} = L_{rpn} + L_{box_c} + L_{box_r} + L_{seg} \tag{4.3}$$

Overall Losses and Training In **VSGD-Net**, the shared intermediate features are learned to characterize features of melanocytes and boost the Sox10 image synthesis at the same time. To facilitate such multi-task learning, we combine L_{GAN} with L_{DET} and backpropagate them to the encoder inside G. The final total loss is defined below:

$$\min_G \max_D \sum_{i=1,2} \big(\log(D_i(X_s)) + \log(1 - D_i(G(X_h))) \big) + \lambda * L_{feat} + L_{DET} \tag{4.4}$$

4.4 CC-WSI-Net

As shown in Figure 4.1, although VSGD-Net achieves state-of-the-art performance in patch-wise virtual staining task, the inconsistencies among patches remain a big challenge. To improve the stitching consistency and generate seamless WSIs, CC-WSI-Net builds color and content consistency modules on VSGD-Net to ensure both stain and clinical effectiveness (shown in Figure 4.4). For content consistency, we employ a context-aware discriminator to assess continuity between synthetic patches and their surrounding content. For color

consistency, we utilize a 2D histogram from an unrelated SOX10 WSI as color condition to supervise uniform color scheme across patches.

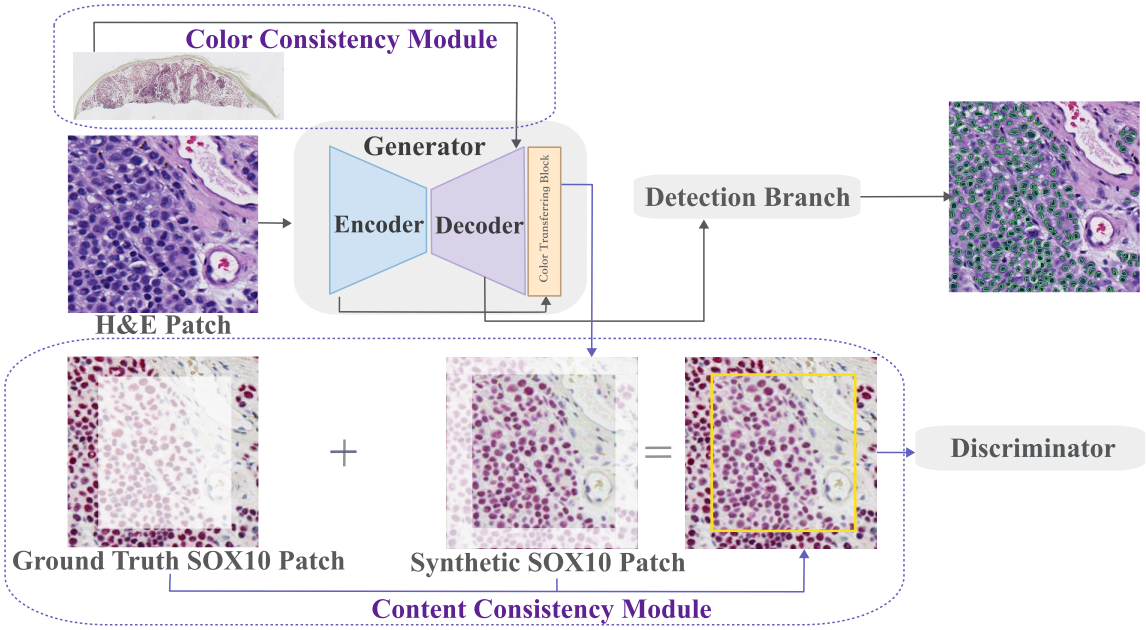


Figure 4.4: **CC-WSI-Net** framework: Content and color consistency modules are proposed to improve the image generation.

4.4.1 Content Consistency

To enforce content consistency over the borders of patches, we use a context-aware discriminator trained on composite images.As shown in Figure 4.4, we first synthesize a SOX10 patch with dimensions 256×256. The central region (192×192) is then cropped and combined with the border region from the ground truth image to create the final composite image, which is subsequently input to the discriminator. This approach forces the generator to produce center regions that seamlessly integrate with the broader surrounding tissue context. During training, the discriminator takes the composite image as the input and distinguishes based on the overall appearance. This design enforces both the visual coherence and the realistic

content generation.

4.4.2 Color Consistency

To address the color inconsistency problem, we propose a color consistency module for the generator. As shown in Figure 4.5, we replace the last convolution layer in the decoder with a color-transferring block. This module is inspired from the recolor head in ReHistoGAN [1]. To preserve the fine structure details in the synthesized images, the first two layers in the encoder are passed to the color-transferring block. In addition, the color-transferring block takes a 2-D color histogram from a 2.5x magnification WSI as a color condition to guide the color distribution of the generated image.

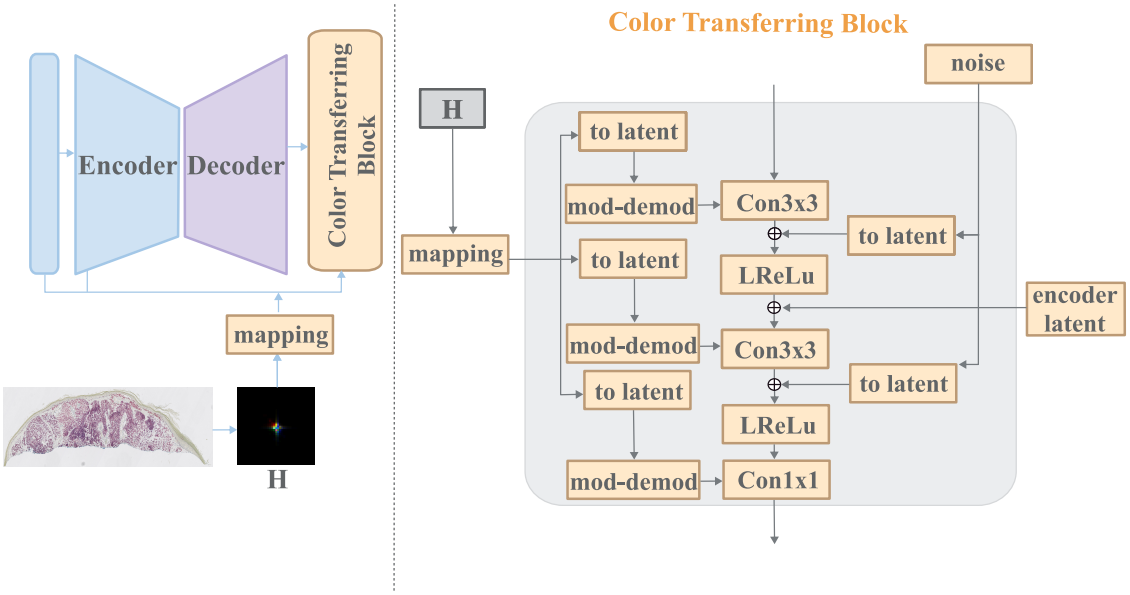


Figure 4.5: The left part illustrates the overall generator with the color consistency module, while the right part shows the details of the color transferring block in the color consistency module. *H* represents the 2-D color histogram of the reference WSI as the color condition, and *mapping* refers to a fully connected network.

4.4.3 Training Process

Analogous to the approach employed in ReHistoGAN, we utilize the color-matching histogram loss to supervise the color scheme of the synthetic Sox10 image. The histogram loss is given as:

$$C(\mathbf{H}_g, \mathbf{H}_s) = \frac{1}{2} \left\| \mathbf{H}_g^{1/2} - \mathbf{H}_s^{1/2} \right\|_2, \tag{4.5}$$

where $\|\cdot\|_2$ is the standard Euclidean norm and $H^{1/2}$ is an element-wise square root, and $\mathbf{H}_g$ and $\mathbf{H}_s$ are the histograms of the ground truth Sox10 patch and synthetic Sox10 patch respectively.

To train CC-WSI-Net, we adopt the loss functions in VSGD-Net [68] (see Equation 4.4)and combine with the histogram loss, the overall loss function for CC-WSI-Net is the following:

$$\begin{aligned} L_{total} = & \min_G \max_D \sum_{i=1}^{2} \left(\log(D_i(X_s)) + \log(1 - D_i(G(X_h))) \right) \\ & + \lambda * L_{\text{feat}} + L_{\text{DET}} + \lambda_{color} * C(\mathbf{H}_g, \mathbf{H}_s) \end{aligned} \tag{4.6}$$

where λ_{color} is the weight for color histogram loss.

4.5 Results and Ablation Study

4.5.1 Melanocyte Detection

Experimental Design and Baseline Methods We compared **VSGD-Net** with two lines of methods. The first group is specialized in nuclei detection, including Radial Line Scanning (RLS)[76], Mask R-CNN[41], U-Net[89], StarDist[92], HoverNet[36], the new state-of-the-art CHR-Net[30], and a "nuclei classification" method we designed. RLS was specifically proposed to study melanocyte detection. It leverages a feature-based approach based on the "halo region" assumption that melanocytes appear with a brighter region surrounding the nuclei under H&E staining. Furthermore, to investigate the local texture around nuclei, we designed the "nuclei classification" method, which first applies a fine-tuned ensemble model [93] to detect nuclei and then trains the open-source ESPNetv2[77] to classify cropped nuclei patches.

The second group of methods consists of GAN-based approaches, including StainGAN [94], PC-StainGAN [71], and a self-implemented GAN-based segmentation model similar to [34]. The segmentation model, whose G and D are the same as **VSGD-Net**, directly feeds the synthesized image to the segmentation net and is trained end-to-end. For the other GAN models that do not incorporate any downstream modules, we tested their performances in a two-stage manner, using the random forest and the NuSeT model in our groundtruth-generating step (Chapter 2.3).

In our experiments, the ResNet-50 backbone in Mask R-CNN and the ResNet-34 backbone in CHR-Net are pre-trained with ImageNet for fair comparisons. We empirically set $\lambda = 10$ in Eq. 4.4. We report precision (P), recall (R), F_1-score, and Jaccard index. All the models are trained and tested on the dataset described in Chapter 2.3.

Results In clinical practice, diagnosing melanoma relies heavily on accurately identifying melanocyte distribution, requiring both high precision and recall. Low recall risks missing malignant melanocytes, leading to under-diagnosis, while low precision may result in over-diagnosis. Consequently, metrics like the F_1-score and Jaccard index, which balance precision and recall, are crucial for evaluation.

As shown in Table 4.1, **VSGD-Net** achieves the highest F_1-score and Jaccard index. While RLS heuristically exploits the "halo region" characteristic of melanocytes, it requires extensive hyperparameter tuning and lacks generalizability. Methods like "Nuclei Classification" and Mask R-CNN exhibit high precision but low recall, as their instance-level learning framework predicts only high-confidence instances. Similarly, StarDist and HoverNet struggle with the shape representation and distance maps due to the similarity between melanocytes and other cells. U-Net performs reasonably well, aided by skip connections, while CHR-Net, with its dual U-Net structure and high-resolution feature extraction, improves on U-Net by 1%, consistent with prior findings [30]. However, both models underperform **VSGD-Net** due to their inability to leverage Sox10 staining.

Figure 4.6 illustrates qualitative comparisons among **VSGD-Net**, CHR-Net, and GAN-

Table 4.1: Comparison with nuclei detection methods.

Method	**P**	**R**	F_1	**Jaccard**
RLS [76]	0.443	0.570	0.499	0.332
Nuclei Classification	0.693	0.506	0.585	0.413
Mask R-CNN [41]	0.735	0.514	0.605	0.434
U-Net [89]	0.630	0.639	0.635	0.465
StarDist[92]	**0.745**	0.426	0.542	0.372
HoverNet[36]	0.729	0.499	0.592	0.421
CHR-Net [30]	0.607	0.688	0.645	0.476
Ours	0.660	**0.710**	**0.684**	**0.520**

based segmentation. **VSGD-Net** predictions align closely with the ground truth, while CHR-Net over-predicts melanocytes in the bottom-left region, and GAN-based segmentation over-predicts those at the top.

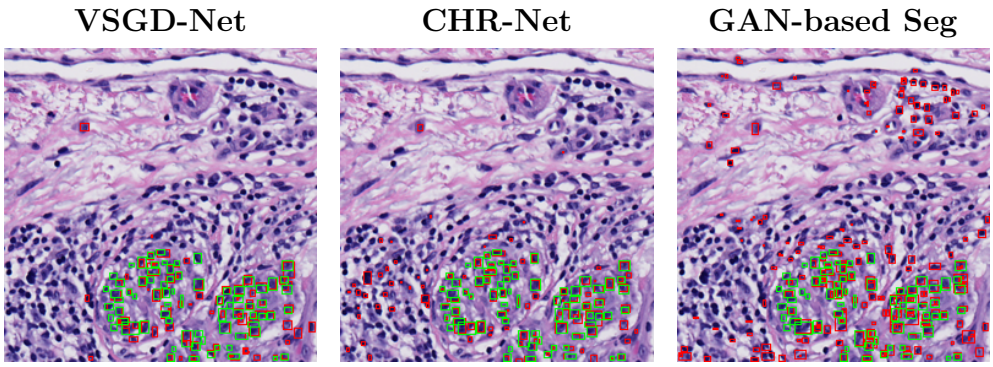


Figure 4.6: The green and red bounding boxes denote the groundtruth and the predicted instances. (Zoom in for best view)

Table 4.2 and Figure 4.7 demonstrate the performance of GAN-based methods. StainGAN [94] and PC-StainGAN [71] were designed based on unsupervised CycleGAN [125]. Without any additional supervision, StainGAN fails to learn the distribution gap between the two stainings. Although PC-StainGAN adds a pathology constraint to the CycleGAN, it still lacks supervision on the conversion between H&E and Sox10. On the other hand, the GAN-based segmentation method has supervision on the synthesized images, but its detection performance is bounded by the image synthesis quality due to its architecture.

Table 4.2: Comparison with GAN-based methods.

Method	**P**	**R**	F_1	**Jaccard**
StainGAN [94]	0.476	0.299	0.367	0.225
PC-StainGAN [71]	0.591	0.343	0.434	0.277
GAN-based Segmentation	0.569	**0.719**	0.636	0.466
Ours	**0.660**	0.710	**0.684**	**0.520**

Ablation Study In Table 4.3, we ablated each key component in **VSGD-Net**, namely the image synthesis features, the location of the detection branch and the attention module's presence. To verify the efficacy of the image synthesis features, we replaced the generator of **VSGD-Net** with the generator in Pix2PixHD[105], which has fewer convolution layers, no skip connections, and no attention module. As Row 1 of Table 4.3 shows, despite the weakness of the Pix2PixHD generator, it still achieves comparable results and outperforms other baselines with the key component of boosting detection with image synthesis features. We assumed the features in the decoders have higher correlations with Sox10 staining and melanocytes, and the attention module refines the intermediate features. Such assumptions are verified by the notable performance gains in Table 4.3 row 5.

Table 4.3: Ablation results.

Generator	**Features From**	**Atten.**	F_1	**Jaccard**
Pix2pixHD	Decoder	-	0.654	0.486
Ours	Encoder	✗	0.641	0.472
Ours	Decoder	✗	0.674	0.508
Ours	Encoder	✓	0.660	0.492
Ours	Decoder	✓	**0.684**	**0.520**

4.5.2 Virtual Staining

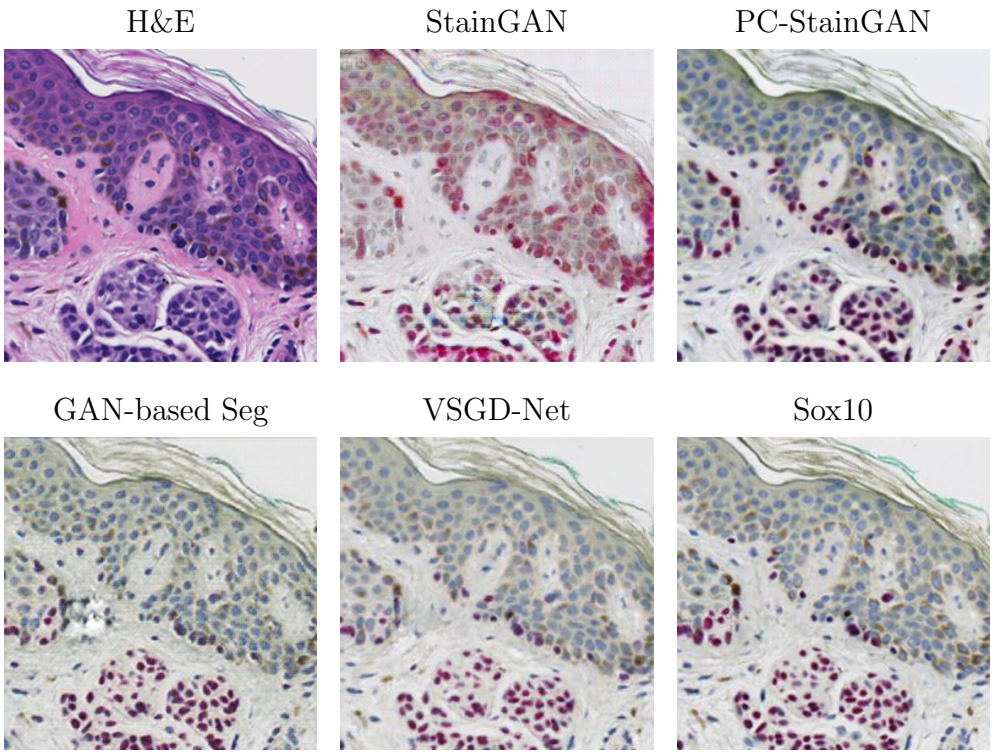


Figure 4.7: Synthesized Sox10 images.

To measure the reliability of the virtual staining, we calculate the average Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM). Larger numbers in PSNR and SSIM indicate better image quality and higher similarity with the groundtruth. As Table 4.4 shows, **VSGD-Net** achieves the highest PSNR and a comparable SSIM to PC-StainGAN. By assessing the mean squared error of the synthesized images, higher PSNR indicates more reliable results with regard to the virtual staining task. We present more virtual staining and detection results in Figure 4.8.

Table 4.4: Synthesized image quality assessment.

Method	**PSNR**(dB)	**SSIM**
StainGAN [94]	19.010	0.577
PC-StainGAN [71]	19.344	**0.618**
GAN-based Segmentation	19.583	0.569
Ours	**19.815**	0.611

4.5.3 Generating Seamless WSIs

Experimental Design In our study, we utilize the melanocyte dataset (detailed in Chapter 2.3) for training and quantitative evaluation of **CC-WSI-Net**. As no established metric exists to measure patch consistency, we perform a subjective assessment of synthesis quality. For this evaluation, we compile the *Consistency Eval* dataset, comprising 25 new cases. Each case includes an H&E-stained WSI and its corresponding Sox10 image (unregistered). Similar to the melanocyte dataset, a WSI contains 6-12 slices from the same patient. For each case, an expert pathologist (Stevan Knezevich) selects one diagnostically representative slice for the subjective survey.

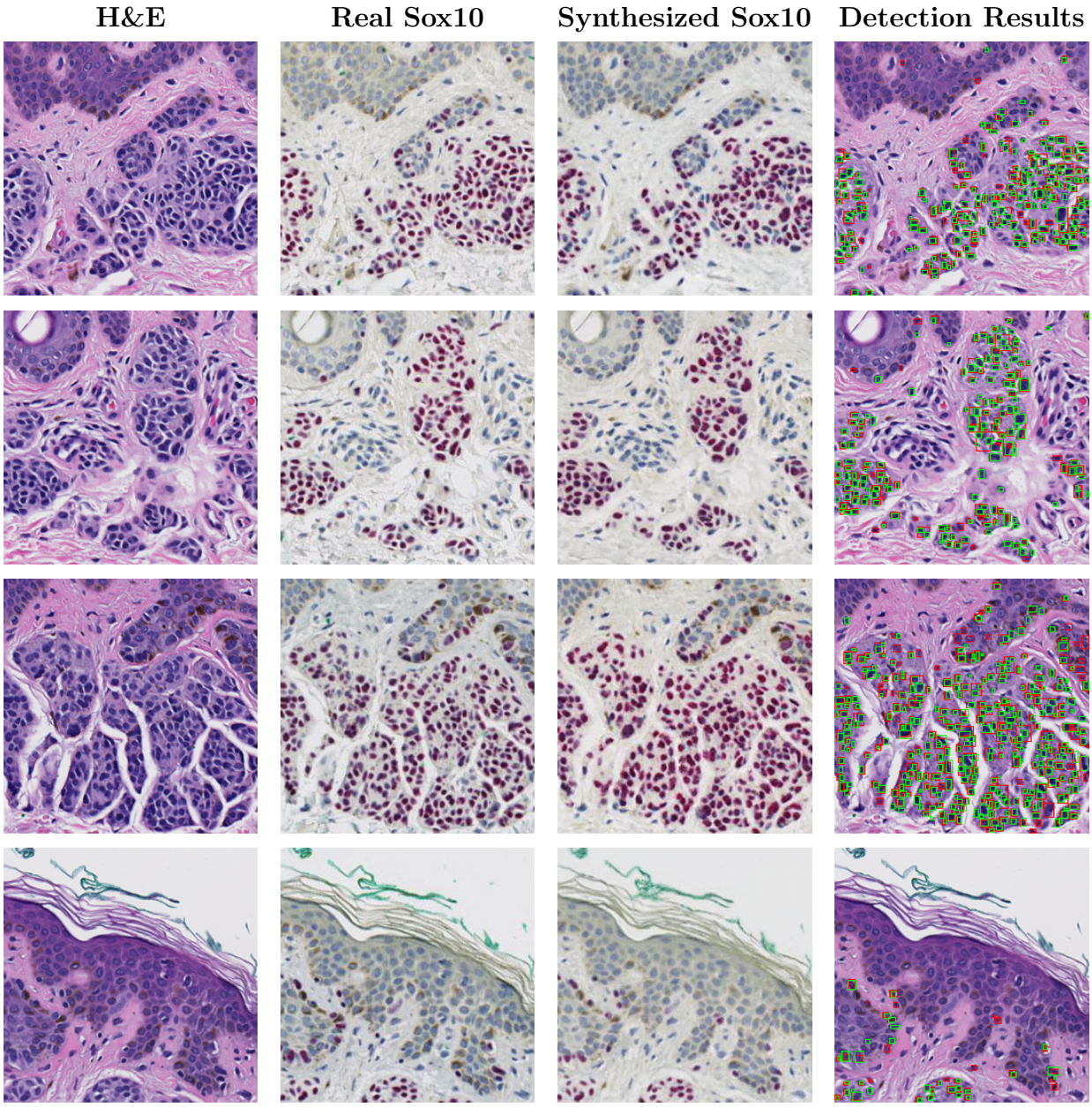


Figure 4.8: More qualitative results from **VSGD-Net**. In the last column, the green bounding boxes denote the groundtruth melanocytes while the red bounding boxes denote the predicted melanocytes. (Zoom in for best view)

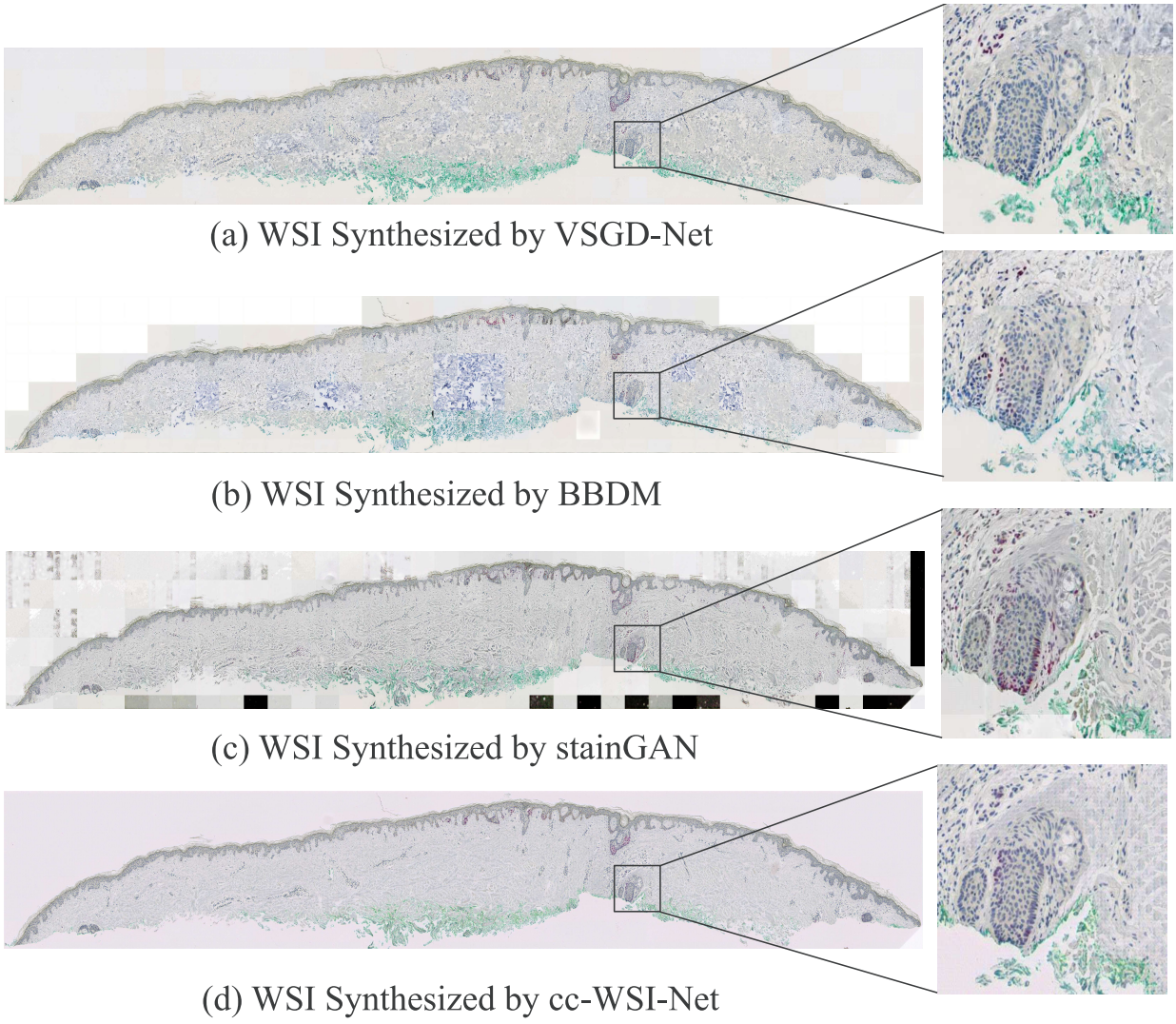


Figure 4.9: **Qualitative comparisons on synthesized WSI:** CC-WSI-Net produces seamless WSIs without the content and color inconsistencies seen in VSGD-Net, BBDM, and stainGAN. (Zoom in for best view)

Qualitative Comparison Figure 4.9 presents a qualitative comparison of WSIs synthesized by different methods, including VSGD-Net, Brownian Bridge Diffusion Model (BBDM) [64], and stainGAN [95]. Unlike CC-WSI-Net, the baseline methods exhibit noticeable issues such as color mismatches and stitching artifacts. These results highlight the effectiveness of

CC-WSI-Net's consistency modules for high-quality synthesis.

4.5.4 Subjective Survey among Pathologists

Survey Design To assess the feasibility and effectiveness of synthetic Sox10 WSIs generated from **CC-WSI-Net**, we design a subjective survey using the *Consistency Eval* dataset. This study aims to strike a balance between the objective measurement and the subjective clinical relevance crucial for real-world applications. In this study, three board-certified pathologists independently reviewed the digital slides. Each pathologist reviewed a total of 50 digital slides: 25 with traditional Sox10 staining and 25 with synthetic Sox10 staining. Each case included both H&E and Sox10 stain images but omitted clinical history to ensure unbiased evaluations.

To further reduce bias, the cases were presented in two randomized blocks. In the first block, the 25 cases were displayed in a random order, with each case assigned a specific staining method (either synthetic or traditional) through simple randomization. In the second block, the same 25 cases were presented again, but in a different random order, and each case was assigned the alternate staining method from the first block. This ensured that each pathologist reviewed every case twice, once with each staining method. Additionally, there were no demarcations to the pathologists that differentiated the two blocks during the review process. The sequence of the 50 slides was maintained consistently across all three pathologists to ensure uniformity in the evaluation.

For each case, the pathologists were provided with an evaluation survey that included the following criteria:

1. **Effectiveness of Sox10 Staining:** Rate how well the Sox10 staining made melanocytes more clearly visible and distinct from the surrounding tissue (1 = poor to 4 = perfect).

2. **Image Quality:** Rate the overall quality of the whole slide image (1 = poor to 4 = perfect).

3. **Staining Identification:** Indicate whether they believed the slide was synthesized, immuno-stained, or if they cannot tell.

Table 4.5: Effectiveness and quality of the synthetic WSIs compared to traditional WSIs.

Review Characteristics	**Traditional Sox10 N (%)**	**Synthetic Sox10 N (%)**
Effectiveness of Sox10 Staining		
1 (poor)	13 (17%)	6 (8%)
2	11 (15%)	14 (19%)
3	29 (39%)	32 (43%)
4 (perfect)	22 (29%)	23 (31%)
Image Quality		
1 (poor)	2 (3%)	0 (0%)
2	1 (1%)	0 (0%)
3	21 (28%)	5 (7%)
4 (perfect)	51 (68%)	70 (93%)

Evaluation Results and Discussion The effectiveness and quality ratings for both traditional ground truth and synthesized Sox10 staining are shown in Table 4.5. The synthetic Sox10 staining received higher mean ratings (standard deviation, sd) for both effectiveness and quality compared to traditional Sox10 staining. The mean (sd) rating for effectiveness for the synthetic images was 3.0 (0.9) and for traditional images was 2.8 (1.1). The mean (sd) rating for image quality for synthetic images was 3.9 (0.3) and for traditional images was 3.6 (0.7). The distribution of these ratings for effectiveness is further shown visually in Figure 4.10.

The results of staining identification are shown in Table 4.6. Pathologists reported that they could not distinguish the Sox10 staining method for 101 (67%) of the images; when they

Table 4.6: Subjective survey results on whether pathologists can identify the staining method.

Accuracy of pathologists in identifying staining method	**N (%)**
Incorrectly identified staining method	37 (25%)
Identified traditional when synthetic	19 (13%)
Identified synthetic when traditional	18 (12%)
Correctly identified staining method	12 (8%)
Correctly identified synthetic	8 (5%)
Correctly identified traditional	4 (3%)
Cannot tell	101 (67%)

placed a response attempting to guess if the image was traditional versus synthetic they were more likely to guess incorrectly.

Overall, these findings suggest that the synthetic Sox10 WSIs generated by our method are indistinguishable from the true, traditional Sox10 WSIs. These evaluation results also indicate that CC-WSI-Net can produce high-quality virtual staining that maintains diagnostic relevance and effectiveness, comparable to traditional IHC methods.

Although these results are promising, the study has some limitations. The study was conducted in a controlled test environment with only three pathologists rather than in a clinical setting. While it enhances the reliability and repeatability of our tests, it does not fully replicate the complexities and unknown variables typical of clinical practice. Further studies are needed before clinical implementation. Additionally, technical elements such as slide preparation, potential variability in digitization techniques, and image processing methods need to be considered. For the traditional Sox10 stains, tissues were first stained with H&E, the coverslip was removed, and then restained with Sox10. While this creates perfectly matched image pairs, it may introduce some imperfections in the Sox10 images. Despite standardized procedures, variability in tissue samples and staining remains possible.

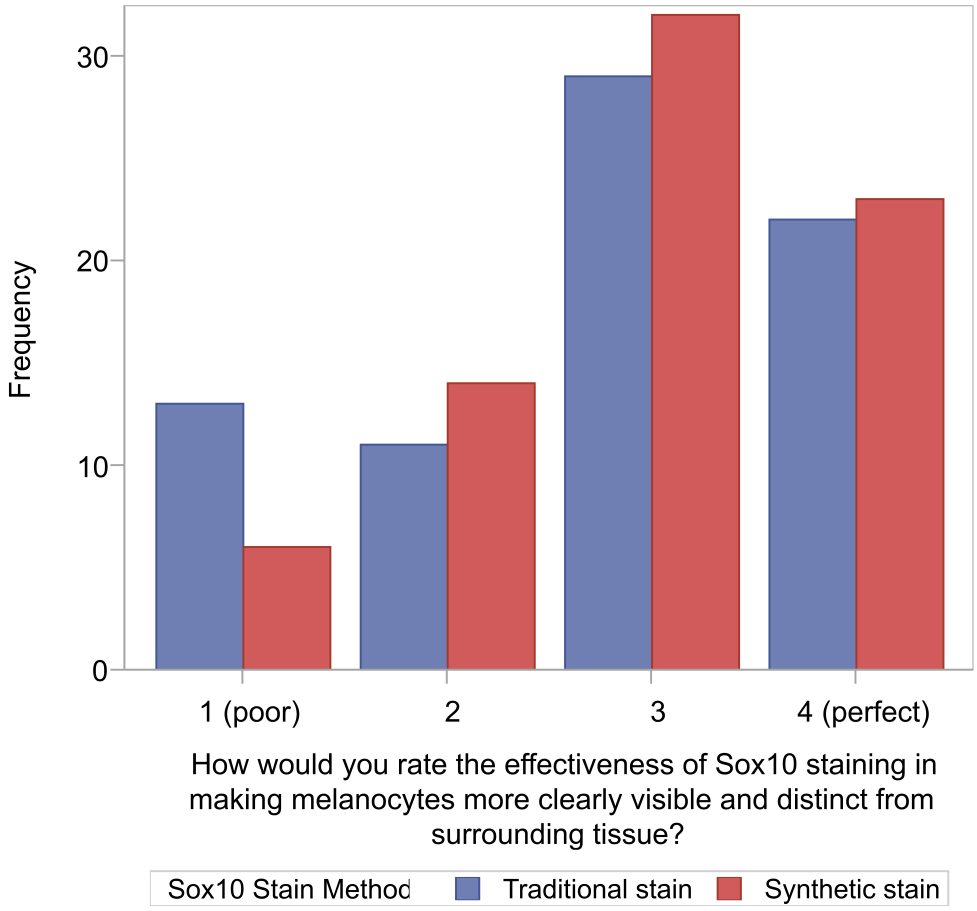


Figure 4.10: Distribution of pathologist ratings of effectiveness of Sox10 staining for traditional and synthetic Sox10 images (N=150 reviews)

Future work should include a broader range of samples and settings to ensure the robustness and generalizability of our approach. Despite these limitations, this study has established a firm foundation for subsequent clinical validation and real-world application, promising to enhance diagnostic processes significantly.

4.6 Summary

In this chapter, we presented two innovative frameworks addressing key challenges in virtual staining and melanocyte detection within melanoma analysis. First, we introduced a novel virtual staining guided detection network, **VSGD-Net**, and investigated cell-type-of-interest detection with the boost of image synthesis features between two distinct stainings on the skin biopsy specimen. During inference, **VSGD-Net** achieves accurate results using only routine H&E staining, validated through extensive experiments on the melanocyte dataset. This method holds promise for broad applicability across diverse tissue types and diseases.

Second, we proposed **CC-WSI-Net**, a framework for synthesizing virtually stained Whole Slide Images (WSIs) from H&E slides, with a focus on addressing color and content consistency issues between adjacent patches. By ensuring IHC stain accuracy, CC-WSI-Net facilitates the generation of high-quality WSIs, offering potential integration into Computer-Aided Diagnosis systems for enhanced WSI analysis. Together, these contributions advance the field of virtual staining and histopathological image analysis, paving the way for more accurate and efficient diagnostic tools.

Chapter 5

SEMANTICS-AWARE ATTENTION GUIDANCE

5.1 Introduction

Building on the focus of Chapter 4, which introduced methods to detect diagnostically significant cell entities for melanoma analysis, Chapter 5 shifts the focus to melanoma diagnosis itself. While existing WSI classification methods have achieved success, their lack of semantic information often results in a limited and less reliable understanding of the underlying histopathological features. Incorporating semantic information, such as the diagnostically relevant tissue structures, can not only improve classification accuracy but also enhance model interpretability, making the results more aligned with clinical reasoning. In this chapter, we explore current classification models and propose a novel semantics-aware attention guidance module to address key limitations in existing approaches.

Automated computerized diagnosis systems for histopathology images requires assessment on WSIs. While deep learning offers a common solution for computer vision tasks, WSIs, as gigapixel images, pose a distinct challenge for deep learning due to their size, making them impractical to directly input into a neural network. The Multiple Instance Learning (MIL) framework provides a solution to this challenge. Within this framework, the input WSI is cropped into patches (instances), enabling the neural network to process the inputs effectively, and the image-level classification result is determined based on the aggregated information from all instances, either through prediction based on the patches' embeddings [40, 116] or the aggregation of patches' prediction labels [16].

MIL models classify images in a way very different from how pathologists approach diagnosis. Pathologists typically begin by scanning a slide at low magnification to identify suspicious regions and form hypotheses, then zoom in to high magnification to analyze cellular

structures, mitotic figures, tissue architectures, *etc.*, , ultimately reaching a definitive diagnosis [78]. MIL models, however, treat patches independently, overlooking the multi-scale and hierarchical nature of pathology. This limitation prevents them from capturing the long-range interactions and nuanced details critical for accurate diagnosis, as they fail to emulate the multi-scale focus pathologists rely on.

In light of these discrepancies, vision transformer models are leveraged to capture the dependencies among patches and generate holistic and effective representations [84, 13, 12, 121, 108]. Specifically, **ScAtNet** [108] introduced a transformer-based end-to-end network that adapts to the information from different input scales using self-attention and predicts the classification label. Results show that **ScAtNet** outperforms other MIL methods by a large margin in the task of melanoma diagnosis.

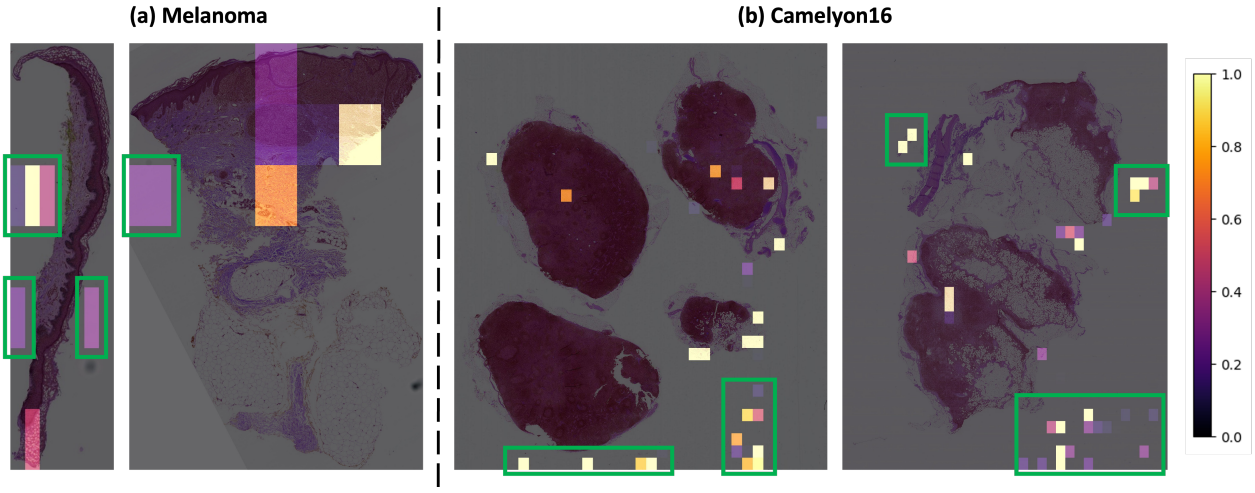


Figure 5.1: Visualization of the baseline model's (ScAtNet [108]) attention on (a) skin biopsy WSIs in the melanoma dataset and (b) breast biopsy WSIs in the Camelyon16 dataset. Green boxes show examples of the baseline model mistakenly focusing on background regions. The signal and attention values are normalized for visualization purposes.

Although it has become a common belief that transformer models, which explore the dependencies among the patches, are suitable for the patch-based WSI analysis task, we

found that sometimes the attention from the transformer models makes no sense by falling on non-cancerous regions, let alone the blank background in the image (see the left images in Figure 5.1 (a) and (b)). While it is unclear why the model bases its prediction on those regions, the attended regions contradict the clinical diagnosis process, and at the same time diminish the model's interpretability and reliability. Previous knowledge, including tissue segmentation, melanocytic proliferation segmentation [69] and melanocyte detection [68], could be potentially incorporated into the diagnosis pipeline to improve interpretability and classification performance. In light of this, we introduce a Semantics-Attention-Guiding transformer network, **SAG**, whose key contributions are:

- Attention Guiding Module: A novel module compatible with attention-based MIL or transformer models, enabling enhanced interpretability and performance.

- Flexible Attention-Guiding Loss: Designed to integrate diverse semantic information, such as tissue and cancerous region masks, into the training process.

- Heuristic Attention Generation: An approach to derive heuristic-guidance signals from diagnostically relevant entities.

- Improved Diagnosis Accuracy: Demonstrated improvements over state-of-the-art models on datasets spanning two cancer types.

5.2 Related Work

In recent years, deep learning technologies have revolutionized histopathological image analysis, offering unprecedented opportunities for automation, precision and scalability in diagnosis [23, 15]. However, analyzing gigapixel WSIs poses unique challenges due to their size and complexity. To address this, MIL models have become a widely adopted solution, offering computational efficiency by segmenting WSIs into smaller, manageable patches [47, 79, 63]. Though MIL is effective, traditional MIL approaches fail to capture the nuanced

and hierarchical diagnostic process used by pathologists, as they often treat all patches equally, overlooking the spatial and contextual relationships between them.

To mitigate this limitation, attention mechanisms have been integrated into MIL frameworks, allowing models to prioritize diagnostically relevant regions. For instance, ABMIL [50] employs an additional MLP layer to learn attention weights for each patch, effectively identifying areas of interest. Building upon this, Additive MIL [54] introduces an additive attribution module, further refining patch importance and enhancing classification performance in pathology. These attention mechanisms enable an effective aggregation of features from patches and improves the diagnosis accuracy.

Beyond MIL, vision transformer models have emerged as a powerful alternative for WSI analysis. Thanks to their self-attention mechanisms, transformers naturally capture long-range dependencies between patches, constructing comprehensive global representations that incorporate both spatial and contextual relationships [84, 13, 12, 121, 108]. Recent studies have further explored multi-resolution approaches, aggregating features hierarchically or through concatenation to enhance diagnostic predictions [108, 38, 97]. For instance, ScAtNet [108], a transformer-based end-to-end network, integrates information across different scales using self-attention mechanisms, and outperforms other MIL methods in melanoma diagnosis. This method emulates how pathologists diagnose by taking WSIs with multiple resolutions as input, and utilizes the attention mechanism to learn attended regions. However, in such methods and the aforementioned attention based MIL models, the learned regions can be very different from diagnostically relevant regions in clinical practice, making the model less reliable and interpretable.

In response, integrating domain-specific knowledge into diagnostic models has emerged as a promising strategy. Such efforts not only enhance classification accuracy but also improve model performance, especially in scenarios where data is scarce. For example, Miao *et al.*[82] introduce spatial prior attention using binary anatomy knowledge maps, demonstrating the potential of integrating prior knowledge. However, their reliance on binary representations highlights the need for richer and more nuanced semantic information to improve accuracy.

Similarly, Chen *et al.* [13] incorporated genomics data alongside WSIs to predict patient outcomes, but their framework lacks flexibility for integrating other modalities or guidance signals.

To address these gaps, we propose a Semantics-Aware Attention Guidance (SAG) module, which enhances both MIL and transformer-based models by leveraging semantic information to supervise attention learning. The proposed module is adaptable to both binary and continuous signals, allowing it to guide model attention toward diagnostically important regions. By embedding domain knowledge into the attention mechanism, SAG significantly improves classification accuracy, model reliability, and interpretability, offering a robust framework for clinically informed WSI analysis.

5.3 Methodology

Our **SAG** framework aims to infuse diagnostic models with relevant knowledge, thereby enhancing the diagnostic performance and the interpretability of attention-supervised representations. This versatile framework is compatible with a broad range of attention-based MIL and transformer methods. Figure 5.2 illustrates our **SAG** pipeline, which includes three main components: 1) generate patchwise embeddings with an off-the-shelf feature extractor, 2) learn diagnostic patterns from these embeddings via a diagnosis network, and 3) utilize an attention-guiding loss that leverages heuristic guidance (**HG**) and tissue guidance (**TG**). In the following sections, we give the details of the proposed attention guidance.

5.3.1 Diagnosis Models

As Figure 5.2 illustrates, after cropping the WSI into patches, we employ a pre-trained feature extractor f to extract patch embeddings. The implementation detail of f is provided in Chapter 5.4.3. To demonstrate the versatility and model-agnostic nature of our **SAG** framework, we apply **SAG** to two state-of-the-art baseline models: a transformer-based model, ScAtNet [108], and an MIL-based model, ABMIL [50].

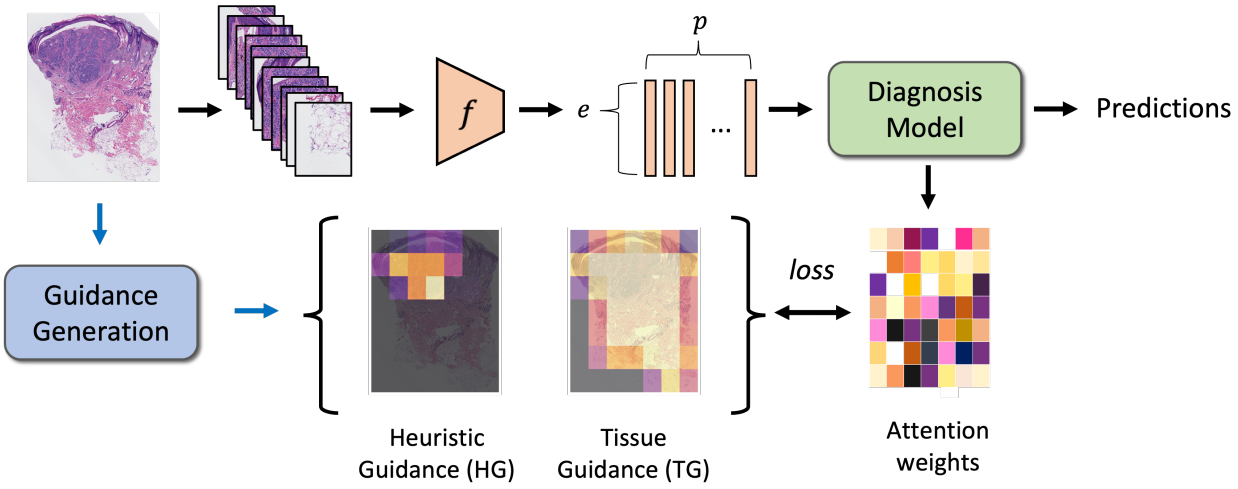


Figure 5.2: Overview of the **SAG** approach for improving WSIs diagnosis models. First, a high-resolution histopathological image is divided into p number of non-overlapping patches. Then, patch embeddings are obtained using an off-the-shelf feature extractor f. Subsequently, a diagnostic network utilizes the $p \times e$-dimensional feature map for classification into distinct categories. During training, heuristic guidance (**HG**) and tissue guidance (**TG**) are leveraged to supervise the attention within the diagnosis model, ensuring the focus on diagnostically relevant regions.

5.3.2 Attention Weights

For transformer-based models, the self-attention module learns the attention weights for the input patches. Typically, a transformer model consists of l layers with h self-attention heads per layer. For each self-attention head, the inputs $\mathbf{X} \in \mathbb{R}^{p \times e}$, where p is the number of patches and e is the dimension of the features, are transformed into query ($\mathbf{q}$), key ($\mathbf{k}$), and value ($\mathbf{v}$) vectors, where $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^{p \times e_k}$. The attention head induces a pairwise similarity ($\mathbf{A}$) from $\mathbf{q}$ and $\mathbf{k}$ to transform $\mathbf{v}$, as shown in Equation 5.1 and 5.2.

$$\mathbf{A} = \text{softmax}(\frac{qk^{\intercal}}{\sqrt{D_h}}) \in \mathbb{R}^{p \times p} \tag{5.1}$$

$$\textbf{Self-attention} = \mathbf{A} \cdot \mathbf{v} = \text{softmax}(\frac{qk^{\intercal}}{\sqrt{D_h}}) \cdot \mathbf{v} \tag{5.2}$$

To obtain the model's attention weights for each patch ($\mathbf{MA_t}$), we average the pairwise similarity as illustrated in Equation 5.3.

$$\mathbf{MA_t} = \frac{1}{p}\sum_{i=1}^{p} A_i \in \mathbb{R}^p \tag{5.3}$$

For MIL models [50], the attention weights ($\mathbf{MA_m}$) are formulated as the weighted aggregation of instance embeddings:

$$\mathbf{MA_m} = f(x) \in \mathbb{R}^p, \tag{5.4}$$

where f denotes the linear layers to learn the attention weights, and $x \in \mathbb{R}^{p \times e}$ denotes the embeddings from p patches.

5.3.3 Attention Guidance Formulation

To accommodate the learning of diagnostically relevant regions and regularize the model's attention **MA**, we design two types of semantic attention guidance: tissue guidance (**TG**) and heuristic guidance (**HG**) (Figure 5.3), each represented as a vector $\in \mathbb{R}^p$. We describe the formulation of attention guidance in two steps: 1) Acquisition of tissue mask and diagnostic heuristics, and 2) Calculation of guidance weights.

Acquisition of semantic masks To obtain the tissue mask for **TG**, Otsu's method [118] is used to perform high-quality segmentation of tissue patches. This process transforms the input image shown in Figure 5.3a into the binary tissue mask shown in Figure 5.3b.

To obtain **HG**, we exploit dataset- and disease-specific prior knowledge, such as structures, tissues, and cells. In the example shown in Figure 5.3, we first perform cell segmentation for a specific cell type (Figure 5.3d). Then, groups of cells are aggregated via the density-based spatial clustering algorithm DBSCAN [27]. Next, the convex hull [102] is generated for each cluster (Figure 5.3e) and utilized as the semantic signal for attention supervision (Figure 5.3f).

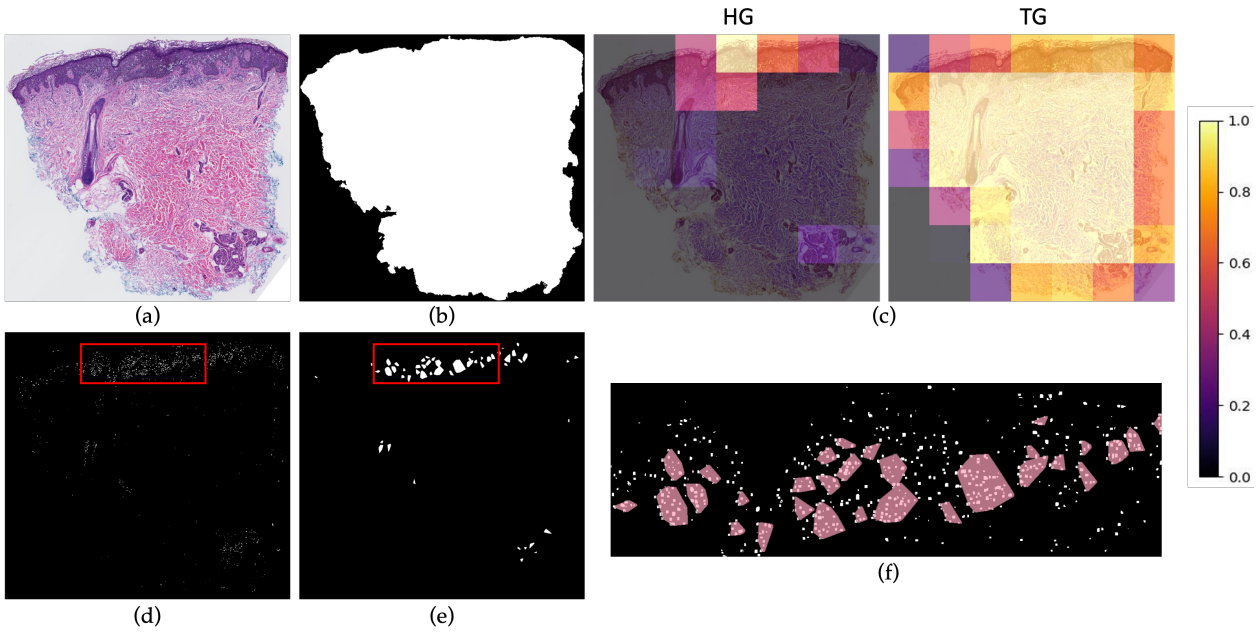


Figure 5.3: Generation of attention guidance: (a) H&E sample image. (b) Tissue segmentation mask. (c) **HG** and **TG**. The values are normalized for visualization purpose. (d) Cellular entities detected (**zoom-in for best view**). (e) Convex hull of cellular clusters. (f) A zoomed-in view of the red boxes in (d) and (e). The convex hull is rendered with red color.

Calculation of guidance weights To calculate the guidance weight $W \in \mathbb{R}^p$ for **HG** and **TG**, we normalize the semantic masks to transform the heuristic signals (**HG**) and the tissue masks (**TG**) into the attention supervision (Figure 5.3c):

$$W_i^k = \frac{M_i^k}{\sum_{j=1}^{p} M_j^k}, \quad k \in \{\mathbf{TG}, \mathbf{HG}\}, \tag{5.5}$$

where W_i^k denotes the guidance weight of patch i, and M_i^k is the mask area ratio of patch i.

5.3.4 Loss Functions

Since heuristic guidance (**HG**) reflects the relevance to the diagnosis, we employ the mean squared error (MSE) loss, L_{mse}, to regularize the model's attention **MA**:

$$L_{mse} = \frac{1}{p}\sum_{i=1}^{p}(W_i^{\mathbf{HG}} - \mathbf{MA}_i)^2. \tag{5.6}$$

On the other hand, tissue guidance (**TG**) is useful in guiding the model to focus on tissue patches and ignore the background and artifact patches. Thus, we employ an inclusion-exclusion loss, $L_{in\&out}$, which sums the attention weights outside of the tissue and the negative attention weights inside the tissue, as defined in Equation 5.7 below. The optimization using this inclusion-exclusion loss penalizes the model's attention weights on background and artifact regions while encouraging focus on tissue areas, thereby preventing classification based on irrelevant patches.

$$L_{in\&out} = \frac{1}{p}(-\sum_{i,W_i^{\mathbf{TG}}>0}^{p}\mathbf{MA}_i + \sum_{i,W_i^{\mathbf{TG}}=0}^{p}\mathbf{MA}_i). \tag{5.7}$$

To integrate the training losses effectively, we adopt uncertainty weighting, $\mathcal{UW}$ [59], which weighs multiple loss functions by considering the homoscedastic uncertainty - task-specific noise that remains constant across different input samples - of each task. The overall loss function is defined as:

$$L = \mathcal{UW} \otimes \{L_{cls}, L_{mse}, L_{in\&out}\}, \tag{5.8}$$

where L_{cls} is the cross entropy loss for the classification task.

5.4 Dataset and Implementation Details

To demonstrate **SAG**'s effectiveness across various cancer types and datasets, we train the model on the M-PATH dataset and the Camelyon16 dataset [8].

5.4.1 M-PATH dataset

As described in Chapter 2.1, the M-PATH dataset comprises skin biopsy images with diagnostic classes spanning from benign to invasive pT1b.

Feature Extraction We leverage an ImageNet pre-trained MobileNetV2 [91] to extract a 1280-dimensional feature vector for each patch within the WSI.

Heuristic Guidance As melanocytes are believed to be highly informative about melanoma diagnosis, we leverage the VSGD-Net [68] (Chapter 4) to generate the cellular entity map that eventually transforms to **HG**, as described in Section 5.3.3. To cluster the cell entities, DBSCAN in the scikit-learn package [9] is used with `eps=20` and `min_samples=5`. **TG** is generated using Otsu thresholding [118].

Soft Labels A WSI often contains multiple tissue slices. However, the diagnosis is often represented by only one or two tissue slice, and the other tissue slices may correspond to other diagnosis categories, which are less severe than the case diagnosis. In this way, assigning the same diagnostic label to all tissue slices could cause more false tissue-label pairs and hinders learning representations. To address this, we adopt the soft labeling method proposed in **ScAtNet** [108]. It applies singular value decomposition on each tissue slice, then acquires similarity scores for each diagnostic class via dot product, and assigns soft labels with floating point numbers for the tissue slices that are not marked with ROIs.

5.4.2 Camelyon16 dataset

Camelyon16 [8] is a public dataset comprising 400 H&E stained WSIs from breast cancer. The WSIs are diagnosed into two classes: normal and tumor. We use the official split of 271/129 slides for training and testing. To train ABMIL, we follow DSMIL [63], which crops the WSI into 224x224 sized non-overlapping patches in 20x magnification, and excludes background patches, leaving around 15K patches per bag on average. To train ScAtNet, we

adapted the original skin biopsy patch size while adjusting the number of patches per WSI (10x magnification) to maintain similar content per patch. The result is 35×35, or $1,225$ number of crops. This ensures consistent representation and preserves model architecture.

Feature Extraction We leverage the same embeddings used in DSMIL [63] , which applies a pretrained SimCLR to extract a 512-dimensional feature vector for each patch.

Heuristic Guidance The dataset provides annotated metastasis masks which denote the cancerous regions. We leverage those labels and the tissue masks for **HG** and **TG**.

5.4.3 Implementation Details

The proposed attention guiding framework, **SAG**, is applied to two models: a transformer model, ScAtNet [108], and a MIL model, ABMIL [50]. We use ABMIL's [50] and ScAtNet's [108] public codebases for implementation and train models under their experimental settings.

ScAtNet: We impose **TG** learning across all attention heads and impose **HG** learning on half of the attention heads. This maintains the model's adaptability and accommodates potential noise in **HG**.

ABMIL: We apply both **HG** and **TG** on the melanoma dataset, while we only apply **HG** to Camelyon16 as the preprocessed dataset already excludes background patches.

5.5 Results and Ablation Study

Table 5.1 compares the overall performance of **SAG** on different datasets and backbone models, demonstrating its consistent ability to enhance diagnostic performance in histopathological image analysis. For each setting, we conduct 15 runs of experiments with randomly sampled seeds for model initialization and report the average.

Notably, incorporating **SAG** into single- and multi-scale ScAtNet models on the melanoma dataset yields significant improvements, particularly with multi-scale inputs achieving a 4.55%

Table 5.1: Experimental Results of SAG across single-scale (SC) and multi-scale (MC) configurations for Melanoma and Camelyon16 datasets. Baseline methods are indicated with a †. Performance metrics include Accuracy (Acc), Precision (P), Recall (R), and Area Under the Curve (AUC).

	SAG		Melanoma				Camelyon16			
Methods	HG	TG	Acc	P	R	AUC	Acc	P	R	AUC
ScAtNet (SC)†[108]			55.03	57.17	55.36	77.38	67.79	58.17	57.51	70.28
ScAtNet (SC)	✓		57.14	59.57	57.31	78.75	68.71	58.50	64.01	72.39
ScAtNet (SC)	✓	✓	56.67	60.27	56.66	79.72	**71.60**	**64.45**	61.22	71.87
ScAtNet (MC)†			58.16	61.54	58.21	79.54	66.82	55.98	61.22	69.45
ScAtNet (MC)	✓		59.95	64.77	60.13	81.58	67.91	57.28	**66.39**	72.26
ScAtNet (MC)	✓	✓	**62.71**	**65.23**	**63.34**	**82.03**	70.13	60.53	62.58	**73.13**
Best Improvement △			+4.55	+3.69	+5.13	+2.49	+3.81	+6.28	+6.50	+3.68
ABMIL†[50]			45.55	48.23	46.42	68.07	93.02	92.47	92.79	97.52
ABMIL	✓		51.59	**57.42**	51.02	**74.68**	**94.73**	**94.61**	**94.17**	**97.80**
ABMIL	✓	✓	**52.01**	56.25	**51.84**	74.35	*Not Applicable*			
Best Improvement △			+6.46	+9.19	+5.42	+6.28	+1.71	+2.14	+1.38	+0.28

accuracy increase (Table 5.1). Similar trends are observed on Camelyon16, where **SAG** boosts accuracy across ScAtNet configurations (3.81% for multi-scale) and increases ABMIL's accuracy by 1.71% (Table 5.1). These improvements highlight **SAG**'s effectiveness in refining focus and enhancing the models' diagnostic performance.

In our analysis, we observe that ABMIL exhibits superior diagnostic performance on the Camelyon16 dataset (94.73% vs. 71.60%), whereas ScAtNet is more effective on the melanoma dataset (62.71% vs 45.52%). This distinction in model efficacy can be attributed to the intrinsic characteristics of these datasets and the models' specific designs. Notably, our melanoma dataset, presenting a four-class classification problem, requires a comprehensive understanding of the entire image at multiple scales and holistic levels. This aligns well with ScAtNet's transformer-based architecture, which excels at capturing long-range dependencies and aggregating multi-scale information through attention mechanisms [108]. In contrast,

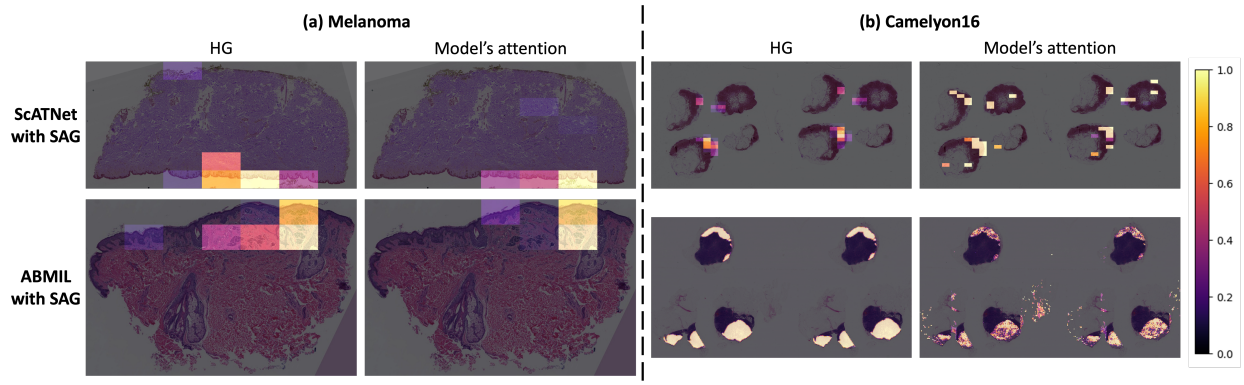


Figure 5.4: Comparative visualizations of **HG** and the models' attention under **SAG**'s training on the melanoma and Camelyon16 datasets. The images are sampled from the test set. The **HG** and attention values are normalized for visualization purpose.

the Camelyon16 dataset, being a binary classification problem, prioritizes local feature identification for diagnosis, which aligns with ABMIL's MIL-based approach, suggesting why ABMIL outperforms in this context. On the other hand, ScAtNet's complexity and multi-scale inputs may not offer significant benefits here due to overfitting risks. This highlights the importance of choosing an appropriate method based on the specific data characteristics.

To qualitatively compare the performance of **SAG**, we visualize the attention patterns of ScAtNet and ABMIL on both datasets compared to **HG** in Figure 5.4. We notice that **SAG** encourages the model to focus on diagnostically relevant regions. These visualizations effectively demonstrate **SAG**'s capacity to guide attention and improve interpretability. We show more comparisons of the **HG**, model's attention without **SAG** (baseline) and model's attention with **SAG** on both datasets in Figure 5.5 and 5.6.

5.6 Summary

Driven by the observation that previous methods often misplace attention on irrelevant regions, we propose this novel framework, Semantics-Aware Attention Guidance (**SAG**). **SAG** incorporates both tissue-based and heuristic attention guidance to more closely emulate

the diagnostic process of pathologists, ensuring that the model focuses on diagnostically meaningful regions and their interconnections within WSIs. By aligning the model's attention with clinically relevant areas, **SAG** enhances performance across diverse datasets, even those with limited size or noisy annotations, thereby significantly advancing the precision and reliability of computational diagnostic systems.

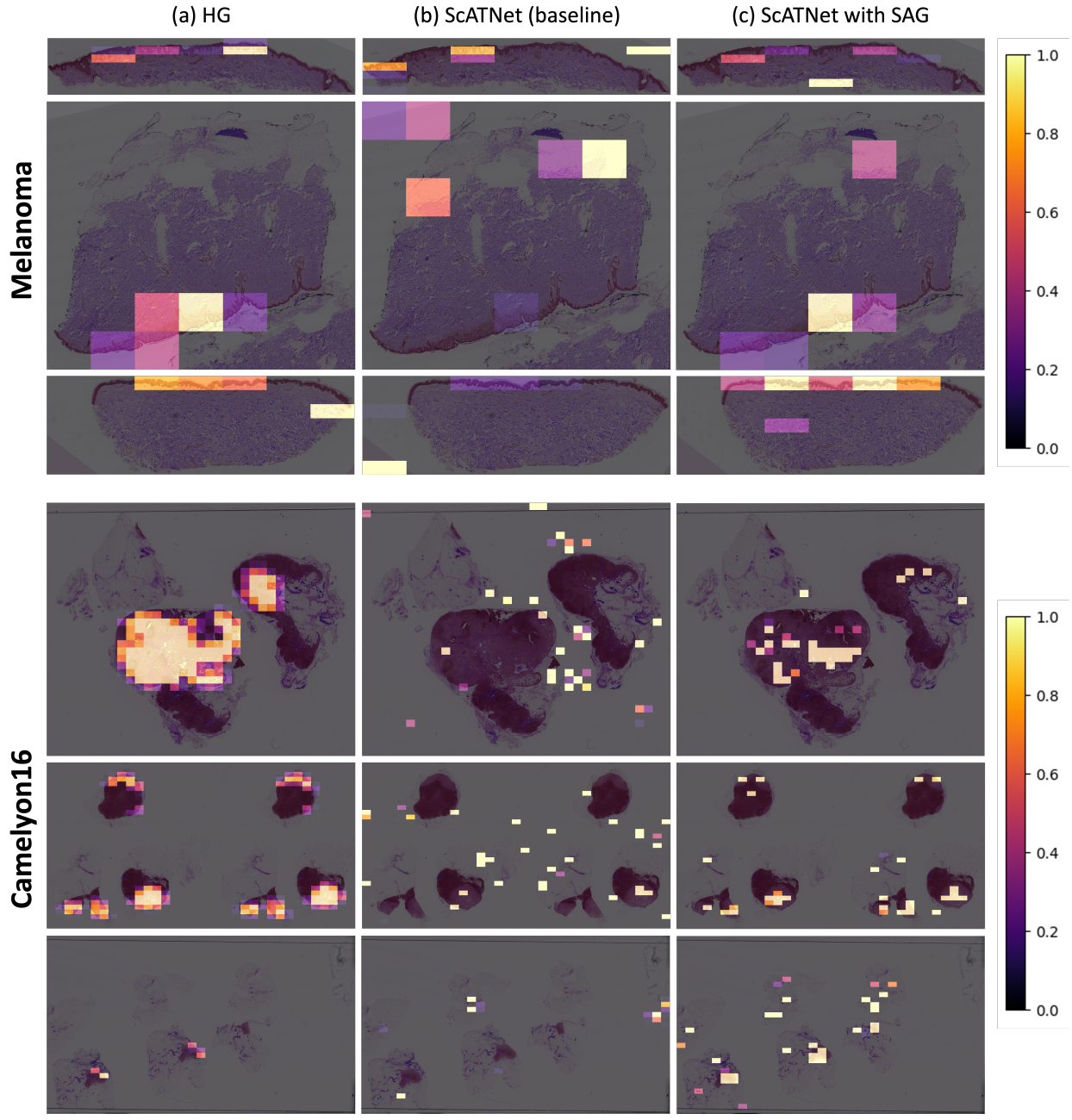


Figure 5.5: Comparison of (a) heuristic guidance (**HG**), (b) ScAtNet (baseline)'s attention, and (c) ScAtNet (with `SAG`)'s attention on the melanoma and Camelyon16 dataset. These images are sampled from the test set. The signal and attention weights are normalized for visualization purpose.

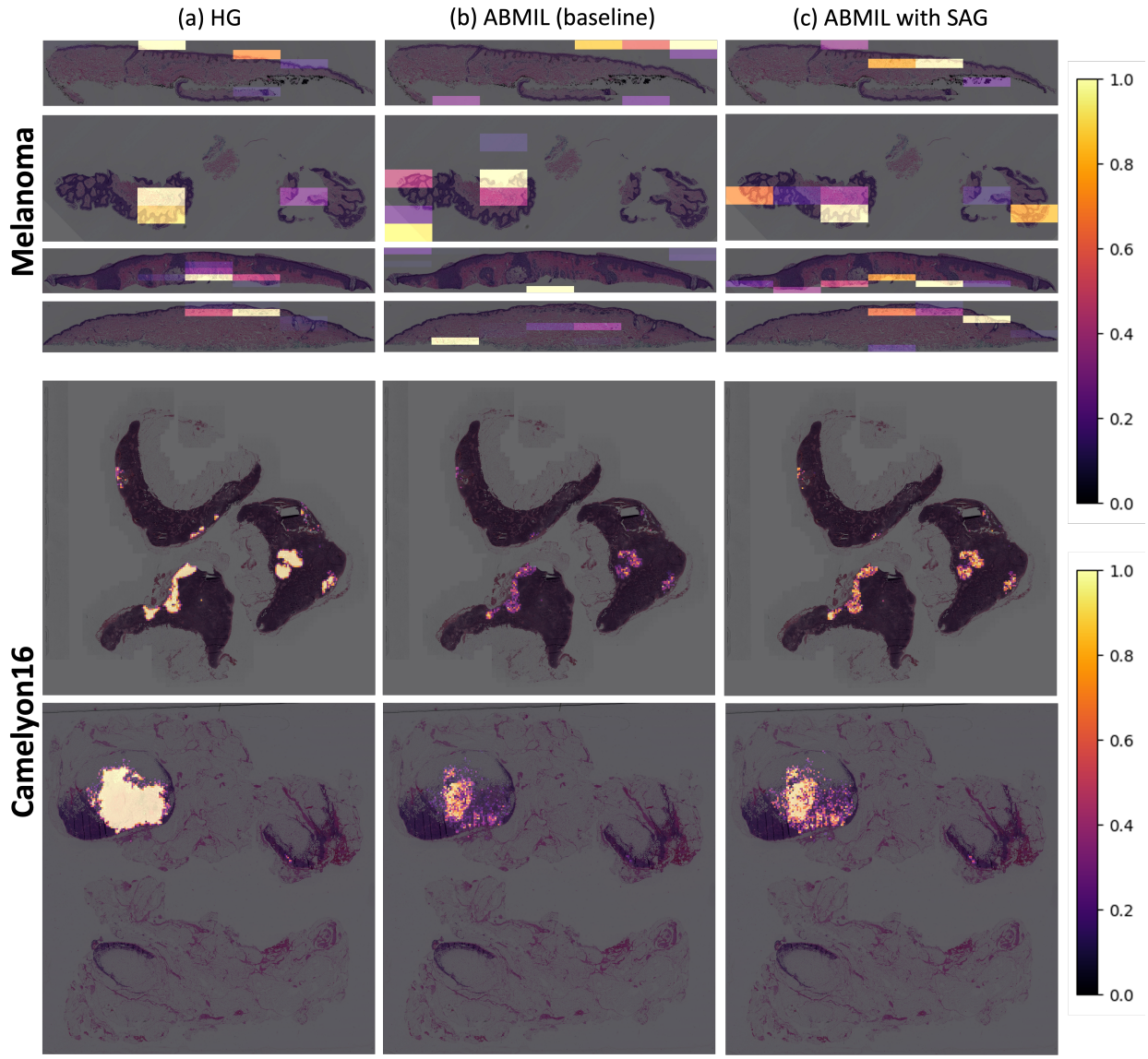


Figure 5.6: Comparison of (a) heuristic guidance (**HG**), (b) ABMIL (baseline)'s attention, and (c) ABMIL (with `SAG`)'s attention on the melanoma and the Camelyon16 dataset. These images are sampled from the test set. The signal and attention weights are normalized for visualization purpose.

Chapter 6

MULTI-LEVEL ROI ATTENDING NETWORK

6.1 Introduction

The automated diagnosis of pathological images presents distinct challenges compared to the classification of natural images or medical images from other modalities in two main aspects: 1) Data scarcity and knowledge transfer limitations pose a significant challenge. While natural image datasets often comprise millions of samples, publicly available datasets for whole slide images (WSIs) are limited in size and diversity. This scarcity hinders the development of generalized foundation models capable of tasks such as zero-shot or few-shot learning. Additionally, the wide range of diseases and organ types in pathology complicates knowledge transfer, as a model trained on one type of pathology may struggle to generalize to others. Effectively leveraging low-resource datasets, therefore, becomes critical for improving diagnostic accuracy, particularly in complex conditions like melanoma. 2) The gigapixel scale and multiscale nature of WSIs present unique computational challenges. WSIs are approximately four orders of magnitude larger than natural images or other medical modalities such as CT scans or X-rays. This massive size necessitates segmenting WSIs into smaller patches for analysis, as processing the entire image in one pass is computationally prohibitive. However, this segmentation process diverges significantly from the clinical workflow of pathologists, who dynamically zoom in and out of digitized WSIs to assess both global tissue-level patterns and fine-grained cellular structures. Moreover, diagnostic information is often represented in small regions-of-interest (ROIs), with the remainder of the WSI contributing little or even misleading information. This multiscale and sparse diagnostic process is not required for natural image classification or other medical imaging modalities, further complicating the development of effective WSI models.

In Chapter 5, we introduced **SAG**, a framework designed to emulate pathologists' diagnostic processes by guiding models to focus on diagnostically relevant regions. This approach demonstrated significant improvements in performance on melanoma datasets, effectively addressing the challenge of data scarcity. However, methods like **SAG** and transformer-based approaches such as **ScAtNet** often take image patches at the same resolution as inputs, failing to capture the intricate correlation between holistic WSI features and localized ROI information. Pathologists, in contrast, rely on both global contextual information and detailed inspection of ROIs, highlighting a critical gap in current computational approaches.

To address this challenge and better emulate the decision-making process of pathologists, it is essential to develop models capable of classifying WSIs by integrating information from both global tissue-level features and localized ROIs. However, the absence of fine-grained annotations poses a significant barrier to achieving optimal performance with deep learning models. Due to the immense size of WSIs, annotating each pixel can require hours of effort from expert pathologists, making the process both time-consuming and prohibitively expensive. As a result, most publicly available WSI datasets are limited to slide-level labels, which restricts the utilization of ROI information.

Existing methods for detecting ROIs in WSIs can be broadly categorized into two main groups: (1) image feature-based approaches and (2) psycho-physical behavior-based approaches. Image feature-based methods leverage statistical or structural tissue characteristics, such as nuclear shape, texture, color distribution, and local binary patterns [85, 7]. These methods are computationally efficient and can provide reliable results; however, they often fail to generalize, particularly in skin cancer analysis, where melanocytes can closely resemble other cell types, leading to challenges in accurate ROI detection. In contrast, psycho-physical behavior-based methods utilize pathologists' interactions with WSIs, such as zooming, panning, and observation duration, to infer the relevance of specific regions [32, 81]. For example, an ROI that has been zoomed into and examined for an extended period is typically interpreted as diagnostically important. While this approach provides insights into pathologists' thought processes, it suffers from several drawbacks. The annotation process

remains labor-intensive, and the resulting labels are often influenced by the annotators' personal biases and habits, reducing the objectivity and consistency of the annotations.

To address the limitations of existing methods, we propose an end-to-end interpretable diagnosis framework, **MiRA** (**M**ult**i**-level **R**egion-of-interest **A**ttended Network). **MiRA** integrates a machine learning-based ROI detection module for automated WSI analysis and combines embeddings from both global WSIs and localized ROIs to produce a robust diagnosis. This dual-level attention approach allows the model to effectively leverage both holistic and detailed information, mimicking the diagnostic workflow of pathologists. The key contributions of this study are summarized as:

- Dynamic and Flexible ROI Retrieval Approaches: We introduce two novel methods, *Top-K* and *Thresholding*, for selecting diagnostically relevant patches. These methods enable adaptive and efficient retrieval of meaningful ROIs.

- Heuristic-Guided ROI Detection: By leveraging **SAG**, we train the ROI detection module to prioritize diagnostically significant areas. This approach is independent of specific image textures, enhancing generalizability across diverse WSI datasets.

- Enhanced Diagnostic Performance: **MiRA** demonstrates significant improvements in melanoma diagnosis accuracy over state-of-the-art models, particularly when applied to single-resolution WSIs. The framework's interpretability and flexibility further contribute to its robustness in clinical applications.

6.2 Related Work

The development of automated diagnostic models for WSIs has been primarily driven by two main methodologies: MIL and transformer-based approaches, with a growing focus on leveraging multi-scale information and incorporating region-of-interest (ROI) embeddings for more precise analysis. MIL has been extensively employed in WSI classification due to its ability to handle weakly labeled data. In this framework, WSIs are segmented into smaller

patches (instances), and classification is performed at either the instance level or the bag level. Instance-level strategies involve training classifiers to predict labels for individual patches and aggregating these predictions to infer the WSI label [16, 56]. Alternatively, embedding-level approaches generate patch-level embeddings, which are then combined to form a bag representation for classification [40, 116, 50]. While effective, MIL methods generally capture local representations and often fail to leverage global inter-patch dependencies. Furthermore, although recent MIL methods analyze patches extracted from multiple resolutions [12, 63], they often utilize all patches and fail to integrate ROI information, which introduces noise and may bias the classification results. Addressing these gaps, **MiRA** incorporates transformer models to capture long-range dependencies and integrates ROI embeddings as additional high-resolution instances to facilitate WSI classification.

Transformer models have revolutionized machine learning, particularly through their success in natural language processing (NLP) tasks [101]. Adaptations of these models, such as vision transformers (ViTs) [22], have demonstrated strong performance in WSI analysis tasks, including cancer diagnosis [29, 108, 96], tissue segmentation [117, 100], and survival prediction [13, 48]. Transformers excel in capturing long-range dependencies through self-attention mechanisms, which are particularly beneficial for gigapixel images. Building upon **ScAtNet**[108], **SAG**[70] introduces a semantics-guiding module that encourages the model to learn diagnostically relevant regions, significantly enhancing both accuracy and interpretability in diagnosis. Inspired by **SAG**, **MiRA** incorporates a novel ROI retrieval module designed to dynamically detect diagnostically significant patches and integrate their embeddings into the classification process. To facilitate this dynamic capability, **MiRA** leverages a token padding technique [20] for handling variable-length inputs, ensuring efficient processing of varying numbers of ROI instances. This integration not only improves computational efficiency but also aligns the model's focus with clinically relevant diagnostic regions, resulting in enhanced diagnostic performance.

The multi-scale nature of WSI analysis reflects pathologists' diagnostic process, where they zoom in and out of tissue samples to examine both cellular and tissue-level features.

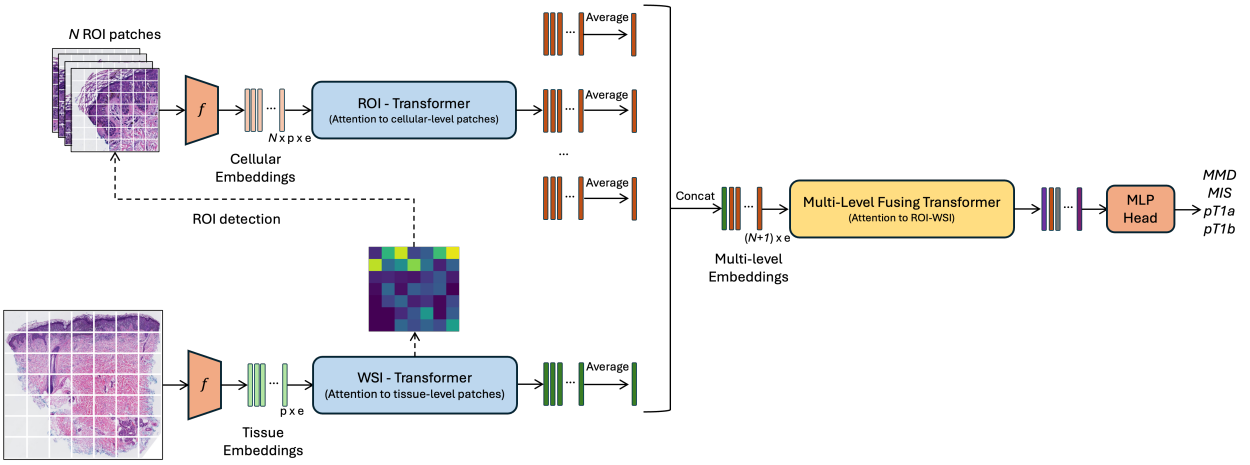


Figure 6.1: Overview of the **MiRA** approach for classifying skin cancer WSIs. p denotes the number of embeddings, while e denotes the length of the embedding. N is the number of retrieved ROIs. Diagnostic terms are defined as: *mild and moderate dysplastic nevi* (MMD), *melanoma in situ* (MIS), *invasive melanoma stage pT1a* ($pT1a$), *invasive melanoma stage* $\geq$ *pT1b* ($pT1b$).

Recent studies have explored multi-scale approaches by integrating hierarchical structures or leveraging transformers to model inter-scale relationships [12, 108, 63]. These methodologies enhance diagnostic accuracy by incorporating features at multiple resolutions. **MiRA** builds on these approaches by selectively attending to diagnostically significant regions and correlating representations across WSI-level and ROI-level using transformers. By fusing cellular-level details with tissue-level patterns, it efficiently integrates multi-level information for robust WSI classification. This approach not only aligns with pathologists' workflows but also enhances computational efficiency by prioritizing diagnostically relevant data.

6.3 Methodology

The **MiRA** model is designed to integrate embeddings from both WSI-level and ROI-level features into biopsy image classification, effectively emulating the clinical decision-making

process of pathologists, who consider both global tissue characteristics and localized, detailed regions. Starting with tissue-level embeddings, **MiRA** employs a WSI transformer to capture representations at the WSI level. A dynamic ROI retrieval module is incorporated to identify diagnostically relevant regions and retrieve their corresponding embeddings within an end-to-end framework. The ROI retrieval module processes precomputed ROI-level features, extracts the embeddings for selected regions, and inputs them into the ROI transformer. These ROI embeddings, together with the WSI-level embeddings, are then concatenated and processed by a multi-level fusing transformer to learn a unified and comprehensive representation for classification. As illustrated in Figure 6.1, the **MiRA** architecture comprises four primary components: 1) generation of multi-level patch-wise embeddings, 2) weakly supervised learning for ROI detection and retrieval of ROI embeddings, 3) efficiently handling varying input lengths, and 4) integration of contextualized ROI and WSI embeddings to facilitate accurate WSI classification. The following sections provide a detailed explanation of each component of the proposed model.

6.3.1 Multi-level Patch Embeddings

As illustrated in Figure 6.2, patch embeddings are extracted separately at the tissue and cellular levels. For tissue-level embeddings, WSIs at a 10x magnification are divided into 7×7 crops, which are further subdivided into 7×7 cellular-level tiles. Similar to the approach in **SAG**, we employ MobileNetV2 [91], an off-the-shelf CNN pretrained on ImageNet, to compute the embeddings at both levels. These precomputed representations are subsequently used as inputs for the classification stages. In our primary study, both tissue- and cellular-level embeddings are generated under the 7×7 cropping scheme and at 10x magnification to ensure consistency. Additional experiments with different magnifications and cropping configurations are conducted as part of an ablation study, which is discussed in detail in Section 6.4.1.

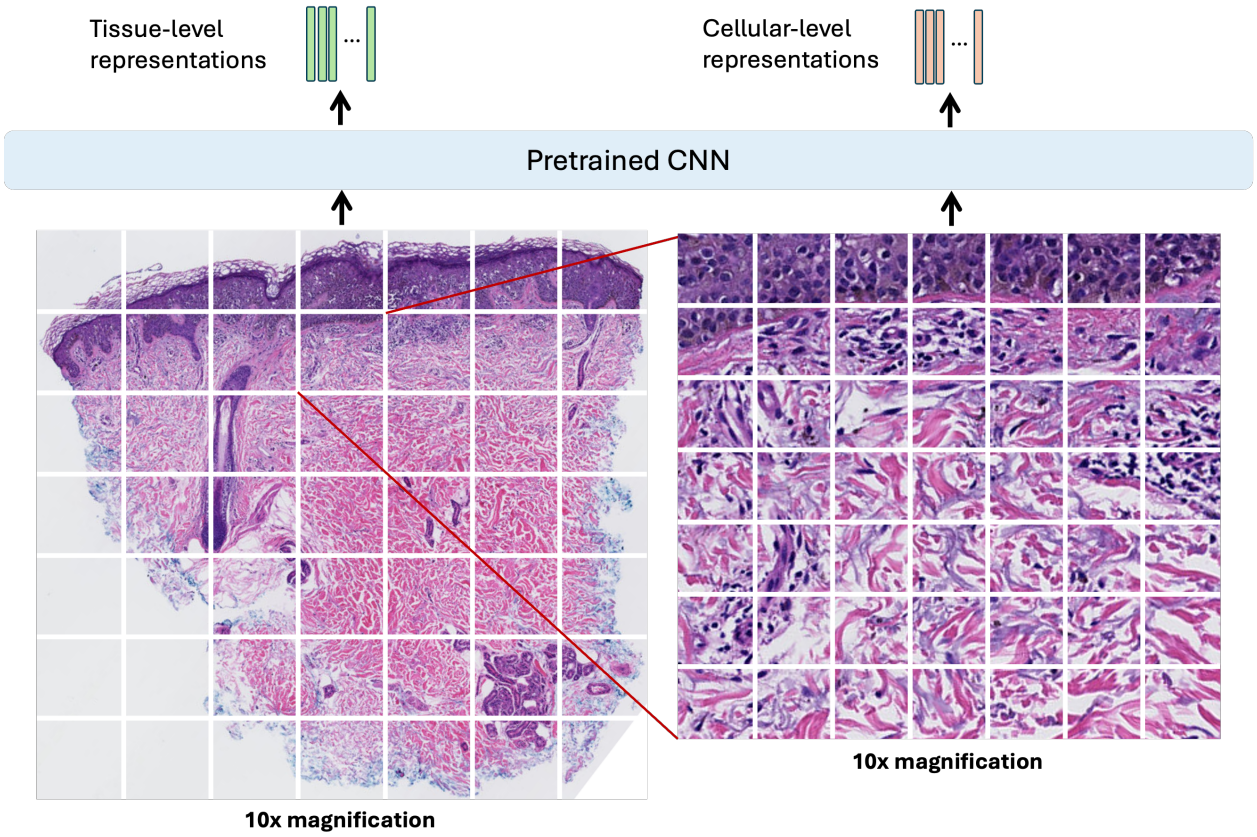


Figure 6.2: Extracting tissue-level and cellular-level patch embeddings: We leverage an off-the-shelf MobileNetV2 [91] to extract different level embeddings from the patch images. For tissue level, we crop the WSI at 10x magnification into 7x7 patches. For cellular level, we crop each tissue level patch into 7x7 sub patches. The embeddings under two separate granularities are utilized in the multi-level branches in **MiRA**.

6.3.2 Weakly Supervised ROI Detection and ROI Embeddings Retrieval

Building upon the approach introduced in **SAG**, we incorporate the attention-guiding loss and the heuristic signals derived from melanocyte masks to train the WSI transformer to focus on diagnostically relevant regions. The attention weights $\mathbf{A}$ computed within the WSI transformer are utilized to localize ROI patches. For extracting ROI patches from $\mathbf{A}$, we

explore two approaches: *Top-K* and *Thresholding*. The *Top-K* method selects the k patches with the highest attention weights, while the *Thresholding* method identifies patches with attention weights exceeding a predefined threshold σ. This weakly supervised ROI detection mechanism allows the model to focus on regions with higher melanocyte density while filtering out irrelevant areas that could introduce noise or bias into the classification process.

We pre-extract the cellular-level embeddings of the fine-grained patches, as described in Section 6.3.1. However, not all embeddings are directly input to the ROI transformer. Instead, we leverage the ROI detection mechanism described above to identify the indices of ROI patches, retrieving only those embeddings. These selected cellular-level embeddings are subsequently processed by the ROI transformer, which learns a distinct representation for each ROI patch.

6.3.3 Contextualized ROI-WSI embeddings

To achieve a comprehensive diagnosis, we integrate holistic WSI-level representations with localized ROI-level representations. These representations are concatenated and processed by the multi-level fusing transformer to produce the final classification.

The *Top-K* selection method ensures a consistent number of ROI embeddings, making it straightforward to manage input dimensions. In contrast, the *Thresholding* approach may result in a variable number of ROIs, depending on the attention weights assigned to the patches. To handle this variability and facilitate batch training, we pad the selected ROI embeddings to a maximum length with placeholder tokens. These tokens are excluded from computation through an attention masking mechanism, following a strategy similar to the padding employed in BERT [20]. This design ensures efficient and consistent processing across varying input lengths, while preserving the model's ability to focus on relevant ROIs.

ROI embeddings, $\mathbf{RE}_i \in \mathbb{R}^e$, are independently generated for each ROI patch i by the cellular-level transformer. These embeddings, along with the WSI embedding $\mathbf{WE} \in \mathbb{R}^e$, are combined to form the multi-level embeddings. However, these embeddings only encode the internal content of their respective patches or the WSI without capturing their relative spatial

relationships. To address this limitation, we introduce backtracked ROI sinusoidal positional encodings $\mathbf{bRPE}_i \in \mathbb{R}^e$. These encodings are added to the ROI embeddings to retain the original spatial positions of the ROI patches within the WSI. For the WSI embedding **WE**, a learnable WSI positional encoding $\mathbf{WPE} \in \mathbb{R}^e$ is incorporated to contextualize its role in the representation.

The resulting multi-level contextualized embeddings **ME**, which combine semantic and spatial information, is defined as:

$$\mathbf{ME} = \text{Concat}(\mathbf{WE} + \mathbf{WPE}, \mathbf{RE}_1 + \mathbf{bRPE}_1, \cdots, \mathbf{RE}_N + \mathbf{bRPE}_N) \tag{6.1}$$

These embeddings are then input to the multi-level fusing transformer, enabling the model to effectively integrate global WSI-level features with localized ROI-level details. This fusion of information supports accurate and interpretable classification outcomes.

6.4 Results and Ablation Study

To evaluate **MiRA**'s effectiveness in melanoma classification, we trained the model on the M-PATH dataset described in Chapter 2.1, utilizing the soft label strategy introduced in Chapter 5.4.1. Since **MiRA** operates on single-scale inputs, we compare its performance against the single-scale versions of **ScAtNet** and **SAG**, ensuring a fair baseline for transformer-based approaches. Additionally, we benchmark **MiRA** against two popular multiple instance learning (MIL) frameworks, ABMIL [50] and ChikonMIL [16], which are widely used in pathology tasks for their ability to aggregate patch-level information into WSI-level representations.

ABMIL applies an attention mechanism to assign learnable weights to patch embeddings, effectively highlighting regions within a WSI while aggregating patch features into a global representation. In contrast, ChikonMIL identifies the top-k patches based on their relevance and subsequently performs instance- and bag-level representation learning. ChikonMIL also incorporates a center loss to minimize intra-class variance and a soft assignment mechanism to map samples to diagnostic centroids, aiming for robust and interpretable classification. Both methods are included to represent distinct MIL paradigms: attention-weighted aggregation

and patch selection followed by feature refinement. All experiments are conducted using WSIs at 10x magnification to maintain consistency across methods and to capture sufficient tissue detail without incurring the computational cost of higher magnifications.

As summarized in Table 6.1, **MiRA** consistently outperforms these baseline methods, achieving superior accuracy in skin cancer WSI classification. Notably, compared to **SAG**, **MiRA** achieves a 2.77% increase in accuracy, underscoring the contribution of its ROI-level branch. This result demonstrates that incorporating fine-grained ROI features enhances the overall image representation, capturing diagnostically critical information that single-scale or holistic methods may overlook.

Table 6.1: Comparison of overall performance with state-of-the-art WSI classification methods across different metrics on the M-PATH test set. Performance metrics include Accuracy, Precision, Recall, and Area Under the Curve. We report the average performance of 10 runs.

Method	Accuracy	Precision	Recall	AUC
ABMIL [50]	45.55	48.23	46.42	68.07
ChikonMIL [16]	56.14	57.22	58.12	75.20
ScAtNet [108]	55.03	57.17	55.36	77.38
SAG [72]	57.14	59.57	57.31	**78.75**
MiRA (Ours)	**59.91**	**60.87**	**61.29**	78.32

6.4.1 ROI Retrieval Methods

We evaluate two methods for retrieving ROI patches: *Top-K* and *Thresholding*. Given that each WSI contains 49 patches in our experimental setup and that most patches represent background or diagnostically irrelevant regions, we select $k = 4, 8, 11, 15$ for the *Top-K* approach. For the *Thresholding* method, we experiment with thresholds $\sigma = 0.05, 0.1, 0.15$ to obtain a similar number of ROI patches as in the *Top-K* approach. To ensure robustness, we

run each configuration 10 times and report the average performance metrics. All experiments are conducted using WSIs at 10x magnification.

The results, summarized in Table 6.2, indicate that the *Top-K* method generally outperforms *Thresholding*. We attribute this to the fixed number of patches used in *Top-K*, which aligns better with scenarios where the cropping count is constant across WSIs. In contrast, the *Thresholding* approach may be more suitable for datasets where the cropping size is uniform, but the number of patches varies significantly between samples.

Table 6.2: Comparison of ROI retrieving approaches. Performance metrics include Accuracy, Precision, Recall, and AUC score. We report the average performance of 10 runs.

Method	Hyperparameter	Accuracy	Precision	Recall	AUC
Top-K	$k = 4$	57.87	60.17	58.12	81.44
	$k = 8$	58.15	60.70	58.34	80.65
	$k = 11$	58.61	60.90	58.74	80.12
	$k = 15$	59.91	60.87	61.29	78.32
Thresholding	$\sigma = 0.05$	55.09	57.25	55.33	78.70
	$\sigma = 0.1$	57.96	60.65	58.04	79.85
	$\sigma = 0.15$	56.67	59.45	56.93	80.61

6.4.2 ROI Resolutions

In our main experiments, cellular-level patches are extracted at 10x magnification. Inspired by multi-scale design principles of **ScAtNet** and **SAG**, we investigate the impact of different resolutions for cellular-level patches while maintaining the same 7×7 cropping configuration. To ensure that critical diagnostic details are retained, we limit our exploration to resolutions exceeding 10x magnification. Specifically, while tissue-level patches are consistently extracted at 10x magnification, we generate cellular-level patches at 10x, 12.5x, and 15x magnification

for comparative analysis.

Table 6.3 presents the classification performance across these varying resolutions. Notably, using 15x magnification achieves an accuracy of 62.13%, closely matching the multi-scale performance of **SAG** (discussed in Chapter 5.1). This demonstrates that aggregating WSI-level and ROI-level representations significantly enhances the model's ability to understand diagnostic features. Moreover, the single-scale approach at higher magnification reduces computational overhead compared to multi-scale inputs, making it a more efficient alternative without compromising accuracy.

Table 6.3: Comparison of cellular-level patches resolution. Performance metrics include Accuracy, Precision, Recall, and AUC score. We report the average performance of 10 runs.

Cellular-level Patch Resolution	Accuracy	Precision	Recall	AUC
10x	59.91	60.87	61.29	78.32
12.5x	58.33	60.31	59.63	79.56
15x	62.13	63.12	63.20	79.75

6.5 Summary

Inspired by the clinical workflow of pathologists, we introduced **MiRA**, a framework that incorporates holistic WSI-level features and localized ROI-level embeddings for comprehensive melanoma classification. Leveraging a weakly supervised ROI detection mechanism, **MiRA** identifies diagnostically significant regions and integrates them within an end-to-end transformer-based architecture. By employing an attention-guiding loss and dynamic ROI selection strategies, **MiRA** ensures a robust focus on relevant areas, refining the diagnostic process.

MiRA demonstrates superior performance on the M-PATH dataset, outperforming **SAG** and other competitive WSI classification methods by effectively aggregating multi-level

embeddings. Through its tailored ROI resolution experiments, **MiRA** balances computational efficiency with high diagnostic accuracy, offering a significant improvement over traditional multi-scale approaches. Furthermore, by emulating the diagnostic approach of pathologists, **MiRA** underscores the potential of computer-aided diagnosis systems and provides more insights into the explainable AI-driven solutions in healthcare.

Chapter 7

CONCLUSION

This dissertation presents a comprehensive exploration of advancing computational diagnostic systems for melanoma using whole slide images. By integrating innovations in machine learning, computer vision, and clinical pathology, each research project has contributed to developing efficient methodologies that align closely with pathologists' diagnostic workflows.

Identifying Melanocytic Proliferation Chapter 3 presented a weakly supervised framework [69] for detecting and segmenting melanocytic proliferations in WSIs. By incorporating weighted loss functions, the model effectively mitigates challenges posed by sparse and noisy annotations. This work serves as an essential initial step in developing automated diagnostic pipelines while underscoring the critical need for high-quality, curated datasets to achieve robust model performance.

Virtual Staining Guided Melanocyte Detection In Chapter 4, the dissertation introduced **VSGD-Net**[68], a pioneering model designed to accurately detect melanocytes in H&E-stained WSIs through virtual staining and knowledge transfer. In this project, we first introduced an automated method to produce pseudo melanocyte labels from SOX-10 stained WSIs. Inspired by the fact that SOX10 staining can highlight melanocytes in different color, we built **VSGD-Net** that leverages shared feature representations to jointly perform virtual staining and melanocyte detection tasks. A significant discovery is the mutual enhancement of both tasks when utilizing shared features, leading to improved melanocyte detection results foundational for subsequent projects.

Expanding on **VSGD-Net**, the dissertation introduced **CC-WSI-Net**[72], a novel framework for generating seamless synthetic WSIs. By employing color prompting and

consistency loss functions, **CC-WSI-Net** breaks the limitations of patch-based synthesis, enabling whole-slide image generation. This innovative approach lays the groundwork for utilizing synthetic WSIs as powerful tools in diagnostic assistance and medical research.

Semantics-Aware Attention Guidance Chapter 5 introduced **SAG** [70], an innovative module that significantly advances the classification performance of attention-based models, including transformers and MIL approaches, for whole slide image (WSI) diagnosis. The primary contributions of **SAG** include a heuristic attention-generation mechanism that translates diagnostically relevant cellular entities into numerical attention weights and an attention-guiding loss designed to supervise the model's learning across diverse semantic signals. By directly steering the model's attention towards diagnostically critical regions, **SAG** not only boosts classification accuracy across two cancer datasets but also enhances the interpretability of WSI diagnostic models. This alignment of model attention with clinical insights highlights the potential of **SAG** to bridge the gap between computational predictions and pathologists' diagnostic workflows, providing both technical improvements and practical relevance in clinical settings.

Multi-Level ROI Attention Network Chapter 6 presented **MiRA**, a multi-level classification framework designed to closely emulate the diagnostic reasoning of expert pathologists. This approach integrates global features derived from WSIs with localized insights obtained from dynamically selected ROIs. By bridging these two perspectives, **MiRA** achieves a holistic understanding of the pathological landscape, enhancing its diagnostic accuracy. Furthermore, **MiRA** improves the interpretability of computational diagnostic (CAD) systems by offering insights into how its decisions are formed, aligning its reasoning with clinical practices.

These projects collectively demonstrate the power of deep learning in analyzing WSIs and mark significant progress toward automated cancer diagnosis. They not only showcase advancements in diagnostic accuracy but also emphasize improving the interpretability of AI

models, paving the way for more transparent and clinically viable tools.

7.1 Limitations and Future Work

While the presented projects demonstrate significant progress on the automated analysis of melanoma WSIs, several limitations remain that highlight opportunities for future research. These include challenges in investigating more extensive datasets, testing the clinical use of the synthetic WSIs, and improving transparency in the interpretation of diagnosis models. Addressing these limitations will be crucial for refining the models and exploring their full potential for widespread clinical application.

Limited Datasets Access to larger and more diverse datasets is crucial for enhancing the generalizability and robustness of the proposed methods. The datasets utilized in the projects are limited in both size and variety of cancer types, which may constrain the broader applicability of the models.

For instance, the melanocyte detection dataset, although comprising 25,314 patches, is derived from only 15 WSIs. This narrow sampling scope may introduce biases stemming from specific cases or variations in staining quality. To mitigate this limitation, we have expanded the paired H&E-SOX10 dataset to include 76 cases, thereby establishing a more representative basis for future research in melanocyte detection.

Additionally, future virtual staining studies could greatly benefit from datasets containing a broader range of diseases and imaging modalities. In our virtual staining research, we leverage melanocyte segmentation masks as supervision and evaluate the model's performance. Expanding on this strategy, future work could incorporate medical foundation models to extract comprehensive cell and tissue entities from additional datasets. This approach would help reduce model hallucinations, ensure the precision of generated virtual staining, and explore new possibilities for computational pathology.

Similarly, the M-PATH dataset used for training and evaluating diagnostic models comprises only 222 WSIs. This limited dataset may not adequately capture the variability

encountered in routine clinical diagnostics. Expanding the dataset to include a larger number of cases, along with samples from diverse cancer types, is a vital step forward. Such efforts would enhance the robustness of the models and ensure their applicability across different clinical settings.

Applications of Synthetic WSIs In **CC-WSI-Net**, we propose a novel methodology for generating seamless WSIs and demonstrate its application to SOX10 virtual staining. Subjective assessment results indicate that the synthetic WSIs are highly realistic and virtually indistinguishable from authentic SOX10-stained WSIs. However, the clinical applicability of these synthetic WSIs, particularly their potential role in aiding CAD systems, remains an open area of research. Future studies could explore integrating synthetic WSIs as an additional input modality in diagnostic models, potentially enriching the models' understanding of cellular and tissue-level patterns. By incorporating synthetic WSIs into multi-modal frameworks, we could investigate their capacity to enhance diagnostic accuracy and interpretability, ultimately paving the way for their broader applications in computational pathology.

Transparent Interpretation Another limitation of the proposed methods is the lack of transparent and intuitive interpretation of the decision-making process. While **SAG** and **MiRA** improve reliability and interpretability by training models to focus on diagnostically relevant regions, they fall short of providing clear, explicit reasoning that could align with pathologists' expectations. Without straightforward explanations into why certain regions are prioritized or how predictions are made, these methods may face challenges in gaining clinical trust and regulatory approval.

To address this, future work could explore integrating large language models (LLMs) to generate human-readable explanations of the model's decision-making process. By linking visual attention mechanisms with natural language descriptions, such as summarizing why specific regions were deemed diagnostically relevant, LLMs could bridge the gap between complex model outputs and clinical human-readable interpretation.

Another promising direction would involve combining segmentation models with malignancy classification to provide localized visual interpretations. By segmenting different tissue types and associating malignancy predictions with specific regions, models could offer a layered explanation of their reasoning. This approach could not only enhance interpretability but also enable pathologists to validate results more effectively, fostering trust and adoption in clinical workflows.

BIBLIOGRAPHY

[1] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7941–7950, 2021.

[2] Kimberly H Allison, Lisa M Reisch, Patricia A Carney, Donald L Weaver, Stuart J Schnitt, Frances P O'Malley, Berta M Geller, and Joann G Elmore. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology*, 65(2):240–251, 2014.

[3] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. *arXiv:1802.06955 [cs]*, May 2018. arXiv: 1802.06955.

[4] American Cancer Society. Cancer facts and figures 2024, 2024. Retrieved from https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures.html.

[5] Axel Andersson, Nadezhda Koriakina, Nataša Sladoje, and Joakim Lindblad. End-to-end multiple instance learning with gradient accumulation. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2742–2746. IEEE, 2022.

[6] 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.

[7] Jocelyn Barker, Assaf Hoogi, Adrien Depeursinge, and Daniel L Rubin. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. *Medical Image Analysis*, 30:60–71, 2016.

[8] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

[9] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[10] 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.

[11] Kemeng Chen, Ning Zhang, Linda Powers, and Janet Roveda. Cell Nuclei Detection and Segmentation for Computational Pathology Using Deep Learning. In *2019 Spring Simulation Conference (SpringSim)*, pages 1–6, April 2019.

[12] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

[13] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021.

[14] 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.

[15] Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C Thai, Kathleen Moore, Robert S Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, 79:102444, 2022.

[16] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 519–528. Springer, 2020.

[17] 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.

[18] Kevin de Haan, Yijie Zhang, Jonathan E Zuckerman, Tairan Liu, Anthony E Sisk, Miguel FP Diaz, Kuang-Yu Jen, Alexander Nobori, Sofia Liou, Sarah Zhang, et al. Deep learning-based transformation of h&e stained tissues into special stains. *Nature Communications*, 12(1):1–13, 2021.

[19] V Della Mea, Fabio Puglisi, Mariella Bonzanini, Stefano Forti, Vito Amoroso, Roberta Visentin, P Dalla Palma, and Carlo A Beltrami. Fine-needle aspiration cytology of the breast: a preliminary report on telepathology through internet multimedia electronic mail. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, 10(6):636–641, 1997.

[20] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer*, 124(4):686–696, 2021.

[24] David E Elder, Megan M Eguchi, Raymond L Barnhill, Kathleen F Kerr, Stevan R Knezevich, Michael W Piepkorn, Lisa M Reisch, and Joann G Elmore. Diagnostic error, uncertainty, and overdiagnosis in melanoma. *Pathology*, 55(2):206–213, 2023.

[25] 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.

[26] Elmore et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*, 2015.

[27] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[28] Navid Farahani, Anil V Parwani, and Liron Pantanowitz. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, pages 23–33, 2015.

[29] Zeyu Gao, Bangyang Hong, Xianli Zhang, Yang Li, Chang Jia, Jialun Wu, Chunbao Wang, Deyu Meng, and Chen Li. Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 299–308. Springer, 2021.

[30] Zeyu Gao, Jiangbo Shi, Xianli Zhang, Yang Li, Haichuan Zhang, Jialun Wu, Chunbao Wang, Deyu Meng, and Chen Li. Nuclei Grading of Clear Cell Renal Cell Carcinoma in Histopathological Image by Composite High-Resolution Network. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 132–142, Cham, 2021. Springer International Publishing.

[31] C Garbe, U Keim, TK Eigentler, T Amaral, A Katalinic, B Holleczek, P Martus, and U Leiter. Time trends in incidence and mortality of cutaneous melanoma in germany. *Journal of the European Academy of Dermatology and Venereology*, 33(7):1272–1280, 2019.

[32] Fatemeh Ghezloo, Oliver H Chang, Stevan R Knezevich, Kristin C Shaw, Kia Gianni Thigpen, Lisa M Reisch, Linda G Shapiro, and Joann G Elmore. Robust roi detection in whole slide images guided by pathologists' viewing patterns. *Journal of Imaging Informatics in Medicine*, pages 1–16, 2024.

[33] 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.

[34] Xuan Gong, Shuyan Chen, Baochang Zhang, and David Doermann. Style Consistent Image Generation for Nuclei Instance Segmentation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3993–4002, Waikoloa, HI, USA, January 2021. IEEE.

[35] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, volume Advances in Neural Information Processing Systems, June 2014. arXiv: 1406.2661.

[36] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

[37] 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.

[38] Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 755–764. Springer, 2023.

[39] Robert Harb, Thomas Pock, and Heimo Müller. Diffusion-based generation of histopathological whole slide images at a gigapixel scale. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5131–5140, 2024.

[40] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3852–3861, 2020.

[41] 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.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016. IEEE.

[43] 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.

[44] Yuta Hiasa, Yoshito Otake, Masaki Takao, Takumi Matsuoka, Kazuma Takashima, Aaron Carass, Jerry L Prince, Nobuhiko Sugano, and Yoshinobu Sato. Cross-modality image synthesis from unpaired data using cyclegan. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 31–41. Springer, 2018.

[45] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[46] Md Shakhawat Hossain, Galib Muhammad Shahriar, MM Mahbubul Syeed, Mohammad Faisal Uddin, Mahady Hasan, Shingla Shivam, and Suresh Advani. Region of interest (roi) selection using vision transformer for automatic analysis using whole slide images. *Scientific Reports*, 13(1):11314, 2023.

[47] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2424–2433, 2016.

[48] Ziwang Huang, Hua Chai, Ruoqi Wang, Haitao Wang, Yuedong Yang, and Hejun Wu. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 561–570. Springer, 2021.

[49] Henning Höfener, André Homeyer, Nick Weiss, Jesper Molin, Claes F. Lundström, and Horst K. Hahn. Deep learning nuclei detection: A simple approach can deliver state-of-the-art results. *Computerized Medical Imaging and Graphics*, 70:43–52, December 2018.

[50] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pages 2127–2136. PMLR, 2018.

[51] 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.

[52] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[53] 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.

[54] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022.

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

[56] Fahdi Kanavati, Gouji Toyokawa, Seiya Momosaki, Michael Rambeau, Yuka Kozuma, Fumihiro Shoji, Koji Yamazaki, Sadanori Takeo, Osamu Iizuka, and Masayuki Tsuneki. Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific Reports*, 10(1):9297, 2020.

[57] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[58] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[59] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[60] 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.

[61] Amal Lahiani, Irina Klaman, Nassir Navab, Shadi Albarqouni, and Eldad Klaiman. Seamless virtual whole slide image synthesis and validation using perceptual embedding consistency. *IEEE Journal of Biomedical and Health Informatics*, 25(2):403–411, 2020.

[62] 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.

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

[64] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1952–1961, 2023.

[65] 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.

[66] 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.

[67] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai. Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4243–4252, 2020.

[68] Kechun Liu, Beibin Li, Wenjun Wu, Caitlin May, Oliver Chang, Stevan Knezevich, Lisa Reisch, Joann Elmore, and Linda Shapiro. Vsgd-net: Virtual staining guided melanocyte detection on histopathological images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1918–1927, 2023.

[69] Kechun Liu, Mojgan Mokhtari, Beibin Li, Shima Nofallah, Caitlin May, Oliver Chang, Stevan Knezevich, Joann Elmore, and Linda Shapiro. Learning melanocytic proliferation segmentation in histopathology images from imperfect annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3766–3775, 2021.

[70] Kechun Liu, Wenjun Wu, Joann G Elmore, and Linda G Shapiro. Semantics-aware attention guidance for diagnosing whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 595–605. Springer, 2024.

[71] Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He. Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE Transactions on Medical Imaging*, 40(8):1977–1989, 2021.

[72] Sitong Liu, Kechun Liu, Samuel Margolis, Wenjun Wu, Stevan R Knezevich, David E Elder, Megan M Eguchi, Joann G Elmore, and Linda Shapiro. Generating seamless virtual immunohistochemical whole slide images with content and color consistency. *arXiv preprint arXiv:2410.01072*, 2024.

[73] Yiming Liu, Pengcheng Zhang, Qingche Song, Andi Li, Peng Zhang, and Zhiguo Gui. Automatic segmentation of cervical nuclei based on deep learning and a conditional random field. *IEEE Access*, 6:53709–53721, 2018.

[74] 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.

[75] Johannes Lotz, Nick Weiss, Jeroen van der Laak, and StefanHeldmann. High-resolution Image Registration of Consecutive and Re-stained Sections in Histopathology. *arXiv:2106.13150 [cs, eess]*, June 2021. arXiv: 2106.13150.

[76] 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.

[77] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. ESPNetv2: A Light-Weight, Power Efficient, and General Purpose Convolutional Neural Network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9182–9192, Long Beach, CA, USA, June 2019. IEEE.

[78] Claudia Mello-Thoms, Carlos AB Mello, Olga Medvedeva, Melissa Castine, Elizabeth Legowski, Gregory Gardner, Eugene Tseytlin, and Rebecca Crowley. Perceptual analysis of the reading of dermatopathology virtual slides by pathology residents. *Archives of Pathology & Laboratory Medicine*, 136(5):551–562, 2012.

[79] Caner Mercan, Bulut Aygunes, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Deep feature representations for variable-sized regions of interest in breast histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(6):2041–2049, 2020.

[80] Caner Mercan, Germonda Reijnen-Mooij, David Tellez Martin, J. Lotz, Nick Weiss, M. V. Gerven, and F. Ciompi. Virtual Staining for Mitosis Detection in Breast Histopathology. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020.

[81] Ezgi Mercan, Selim Aksoy, Linda G Shapiro, Donald L Weaver, Tad T Brunyé, and Joann G Elmore. Localization of diagnostically relevant regions of interest in whole slide images: a comparative study. *Journal of Digital Imaging*, 29:496–506, 2016.

[82] Kevin Miao, Akash Gokul, Raghav Singh, Suzanne Petryk, Joseph Gonzalez, Kurt Keutzer, and Trevor Darrell. Prior knowledge-guided attention in self-supervised vision transformers. *arXiv preprint arXiv:2209.03745*, 2022.

[83] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[84] Andriy Myronenko, Ziyue Xu, Dong Yang, Holger R Roth, and Daguang Xu. Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 329–338. Springer, 2021.

[85] Anupiya Nugaliyadde, Kok Wai Wong, Jeremy Parry, Ferdous Sohel, Hamid Laga, Upeka V Somaratne, Chris Yeomans, and Orchid Foster. Rcnn for region of interest detection in whole slide images. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part V 27*, pages 625–632. Springer, 2020.

[86] 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.

[87] Michael W Piepkorn, Raymond L Barnhill, David E Elder, Stevan R Knezevich, Patricia A Carney, Lisa M Reisch, and Joann G Elmore. The mpath-dx reporting schema for melanocytic proliferations and melanoma. *Journal of the American Academy of Dermatology*, 70(1), 2014.

[88] Shan E Ahmed Raza, Linda Cheung, Muhammad Shaban, Simon Graham, David Epstein, Stella Pelengaris, Michael Khan, and Nasir M Rajpoot. Micro-net: A unified model for segmentation of various objects in microscopy images. *Medical Image Analysis*, 52:160–173, 2019.

[89] 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.

[90] 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.

[91] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[92] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 265–273. Springer, 2018.

[93] Selim Seferbekov. DSB2018 [ods.ai] topcoders 1st place solution, February 2022. original-date: 2018-04-10T20:06:11Z.

[94] M. Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain Style Transfer for Digital Histological Images. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019.

[95] M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 953–956. IEEE, 2019.

[96] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

[97] Jiangbo Shi, Lufei Tang, Yang Li, Xianli Zhang, Zeyu Gao, Yefeng Zheng, Chunbao Wang, Tieliang Gong, and Chen Li. A structure-aware hierarchical graph-based multiple instance learning framework for pt staging in histopathological image. *IEEE Transactions on Medical Imaging*, 2023.

[98] Rebecca L. Siegel, Kimberly D. Miller, Heather E. Fuchs, and Ahmedin Jemal. Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(1):7–33, 2022.

[99] Kexin Sun, Zhineng Chen, Gongwei Wang, Jun Liu, Xiongjun Ye, and Yu-Gang Jiang. Bi-directional feature fusion generative adversarial network for ultra-high resolution pathological image virtual re-staining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3904–3913, 2023.

[100] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, part I 24*, pages 36–46. Springer, 2021.

[101] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[102] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[103] Aarno Oskar Vuola, Saad Ullah Akram, and Juho Kannala. Mask-RCNN and U-Net Ensembled for Nuclei Segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 208–212, April 2019. ISSN: 1945-8452.

[104] Shidan Wang, Ruichen Rong, Donghan M Yang, Junya Fujimoto, Shirley Yan, Ling Cai, Lin Yang, Danni Luo, Carmen Behrens, Edwin R Parra, et al. Computational staining of pathology images to study the tumor microenvironment in lung cancer. *Cancer Research*, 80(10):2056–2066, 2020.

[105] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, Salt Lake City, UT, USA, June 2018. IEEE.

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

[107] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[108] Wenjun Wu, Sachin Mehta, Shima Nofallah, Stevan Knezevich, Caitlin J May, Oliver H Chang, Joann G Elmore, and Linda G Shapiro. Scale-aware transformers for diagnosing melanocytic lesions. *IEEE Access*, 9:163526–163541, 2021.

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

[110] Xide Xia and Brian Kulis. W-net: A deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*, 2017.

[111] 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.

[112] Zhaoyang Xu, Carlos Fernández Moro, Béla Bozóky, and Qianni Zhang. GAN-based Virtual Re-Staining: A Promising Solution for Whole Slide Image Analysis. *arXiv:1901.04059 [cs]*, January 2019. arXiv: 1901.04059.

[113] 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.

[114] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Zongben Xu, and Jerry Prince. Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 174–182. Springer, 2018.

[115] Linfeng Yang, Rajarshi P. Ghosh, J. Matthew Franklin, Simon Chen, Chenyu You, Raja R. Narayan, Marc L. Melcher, and Jan T. Liphardt. NuSeT: A deep learning tool for reliably separating and analyzing crowded cells. *PLOS Computational Biology*, 16(9):e1008193, September 2020. Publisher: Public Library of Science.

[116] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022.

[117] Jiawei Zhang, Yanchun Zhang, and Xiaowei Xu. Pyramid u-net for retinal vessel segmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1125–1129. IEEE, 2021.

[118] Jun Zhang and Jinglu Hu. Image segmentation based on 2d otsu method with histogram analysis. In *2008 International Conference on Computer Science and Software Engineering*, volume 6, pages 105–108. IEEE, 2008.

[119] Bingchao Zhao, Xin Chen, Zhi Li, Zhiwen Yu, Su Yao, Lixu Yan, Yuqian Wang, Zaiyi Liu, Changhong Liang, and Chu Han. Triple u-net: Hematoxylin-aware nuclei segmentation with progressive dense feature aggregation. *Medical Image Analysis*, 65:101786, 2020.

[120] Jianfeng Zhao, Dengwang Li, Zahra Kassam, Joanne Howey, Jaron Chong, Bo Chen, and Shuo Li. Tripartite-GAN: Synthesizing liver contrast-enhanced MRI to improve tumor detection. *Medical Image Analysis*, 63:101667, July 2020.

[121] Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 41(11):3003–3015, 2022.

[122] Yushan Zheng, Zhiguo Jiang, Fengying Xie, Jun Shi, Haopeng Zhang, Jianguo Huai, Ming Cao, and Xiaomiao Yang. Diagnostic regions attention network (dra-net) for histopathology wsi recommendation and retrieval. *IEEE Transactions on Medical Imaging*, 40(3):1090–1103, 2020.

[123] Yanning Zhou, O. F. Onder, Q. Dou, E. Tsougenis, Hao Chen, and P. Heng. CIA-Net: Robust Nuclei Instance Segmentation with Contour-aware Information Aggregation. *IPMI*, 2019.

[124] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

[125] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

VITA

Kechun Liu received her B.Eng. degree in Electronic Engineering from Tsinghua University in China. She is currently a Ph.D. candidate in Computer Science and Engineering at the University of Washington. Her research focuses on tackling real-world clinical problems with computer vision and machine learning techniques.

She welcomes your comments to kechun@uw.edu.