Transformative Diagnostics: Applying Transformer Networks and Semantic Guidance to Whole Slide Images

Wenjun Wu

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

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Linda Shapiro, Chair
Sachin Mehta
Joann Elmore

Program Authorized to Offer Degree:
Biomedical Informatics and Medical Education

University of Washington

Abstract

Transformative Diagnostics: Applying Transformer Networks and Semantic Guidance to Whole Slide Images

Wenjun Wu

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

This dissertation advances the field of digital pathology by introducing innovative deep learning approaches to improve the analysis and diagnosis of skin and breast cancers from whole slide images (WSIs). Given the complexity and variability inherent in WSIs, traditional diagnostic methods often struggle with accuracy and efficiency. This work addresses these challenges through a series of projects leveraging advanced segmentation techniques, transformer-based models, and a novel Semantics-Aware Attention Guidance (`SAG`) framework.

The initial focus of the research is on enhancing the detection and segmentation of diagnostically significant structures within WSIs. The introduction of VSGD-Net and a two-stage segmentation approach demonstrates significant improvements in identifying melanocytes and other critical features with minimal reliance on extensive annotated data. Building on this foundation, the dissertation explores the application of transformer networks, such as HATNet and ScATNet, utilizing self-attention mechanisms to effectively learn contextual relationships across different scales in WSIs.

The culmination of this research is the development of the `SAG` framework, which integrates semantic information into the diagnostic process, guiding attention mechanisms to focus on areas of potential malignancy. This approach not only enhances the accuracy and

precision of the models but also improves their interpretability, a critical factor in clinical settings.

Empirical evaluations across multiple cancer datasets demonstrate that the proposed methods outperform existing state-of-the-art models in terms of diagnostic accuracy, robustness, and efficiency. These advancements hold significant promise for transforming cancer diagnosis, providing pathologists with powerful tools to enhance decision-making and potentially improve patient outcomes.

By bridging the gap between computational models and clinical applications, this dissertation contributes to the broader goal of utilizing artificial intelligence in medicine to facilitate early detection, accurate diagnosis, and personalized treatment of cancer.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

As I navigate the challenging path of my PhD, I feel deeply grateful to the many individuals who have supported me along the way. Foremost is my PhD advisor, Linda Shapiro. Her exceptional dedication to her students and her unwavering confidence in my abilities have been incredibly inspiring. I know that completing this PhD would have been much more difficult without her guidance. Linda's support has been a consistent source of motivation for me, and I am profoundly thankful for everything she has done.

I am profoundly grateful to Dr. Sachin Mehta and Dr. Joann Elmore for their invaluable contributions to my doctoral journey. Dr. Sachin Mehta, a senior labmate, has been a pillar of support and an exceptional mentor throughout my PhD. His guidance, patience, and deep understanding of our research field have significantly shaped my academic experience, providing me with the insights and encouragement needed to navigate the complexities of this journey. I also owe a great debt of gratitude to Dr. Joann Elmore for her mentorship and for sharing her extensive medical domain knowledge. As a co-PI of my advisor, Dr. Elmore has enriched my research with critical perspectives on medical applications, which have been instrumental in grounding my work in real-world clinical relevance. Her guidance on various papers and projects has not only enhanced my research outcomes but also my growth as a scholar. Their support has been pivotal in my development and achievements, and I am profoundly thankful for their mentorship.

I extend my deepest thanks to my committee members, Dr. Kat Steele and Dr. Neil F. Abernethy, for their insightful feedback on my talks and dissertation, which has significantly improved my work. Additionally, I am grateful to my early academic mentors, Dr. Janani Venugopalan and Dr. Zhenglun Wei, for introducing me to the world of research during my

undergraduate studies and influencing my research interests and career trajectory.

I am also indebted to my fantastic labmates who have contributed to various projects throughout this journey: Nicholas Nuechterlein, Ezgi Mercan, Bindita Chaudhuri, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Jie Gao, Kalyani Sunil Marathe, Anandithaa Raghunath, Zixuan Liu, Wisdom Ikezogwo, and Rajesh Rao.

Moreover, I would like to thank all my co-authors: Kechun Liu, Ximing Lu, Beibin Li, Shima Nofallah, Donald Weaver, Mojgan Mokhtari, Stevan Knezevich, Caitlin May, Oliver Chang, and other members of the MPATH and IMPACT groups. Without their collaboration, my publications would not have been possible.

I owe my mental well-being to my friends who have been pillars of support throughout the challenging phases of my PhD, especially during the pandemic. Special thanks to Xingfan Huang, Ellen Xu, Jasmine Li, Sonic Yao, Ernasto Chen, Di Huang, and Aleksander Holynski for our victorious moments as the "Little Giants" on the volleyball court. I will always cherish those memories.

I also thank Sinan Xie, Xingfan Huang, Qisheng Li, Beibin Li, Chien-yu Lin, Jiayin Qu, Yue Guo, Hao Peng, Xiaoshi Quan, and Aleksander Holynski for exploring the roads and mountains of the Pacific Northwest with me.

My sincere thanks to Lianhui Qin, Raven Yan, Qianye Mei, Dylan Du, and Yiran Zhao for the countless enriching conversations we've had over the years.

Special thanks go to Gwen Yimiao the Cat, Hulu and Sydney the Cats, Sweet Potato the Corgi, and Popcorn the Goldendoodle. Their endearing presence has been a constant source of comfort and delight.

Lastly, my deepest gratitude goes to my family — my mother, Yiqing Lu, and my father, Qingping Wu. Their encouragement during challenging times has been invaluable. I will always be grateful for their unconditional support.

Chapter 1

INTRODUCTION

Cancer occurs when normal cell growth goes awry. Instead of following the body's instructions, some cells multiply uncontrollably, forming tumors that can be malignant or benign. This abnormal growth can occur in various tissues throughout the body, leading to a diverse range of cancers such as lung, brain, and skin cancers. For many cancers, the "gold standard" for diagnosis relies on the visual assessments of biopsy specimens from pathologists. Unfortunately, diagnostic errors are common, and even expert pathologists may not reach consensus on diagnostically challenging cases in many areas within pathology [5–8]. For instance, pathologists disagree in up to 60% of cases when diagnosing invasive melanoma [9]. Variability in diagnostic decisions is a serious problem and can cause substantial patient harm. To address this problem, computer-aided diagnostic (CAD) systems are being explored as a potential second reader to assist pathologists. These systems can leverage various techniques, including machine learning algorithms, to analyze digital whole slide images (WSIs) and provide insights to support pathologists' diagnoses [10–12].

In recent years, deep learning, a subfield of machine learning, has emerged as a powerful tool for image analysis. Deep learning models can learn complex patterns from large datasets, achieving impressive results in various computer vision tasks [13–17], but learning from gigapixel whole slide images (WSIs) remains a difficult problem due to their massive size. One of the most widely used natural image dataset, ImageNet [18] has $1,281,167$ training images, $50,000$ validation images and $100,000$ test images. The average size of ImageNet images is 482×418 pixels. Each image comes with clearly marked and easily comprehensible labels. Whole slide images, on the opposite, can be as large as $100,000 \times 100,000$ pixels.

One of the largest WSI datasets, a prostate core biopsy dataset of Campanella *et al.*[19] has 44,732 images, about 100× smaller than the ImageNet dataset. Moreoever, diagnostic labels for WSIs may have the following issue: 1) diagnostic labels may be only represented by small regions or a few cells in gigapixel WSIs, while other regions may correspond to other diagnostic categories; 2) variability in diagnostic decisions means labels contain noise; and 3) diagnostic labels may be represented by the entire composition of the slides and require an understanding of the interaction between different components (e.g., stroma, melanocytic nests, mitotic counts, and structures like ducts) [3, 20–23].

Due to the enormous size of WSIs, end-to-end learning has high computational complexity. Thus, WSI classification methods often follow a bag-of-words (BoW) model for learning representations, wherein a bigger patch of a whole slide image is treated as a bag while smaller image patches inside a bag are treated as words (or instances). Following this BoW model, many studies adopt a multiple instance learning-based (MIL) approach, which involves first extracting word-level feature representations and then applying global aggregation to bags of word-level representations to obtain WSI-level representation. These approaches are good at reducing the computational cost and solving the first issue of diagnostic labels (i.e. they are only representative in small regions). Now, the problem is reduced to discriminating word-level visual concepts (e.g. classification of cancerous vs. non-cancerous regions). However, diagnostic standards used by pathologists usually require modeling various complicated visual concepts that cannot be discerned by the MIL framework. For example, one of the key factors in skin lesion diagnosis, according to our expert pathologists, is melanocyte maturation, which refers to melanocytes becoming smaller and less pigmented with progressive descent into the dermis. To address this limitation and learn representations that capture such intricate relationships within WSIs, we need to achieve two goals: 1) identify clinically relevant entities, and 2) enable mid-to-long-range interactions between those entities.

My research addresses challenges in applying deep learning to whole slide image analysis for skin and breast cancer diagnosis. Limited labels, small datasets, and the need for efficient learning were key hurdles. We explored methods for overcoming these limitations, includ-

ing virtual staining for melanocyte detection, efficient WSI segmentation labeling, and self-attention networks for cancer classification. Notably, a novel framework (Semantics-Aware Attention Guidance) incorporates diagnostically relevant information into attention-based models, demonstrating promise for improving the accuracy and efficiency of future cancer diagnoses.

In this dissertation, projects are introduced to showcase the power of cutting-edge models in addressing different challenges presented by whole slide image analysis in the area of skin and breast cancer diagnosis.

Chapter 2 provides an overview of the three main datasets explored in this dissertation: a dataset of skin biopsies used for melanoma classification and two additional datasets for breast cancer diagnosis. By employing datasets with distinct characteristics and purposes related to different cancer types, this dissertation demonstrates the generalizability of the proposed deep learning methods across cancer diagnosis applications.

Chapter 3 describes two deep learning projects focused on segmenting diagnostically important structures within WSIs. The first project introduces VSGD-Net [24], a novel detection network that leverages virtual staining for melanocyte identification in skin cancer images. VSGD-Net learns by transferring knowledge from H&E stained images to virtual stains mimicking the appearance of Sox10, a specific biomarker for melanocytes. The second project, *Skin Biopsy Segmentation* [3], tackles the challenge of limited annotated training data in deep learning for skin cancer diagnosis. This project employs a two-stage segmentation approach that leverages coarse and sparse annotations on a small region of the WSI for training. Despite requiring minimal annotations, this method demonstrates promising performance on whole slide image segmentation tasks.

Chapter 4 summarizes two projects that leverage transformer networks for breast cancer diagnosis. The first project, `HATNet` [25], tackles breast cancer classification by directly

learning image representations from whole slide images using a holistic attention mechanism. This approach leverages a modified bag-of-words concept with self-attention to capture global information and identify clinically relevant structures within the tissue, all without explicit supervision. Building upon the success of `HATNet`, we developed `ScATNet` [4], which learns multi-scale representations directly from whole slide images, capturing both fine-grained details and broader contextual information.

Chapter 5. Expanding on our prior work in segmentation and transformer-based diagnosis models, Chapter 5 introduces a novel framework called Semantics-Aware Attention Guidance (`SAG`). `SAG` incorporates two key elements:

- Entity-to-Attention Conversion: This innovative technique transforms diagnostically relevant entities, like tissue anatomy and cancerous regions, into attention signals. These signals guide the model's focus on crucial areas within the WSI.

- Flexible Attention Loss Function: `SAG` employs a flexible attention loss function that efficiently integrates this semantically significant information. This approach optimizes the model's learning process by leveraging the inherent relationships between entities and the overall diagnosis task.

This innovative approach demonstrably improves accuracy, precision, and recall compared to state-of-the-art models on two independent cancer datasets. The content of this chapter is based on Liu and Wu *et al.*[26].

This dissertation draws upon my published work (Chapters 3 & 4) and under review work (Chapter 5). Each chapter commences with a review of relevant prior work, establishing a firm foundation and contextualizing my contributions within the existing research landscape. This context will showcase the existing research landscape and underscore the innovative contributions of my projects.

Chapter 2

DATASETS AND EVALUATION METRICS

2.1 Introduction

This chapter introduces the cornerstone of our investigation: the datasets utilized in this dissertation to develop deep learning models for whole slide image (WSI) analysis in cancer diagnosis. We focus on two prevalent cancer types: skin cancer and breast cancer. With the diversity in data sources, we aim to establish a foundation for exploring the generalizability of our approaches across various cancer diagnoses in later chapters. Understanding the inherent properties and challenges associated with each dataset is crucial for designing effective and adaptable deep learning models.

2.2 Skin Biopsy Dataset

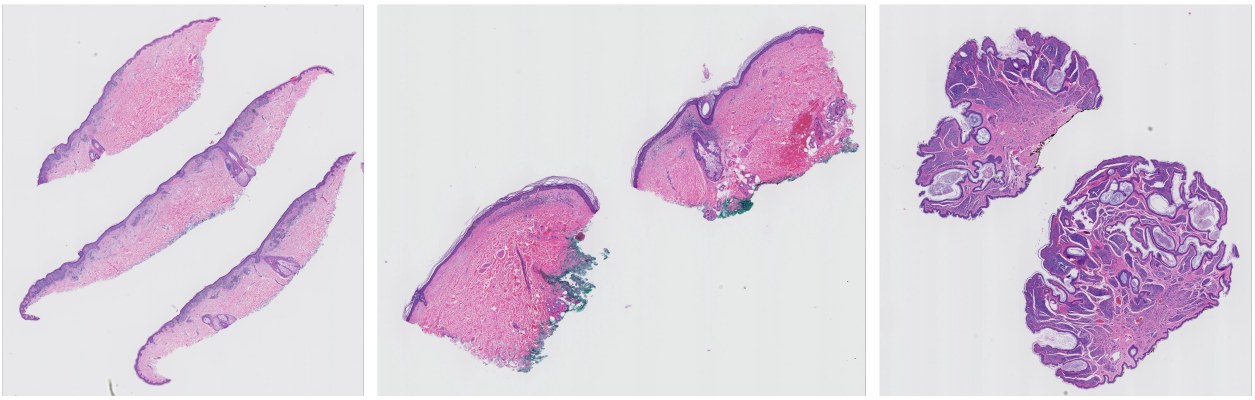


Figure 2.1: Three examples of WSIs in the MPATH Study

Diagnostic	Number of WSIs				Average WSI size
Category	Training	Validation	Test	Total	(in pixels)
MMD	26	6	29	61	11843 × 10315
MIS	25	5	30	60	9133 × 8501
pT1a	33	6	34	73	9490 × 7984
pT1b	18	6	22	46	14858 × 12154
Total	102	23	115	240	11130 × 9603

Table 2.1: Statistics of the skin biopsy whole slide image (WSI) dataset. The average WSI size is computed at a magnification factor of 10×. Diagnostic terms for the dataset used in this study are as follows: *mild and moderate dysplastic nevi* (**MMD**), *melanoma in situ* (**MIS**), *invasive melanoma stage pT1a* (**pT1a**), *invasive melanoma stage* $\geq$ *pT1b* (**pT1b**).

MPATH Dataset. The dataset used for melanoma diagnosis experiments was acquired as a part of the MPATH study (R01CA151306) and consists of 240 skin biopsy images (Figure 2.1) with hematoxylin and eosin (H&E) staining [9]. The MPATH study was approved by the Institutional Review Board at the University of Washington with protocol number STUDY00008506. These biopsy images were interpreted by a consensus panel of three experienced dermatopathologists using the modified Delphi approach [27]. The consensus panel assessments were grouped into five different MPATH-Dx (Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis) [28] simplified categories based on perceived risk for progression. 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. (IV) invasive melanoma stage T1a, and (V) invasive melanoma stage $\geq$ T1b.

These five classes were regrouped to four diagnostic classes for the classification task due to limited sample size in Classes I and II and because the clinical risk for progression of both

Class I and Class II is extremely low. The diagnostic terms we use for each class are as follows: 1) Class I-II: *mild and moderate dysplastic nevi* (**MMD**), which is very low risk to low risk, 2) Class III: *melanoma in situ* (**MIS**), which is higher risk than *MMD*, 3) Class IV: *invasive melanoma stage pT1a* (**pT1a**), which is higher risk for local/regional progression, and 4) Class V: *invasive melanoma stage* $\geq$ *pT1b* (**pT1b**), which is the greatest risk for regional and/or distant metastases. For diagnosis studies, we randomly split 240 WSIs into 102 training, 23 validation, and 115 test WSIs (see Table 2.1).

The consensus panel of three experienced dermatopathologists and an additional dermatopathologist on the MPATH research team marked a total 240 regions of interest (ROIs) that best defined the diagnostic classification of each case during the review process. These ROIs were annotated coarsely and sparsely by an expert pathologist (Mojgan Mokhtari) for training the semantic segmentation model described in Chapter 3. The tissue structures of skin biopsy WSIs used in this report are:

- *background* (*BG*): the pixels that do not contain any tissue.

- *epidermis* (*EP*): the outermost layer of the skin. It's the part you can see and touch. The epidermis provides a protective barrier for the body.

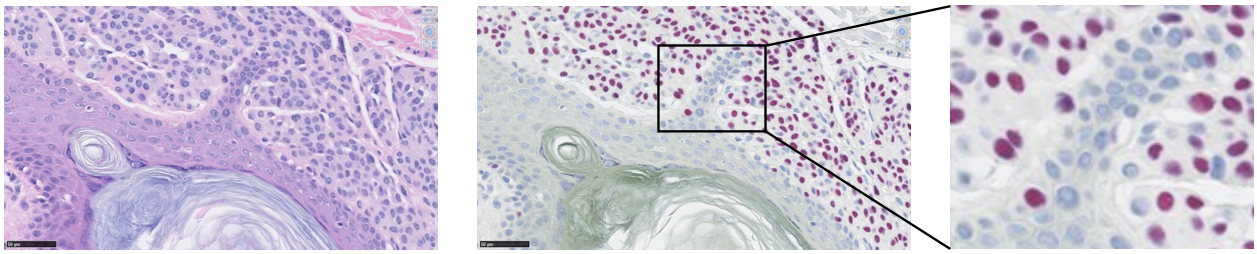
- *dermis* (*DM*): located just below the epidermis, the dermis is a thicker layer of skin. It contains blood vessels, nerves, hair follicles, and glands. This layer supports and strengthens the skin.

- *stratum corneum* (*COR*): the outermost part of the epidermis. It's made up of dead skin cells that have flattened and become tough to protect the underlying living cells.

- *epidermal melanocytic nest* (*EPN*): clusters of melanocytes (cells that produce pigment) found within the epidermis. They are often examined to check for skin abnormalities or conditions like moles or melanoma.

- *dermal melanocytic nest* (*DMN*): similar to the epidermal nests, these are clusters of melanocytes, but they are located within the dermis. These nests are also important for diagnosing skin conditions.

In summary, these describe different parts and features of the skin that are visible in the biopsy images. Understanding these structures helps in diagnosing and studying skin conditions.



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

Figure 2.2: 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.

Melanocyte Detection. In addition to the MPATH dataset, there is a supplementary melanocyte dataset, specifically collected for studying melanocyte detection. This dataset comprises 15 additional skin biopsy images, containing samples from 3 of the original 5 diagnostic categories present in MPATH [28]. Each biopsy was meticulously sectioned into 4-6 thin slices for magnified microscopic examination at 20x. This meticulous sectioning process resulted in a total of 75 image slices for analysis. Each WSI is stained with H&E first (see Figure 2.2a). Then they are carefully destained and re-stained with Sox10 (Figure 2.2(b)), a stain that differentiates melanocyte nuclei (red) from other cell nuclei (blue) (ground truth for melanocyte vs. non-melanocyte classification, (Figure 2.3(e)).

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

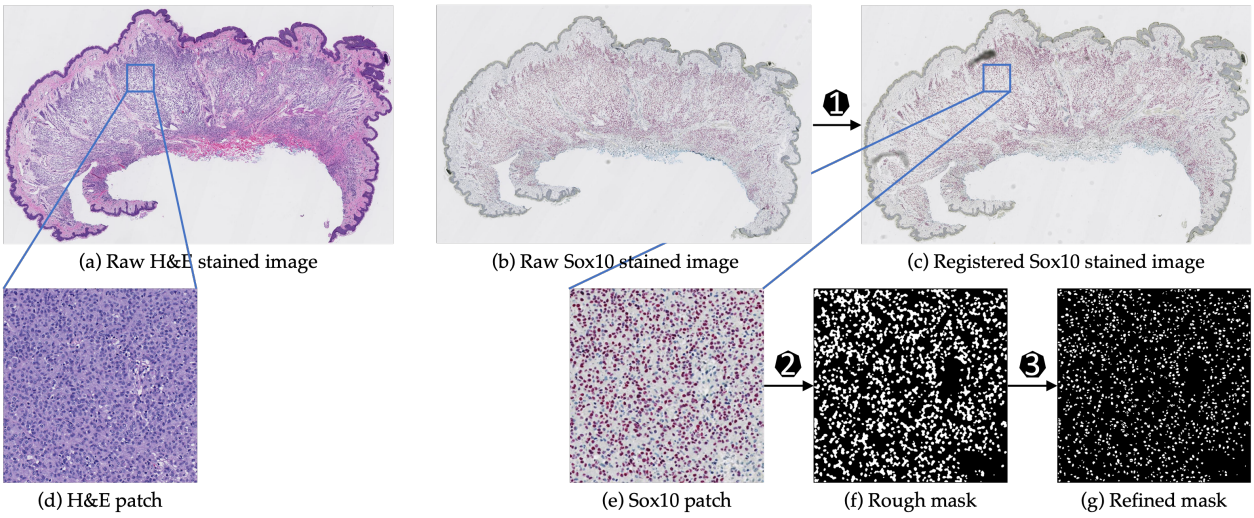
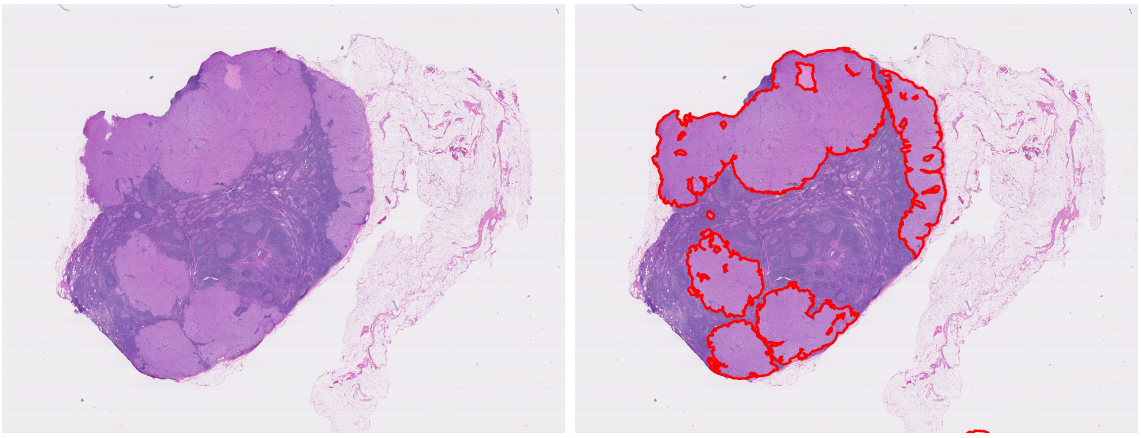


Figure 2.3: Preprocessing steps: Initially, raw Sox10 images (b) are aligned to template H&E images (a) using Histokat software[1][1], resulting in aligned Sox10 images (c). Subsequently, a Random Forest classifier distinguishes melanocytes from non-melanocyte pixels. Finally, the pretrained NuSeT [2] separates adjacent nuclei and refines the nuclear masks.

To generate ground truth labels for melanocyte detection, we developed a semi-automated procedure. A Random Forest classifier, trained on a set of 100 manually labeled melanocytes in Sox10 images, is used to create preliminary melanocyte masks. Next, a pre-trained nuclei detection model, NuSeT [2], is employed to refine these masks by separating clusters of touching nuclei. This two-step process effectively generates accurate melanocyte masks that serve as reliable ground truth labels for our study (Figure 2.3). For computational efficiency, while preserving image details, we cropped the registered-pair images into 256x256 patches at 10x magnification. Background patches were excluded, resulting in a usable dataset of 25,314 patches. To ensure a fair evaluation, we reserved 9652 paired image patches from 5 patients for the testing set. The remaining patches were used for training and validation, guaranteeing that data from the testing set was entirely separate from the training and validation data. Both the training and testing sets encompass the full spectrum of MPATH-Dx diagnostic



(a) An Example WSI (b) Example metastasis annotations

Figure 2.4: Examples from Camelyon16 dataset

classes.

2.3 Breast Biopsy Datasets

Two breast biopsy datasets are used in this dissertation — the Camelyon16 dataset [29] and the Digipath dataset [8].

Camelyon16 The Camelyon16 dataset [29] is a standard benchmarking dataset in the field of medical image analysis, especially in the area of cancer pathology. Originating from the Camelyon16 challenge, this dataset includes 400 whole-slide images (WSIs) of lymph node sections, which are annotated to show regions with metastatic cancer (see Figure 2.4). These annotations were meticulously provided by experienced pathologists, making the dataset a gold standard for training and testing algorithms aimed at detecting cancer metastases, specifically in breast cancer patients. There are two classes in this dataset: normal and tumor.

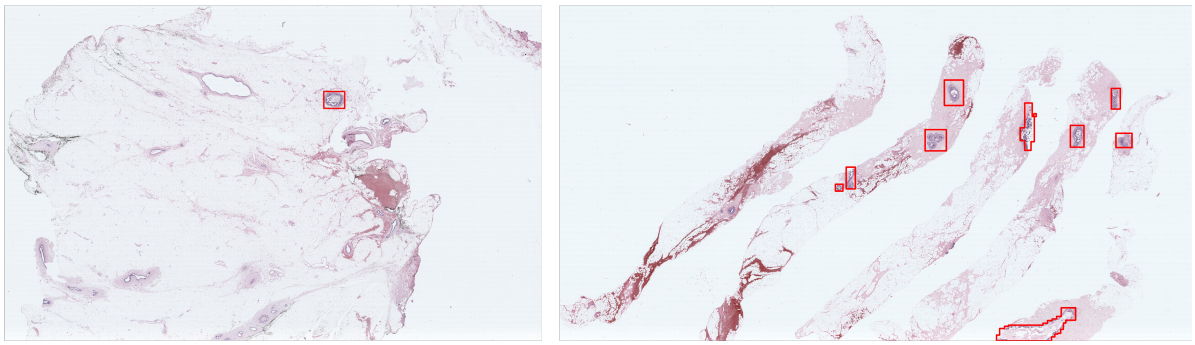


Figure 2.5: Two examples of WSIs in the Digipath dataset. Regions of interest (ROIs) that helped pathologists in diagnosis are shown in red boxes.

Digipath This dataset consists of 240 whole slide images stained with hematoxylin and eosin (H&E) [8]. Three expert pathologists independently reviewed these cases, followed by a consensus meeting using a modified Delphi method to establish a reference consensus label for each slide [30]. The pathologists' assessments were categorized into four groups: (1) benign without atypia, (2) atypia, (3) ductal carcinoma in situ (DCIS), and (4) invasive carcinoma. These consensus labels serve as our ground truth diagnoses.

	Benign	**ADH**	**DCIS**	**INV**	**Total**
Slide					
Training set	34	35	41	10	120
Test set	22	48	38	12	120
ROI					
Training set	60	58	85	17	220
Test set	37	81	80	19	217

Table 2.2: Class distribution of slides and ROIs in the Digipath dataset.

The expert pathologists also identified 422 regions of interest (ROIs), that best supported their diagnoses (Figure 2.5. Consistent with previous studies utilizing this dataset for computer-aided second opinion systems [31–35], we used these ROIs to train and evaluate our method. Following previous studies on this dataset that aims to build directed computer-aided second opinion systems [31, 35–37], the dataset was randomly split into 164 training ROIs, 42 validation ROIs, and 216 test ROIs (See Table 2.2). It's important to note that clinically, each slide can contain multiple ROIs. To prevent information leakage, we ensured all ROIs from a single slide were assigned to the same set (either training and validation or test set).

2.4 Evaluation Metrics

In evaluating the performance of classification models, outcome metrics provide insights into different aspects of the model's accuracy and reliability. Below are the definitions and formulas for the key metrics relevant to this thesis:

- Classification (or Top-1) accuracy counts the number of times the predicted label is the same as the ground truth label and is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

  where TP, FP, TN, and FN denote the true positives, false positives, true negatives, and false negatives, respectively.

- F1-score is a harmonic mean of precision P and recall R and is defined as:

$$\text{F1-score} = \frac{2PR}{P + R}$$

  where $P = \frac{\text{TP}}{\text{TP} + \text{FP}}$ and $R = \frac{\text{TP}}{\text{TP} + \text{FN}}$.

- Sensitivity (also known as Recall) measures the proportion of the positive cases that are correctly classified and is defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Specificity measures the proportion of the negative cases that are correctly classified and is defined as:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- Area under the receiver operating characteristics curve (ROC-AUC) is a graph obtained by varying the threshold for diagnostic decision, illustrating the discrimination ability of the classifier. We use a One-vs-Rest scheme, which computes the AUC of each class against the rest [38].

- Precision-Recall (P-R) curve is a plot that illustrates the trade-off between precision (positive predictive value) and recall (sensitivity) for different threshold values. The curve helps understand a model's ability to balance precision and recall.

- Jaccard Index (also known as Intersection over Union) measures the similarity between two sets and is defined as:

$$\text{Jaccard Index} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

  It quantifies the overlap between the predicted positives and the actual positives.

- Peak Signal-to-Noise Ratio (PSNR): A metric used to quantify the quality of a reconstructed image compared to a noise-free original. It is expressed in decibels (dB) and calculated as:

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right), \tag{2.1}$$

  where MAX_I is the maximum possible pixel value of the image (e.g., 255 for an 8-bit image) and MSE is the Mean Squared Error between the original and reconstructed images. Higher PSNR indicates a better reconstruction/image generation quality, with minimal noise introduced in the process.

Chapter 3

SEMANTIC SEGMENTATION OF DIAGNOSTIC ENTITIES IN WHOLE SLIDE IMAGES

3.1 Introduction

This chapter explores the integration of deep learning technologies to enhance diagnostic precision. It emphasizes identifying key structures such as the epidermis, dermis, melanocytes, and nuclei within skin biopsy specimens and discusses how advanced computational techniques, such as deep neural networks, can address the challenges inherent in traditional diagnostic methods.

Nuclei Detection Recent advances in instance segmentation have significantly improved nuclei detection by allowing for more precise and accurate identification of individual nuclei within complex tissue samples. Techniques like Mask R-CNN [14] and StarDist [39], which employs star-convex polygons for precise localization, are complemented by Hover-Net [40] which introduces a "Hover" branch for segmenting connected nuclei. Additionally, CHR-Net [41] enhances nuclei mapping through a two-stage process. However, these methods often rely on extracting relevant features directly from the H&E stained image, which can be challenging due to limitations in stain variations and overlapping structures.

Image-to-Image Translation Image-to-image translation techniques such as Pix2Pix [42] and CycleGAN [43] have revolutionized stain normalization and virtual staining tasks in histopathology. Despite their potential, these methods often underutilize informative features critical for tasks like nuclei detection. VSGDNet addresses these limitations by simultaneously optimizing image synthesis and nuclei detection, thereby enhancing both processes.

Semantic and instance segmentation The challenge of accurately identifying and differentiating between tissue layers such as the dermis and epidermis, as well as detecting abnormal structures like dermal nests, is critical for diagnosing melanoma. Semantic segmentation is an invaluable tool in this context, though its effectiveness is limited by the availability of extensive, pixel-level annotations. While instance segmentation techniques like Mask R-CNN offer more granular object delineation, they often require a significant amount of annotated data for each individual instance within the image. In scenarios with limited annotations, as in this project, simpler semantic segmentation models can be more effective for learning robust representations from scarce data. Additionally, semantic segmentation methods can be computationally more efficient, making them a more practical choice for real-world applications with large WSIs.

Projects Two methods were proposed to address the aforementioned challenges. The first project, `VSGDNet` [24], revolutionizes melanocyte detection by employing virtual staining techniques, which facilitate the identification of specific biomarkers without the need for additional physical staining processes. By transferring knowledge from standard H&E stained images to virtual stains that mimic the appearance of Sox10, `VSGDNet` offers a novel approach to overcoming the limitations of traditional staining methods in dermatopathology.

The second project, *Skin Biopsy Segmentation* [3], addresses the challenge posed by the scarcity of annotated training data, which is a common bottleneck in the application of deep learning models. By implementing a two-stage segmentation strategy that utilizes coarse and sparse annotations for training, this method demonstrates significant potential in improving the segmentation accuracy of WSIs with minimal annotations.

Together these two projects highlight the innovative integration of deep learning techniques in skin cancer diagnosis, illustrating their potential to transform traditional practices by enhancing diagnostic precision and reducing reliance on extensive annotated datasets.

3.2 *VSGD-Net: Virtual Staining Guided Melanocyte Detection on Histopathological Images*

This section introduces VSGDNet, a state-of-the-art virtual-staining-guided detection architecture designed to enhance the accuracy of melanocyte detection in skin biopsy images. The development and effectiveness of VSGDNet, detailed in the work of Liu, Li, and **Wu** et al. [24], leverage both H&E and Sox10 stains to integrate and analyze complex diagnostic features (Figure 3.1).

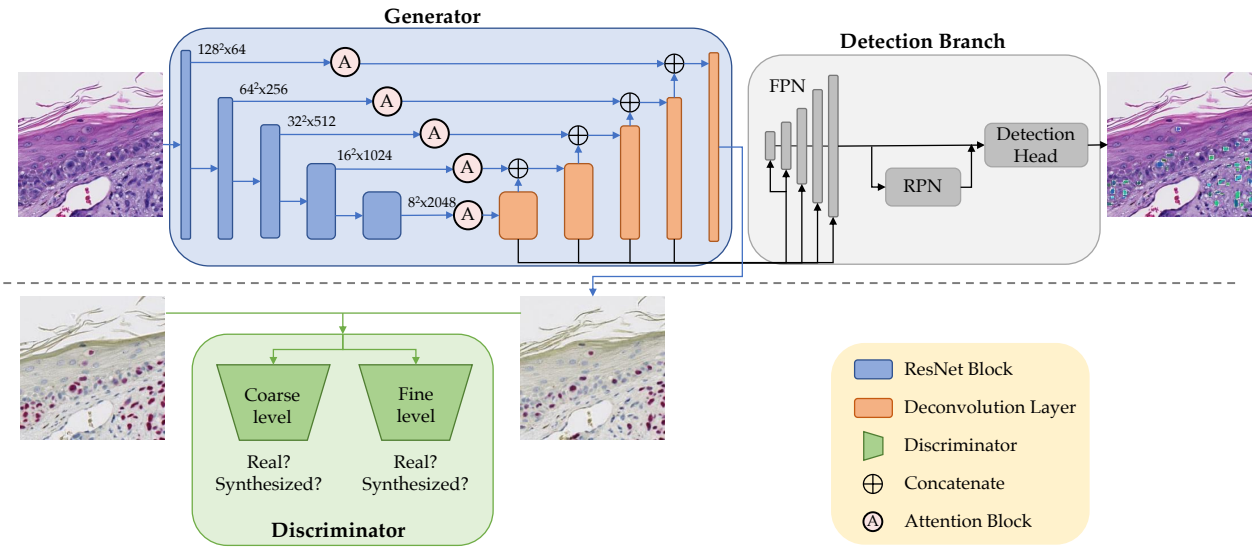


Figure 3.1: `VSGDNet` framework: H&E images are virtually stained to Sox10. The jointly trained detection branch utilizes the intermediate features in the generator to detect melanocytes and provides feedback to the generator to enhance synthesis quality. The inference phase only uses the upper part of the architecture.

3.2.1 `VSGDNet`

At its core, VSGDNet incorporates a U-Net [44] architecture with a ResNet-50 [13] encoder to extract features efficiently. It also includes attention mechanisms [45] within the decoder

to focus specifically on melanocytes during the virtual staining process. The realism of the generated Sox10 images is further refined by a multi-scale discriminator, inspired by Pix2PixHD [46], which ensures the generator (G) produces high-fidelity outputs.

The detection component of VSGDNet mirrors the structure of Mask R-CNN [14], incorporating a Feature Pyramid Network (FPN), a Region Proposal Network (RPN), and specialized detection heads. These elements are strategically embedded within the decoder to utilize the enriched Sox10-aligned features, optimizing the system for precise melanocyte identification. This strategic placement has been substantiated by extensive ablation studies, indicating a higher correlation and effectiveness in utilizing Sox10 features for melanocyte detection [24].

3.2.2 Results and Discussion

Table 3.1: Comparison with nuclei detection methods.

Method	**P**	**R**	F_1	**Jaccard**
RLS [47]	0.443	0.570	0.499	0.332
Nuclei Classification	0.693	0.506	0.585	0.413
Mask R-CNN [14]	0.735	0.514	0.605	0.434
U-Net [44]	0.630	0.639	0.635	0.465
StarDist [48]	**0.745**	0.426	0.542	0.372
HoverNet [40]	0.729	0.499	0.592	0.421
CHR-Net [49]	0.607	0.688	0.645	0.476
Ours	0.660	**0.710**	**0.684**	**0.520**

We evaluated VSGDNet on the melanocyte dataset (described in Chapter 2.2). Melanocyte detection was assessed quantitatively using precision, recall, F1-score, and Jaccard index (def-

Table 3.2: Comparison with GAN-based methods.

Method	**P**	**R**	F_1	**Jaccard**
StainGAN [50]	0.476	0.299	0.367	0.225
PC-StainGAN [51]	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**

initions in Chapter 2.4). Sox10 image quality was evaluated with both Peak Signal-to-Noise Ratio (PSNR), as defined in Chapter 2.4), and qualitative analysis.

Melanocyte Detection Our results in Table 3.1 shows that `VSGDNet` outperformed specialized detection methods (RLS [47], Mask R-CNN [14], U-Net [44], StarDist [48], HoverNet [40], CHRNet [49]). As shown in Table 3.2, it also achieved a higher F_1 -score and Jaccard index compared to GAN-based segmentation methods (StainGAN [50], PC-StainGAN [51], our self-implemented model). While StarDist and the GAN segmentation method showed high precision and recall, respectively, their F_1 and Jaccard scores suffered due to imbalanced prediction (over-prediction or under-prediction).

Image Synthesis Though secondary to detection, we evaluated the quality of synthesized Sox10 images to demonstrate the effectiveness of shared features. VSGDNet achieved the highest PSNR (indicating the most reliable virtual staining) and comparable SSIM to PC-StainGAN (measuring structural similarity). Qualitative results (Figure 3.2) showcase high-fidelity virtual staining by VSGDNet.

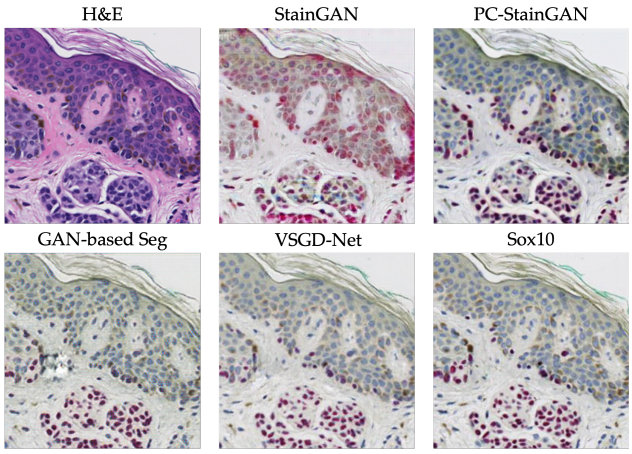


Figure 3.2: The qualitative comparison of `VSGDNet` with other virtual staining methods

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

In this section, we present a novel two-step segmentation pipeline. Our approach leverages coarse and sparse annotations, allowing for the segmentation of larger anatomical structures (like epidermis and dermis) and smaller entities (like melanocytes) within the WSI separately (Figure 3.3). By reducing the annotation burden, our method paves the way for more efficient development of deep learning-based tools for skin cancer diagnosis. The content of this chapter is based on the work of Nofallah, Mokhtari and **Wu** *et al.*[3].

3.3.1 Related Work

Limited data annotation poses a significant challenge in medical image segmentation. To mitigate this, various strategies such as data augmentation (e.g., image sharpening, mixup) have been utilized to enhance segmentation [52]. Although active learning shows potential, it requires a base model and extensive annotation efforts, which are often impractical with sparse data [53]. Adjustments to loss functions, like class-balancing techniques, have also proven helpful in contexts with limited annotations [53].

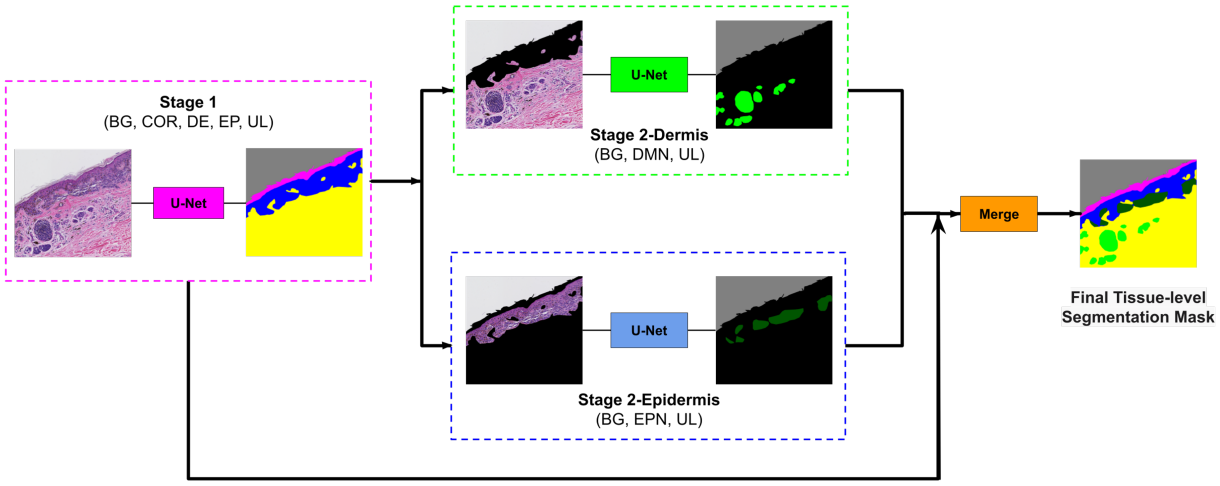


Figure 3.3: Overview of our approach. The image first goes to stage 1, and the segmentation mask of entities (COR, stratum corneum; EP, epidermis; DE, dermis; BG, background; and UL, unlabeled) in stage 1 is generated. Then this mask is used to remove the epidermis from stage 2-Dermis input and remove the dermis from stage 2-Epidermis input. The modifed images are fed to their corresponding trained model. Stage 2-Dermis generates the segmentation masks of entities present in the dermis (DMN, dermal nests), and stage 2-Epidermis generates the entities in the epidermis (EPN, epidermal nests). In the end, stage 2-Dermis and stage 2-Epidermis segmentation masks are overlaid on the stage 1 mask, and the final tissue-level segmentation mask is generated.

While domain adaptation and external data usage have advanced segmentation in other fields, their application to skin biopsy images is hindered by the scarcity of large, labeled datasets and notable morphological diversity [54]. Research in skin biopsy segmentation has generally focused on individual structures such as the epidermis or dermis, with semantic segmentation of a broader range of structures still largely unexplored, particularly in datasets with limited and imperfect annotations.

Table 3.3: Evaluation of the segmentation model on ROI testing set

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

3.3.2 Method

Our approach allows training a convolutional neural network (CNN) on images with coarse and sparse annotations to segment multiple clinically important structures within whole slide images (WSIs). The system uses a modified U-Net architecture for segmenting skin biopsy images in a two-stage approach (See Figure 3.3). The first stage utilizes a U-Net pre-trained on ImageNet to segment large anatomical structures (background, epidermis, dermis). The second stage tackles smaller entities (e.g., melanocytic nests) by focusing separately on the dermis and epidermis regions, using ground truth labels specific to each area. This approach addresses the challenge of sparse annotations for smaller structures by training separate models for different size ranges, aiming to improve segmentation accuracy for all entities of interest in skin biopsies.

3.3.3 Results and Discussion

During testing, the Stage 1 segmentation mask (all tissues) from the model was used to extract dermis and epidermis regions for the Stage 2 branches (focusing on DMN and EPN, respectively). These entities were then overlaid on the Stage 1 mask to create the final segmentation (Figure 3.3). Quantitative results are shown in Table 3.3. Figure 3.4 shows example ROIs with annotations and the corresponding model predictions.

The final goal was to generate segmentation masks for whole slide images (WSIs) using

the model trained on ROIs with coarse annotations. We adopted the validation pipeline from Figure 3.3. WSIs were first divided into slices to reduce image size and eliminate the effect of slide orientation. These slices underwent preprocessing similar to ROIs and were fed into the model. The resulting segmentation masks were merged to create the final WSI segmentation mask.

While WSI segmentation masks (Figure 3.5) suggest promising results using sparse and coarse annotations, limitations were identified. Noise in the training data, such as inaccurate borders and missing labels for similar structures (e.g., inflammatory cells), led to over-labeling of nests, particularly epidermal nests (reflecting the noisier annotations for them). Accurate diagnosis requires both high sensitivity (finding all nests) and high specificity (reducing false positives). Therefore, mitigating noise in sparse annotations is crucial. This can be achieved by having a separate pathologist review the ground truth data.

3.4 Summary

This chapter highlights two innovative projects designed to segment diagnostically important entities within whole slide images (WSIs), each addressing specific challenges in dermatopathology through advanced deep learning techniques:

1. `VSGDNet` [24] introduces an innovative virtual staining technique. By simulating the appearance of specific biomarkers, such as Sox10, it overcomes the limitations of traditional staining methods. This approach not only enhances the detection accuracy but also reduces the reliance on physical staining processes, which are resource-intensive and time-consuming.

2. **Skin Biopsy Segmentation** [3]: this project tackles the challenge of limited annotated data availability, employing a two-stage segmentation approach that effectively utilizes sparse annotations to improve the accuracy of whole slide image (WSI) segmentation.

Together, these projects demonstrate how modern computation methods can reduce dependency on extensive annotated data sets. The structures segmented by these projects will later provide crucial cues for the diagnosis network (Chapter 5).

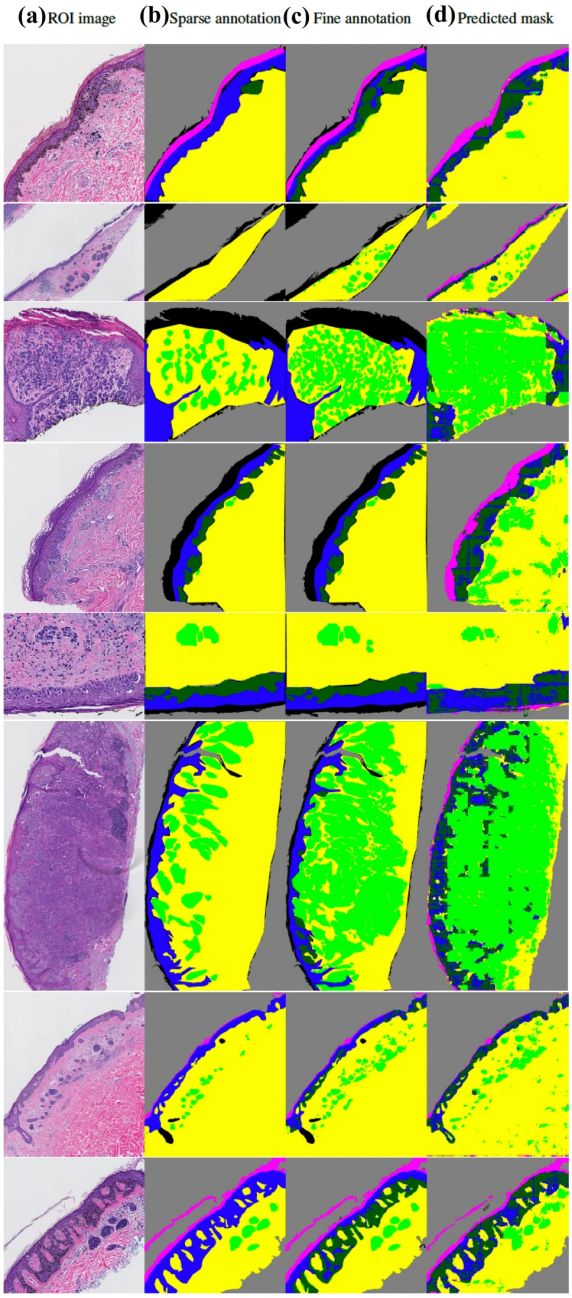
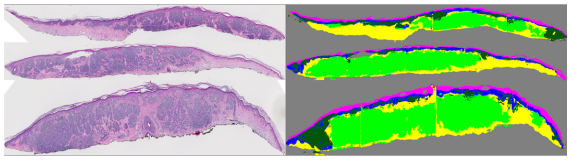
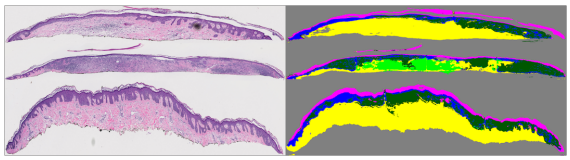


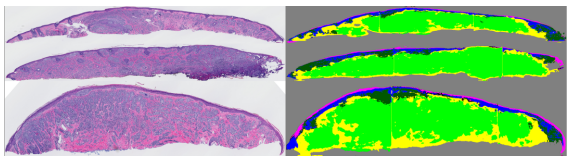
Figure 3.4: Visualization of results by our published two-step segmentation method [3]. (a) Examples of original image; (b) sparse expert annotation; (c) fine-detailed epidermal and dermal nest annotation by expert for evaluation; (d) segmentation results by our published work. The annotation and segmentation images contain the *dermis* (*DM*-yellow), *epidermis* (*EP*-blue), *stratum corneum* (*COR*-pink), *background* (*BG*-gray), *dermal nests* (*DMN*-light green), and *epidermal nests* (*EPN*-dark green).



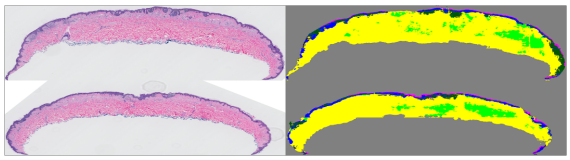
(a) DMN: High sensitivity, high specificity, EPN: High sensitivity, low specificity)



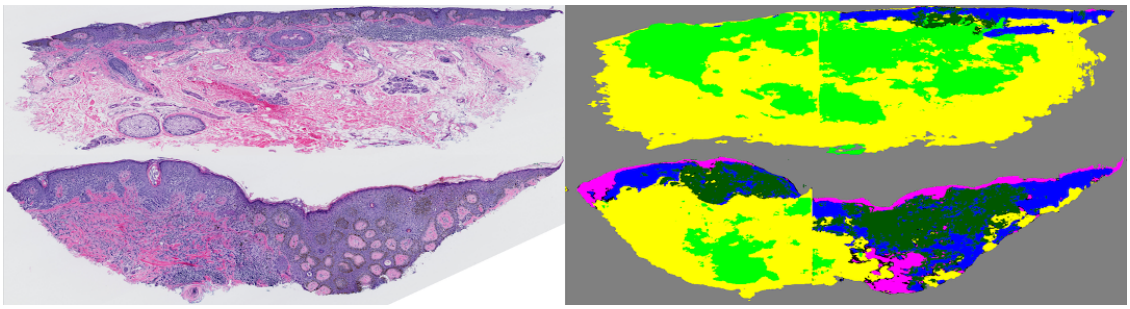
(b) DMN: High sensitivity, medium specificity, EPN: High sensitivity, low specificity



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



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



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

Figure 3.5: Examples of original WSI (left) and its corresponding segmentation mask(right). Slices of each WSI are extracted and concatenated vertically. The segmentation images contain Dermis (DE-yellow), Epidermis (EP-blue), Stratum Corneum (COR-pink), Background (BGgray), Dermal Nests (DMN-light green), and Epidermal Nests (EPN-dark green). The model has been trained on coarse and sparse annotations. The captions show the dermatopathologists' qualitative grading on each WSI segmentation mask for dermal nests (DMN) and epidermal nests (EPN).

Chapter 4

TRANSFORMER NETWORK FOR DIAGNOSIS

4.1 Introduction

While Chapter 3 focuses on the application of deep learning to detect diagnostically important structures (e.g. tissues and cellular entities), Chapter 4 shifts the focus to diagnosis based on the entire whole slide image, where the complexities of diagnosis require not only high precision but also a nuanced understanding of diverse histopathological features. In this chapter, we explore the deployment of transformer networks, renowned for their capability to handle sequential data in natural language processing, to the domain of medical imaging [55]. This transition is motivated by the need to develop more sophisticated models that can learn from the vast and intricate data presented by whole slide images without the constraints of traditional supervised learning frameworks.

The first project detailed in this chapter, `HATNet` [25], introduces a groundbreaking approach to breast cancer classification. It utilizes a holistic attention mechanism that allows for the direct learning of image representations from WSIs. This method incorporates a modified bag-of-words model combined with self-attention to effectively capture and utilize global contextual information from the images, highlighting clinically significant structures without the need for explicit annotation.

Building on the foundations laid by `HATNet`, the subsequent project, `ScATNet` [4], extends these concepts to include the learning of multi-scale representations. By doing so, `ScATNet` is not only able to discern fine-grained details critical for accurate diagnosis but also integrates broader contextual understanding, thereby enhancing the model's diagnostic capabilities across varying levels of detail and scale.

The advancements in transformer technology presented in this chapter signify a pivotal

development in medical imaging, offering the potential to greatly enhance the accuracy and efficiency of cancer diagnosis processes. This chapter aims to demonstrate how transformer-based models can be innovatively applied to the challenges of medical image analysis, particularly in handling the scale and complexity of WSIs.

4.2 End-to-End diagnosis of breast biopsy images with transformers (`HATNet`*)*

The content of this section is based on the work of Mehta, Lu and **Wu** *et al.*[25].

4.2.1 Introduction

Breast cancer is the most common non-skin cancer among women, accounting for approximately 25% of all cancer cases worldwide. Diagnosis largely relies on the subjective visual evaluation of breast biopsy specimens by pathologists, leading to significant variability and potential misdiagnoses with serious clinical repercussions [5–8, 56–58].

Addressing these challenges, this section describes the Holistic ATtention Network (`HATNet`), a self-attention-based neural network that enhances breast biopsy image classification. Inspired by Vaswani *et al.* [55], `HATNet` uses a modified bag-of-words model and self-attention to decompose and analyze biopsy images, enabling the integration of global contextual information for accurate diagnosis. Empirical results show that `HATNet` improves diagnostic accuracy by 8% and doubles the processing speed compared to Y-Net [33], effectively focusing on crucial diagnostic features like stromal tissue and ducts.

4.2.2 `HATNet`: Holistic Attention Network

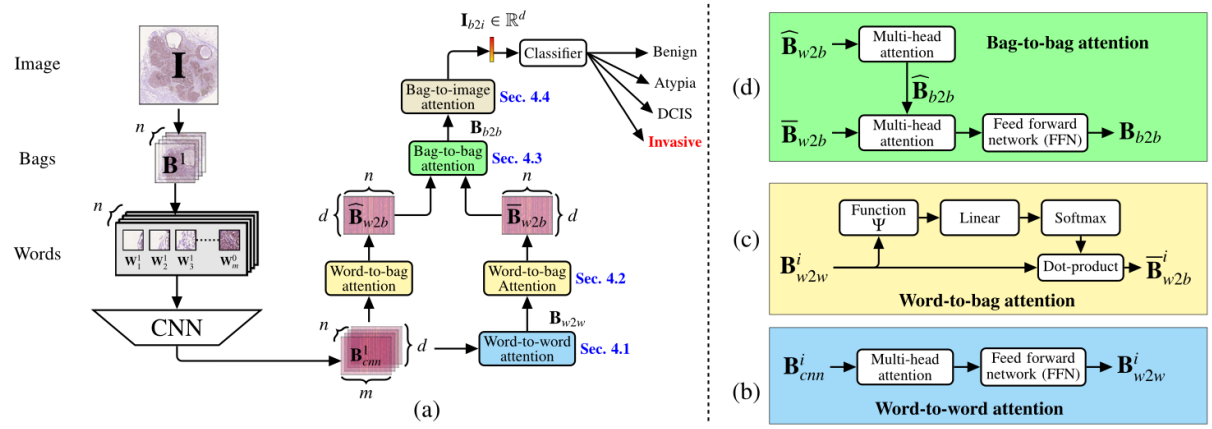


Figure 4.1: (a) `HATNet`: Our end-to-end holistic attention network for classifying breast biopsy images models the relationships between bags and words in a hierarchical manner using self-attention. (b-d) Word-to-word, word-to-bag, and bag-to-bag attention modules are visualized; they allow the learning of relationships between bags and words using a bottom-up method. Note that the word-to-bag attention module for processing B_{cnn} and the bag-to-image attention module for processing B_{b2b} are similar to (c) and therefore, we do not visualize them here.

`HATNet`, depicted in Figure 4.1, enhances the transformer architecture by incorporating a bag-of-words approach. The architecture of `HATNet` strategically refines data from the granular level of words to bags and eventually to the entire image to finalize the classification decision. Initially, `HATNet` processes inter-word relationships through self-attention, amalgamating these into coherent bag-level data points. Subsequent layers of attention synthesize these data points into a unified image-level representation, which is then utilized to determine the diagnostic category.

The progressive, bottom-up approach—from words to bags to the complete image—enables `HATNet` to capture and express detailed and clinically significant patterns within the WSIs

Row No.	Model	Parameters			Evaluation metrics				
		CNN	Attn.		Accuracy	F1-score	Sensitivity	Specificity	ROC-AUC
R1	Pathologists (avg. of 87 practicing pathologists)			0.70	0.71	0.70	0.90		
R2†	LAB & LBP hand-crafted features (w/o saliency)			0.28					
R3†	LAB & LBP hand-crafted features (w/ saliency)			0.45					
R4†	Bag-of-word (majority voting w/o saliency)			0.23					
R5†	Bag-of-word (majority voting w/ saliency)			0.55					
R6†	Bag-of-word (learned fusion w/o saliency)			0.38					
R7†	Bag-of-word (learned fusion w/ saliency)			0.55					
R8	MRSegNet with histogram and co-occurence features	26.03 M	NA		0.55	0.56	0.55	0.85	
R9	MRSegNet with structural features	26.03 M	NA		0.56	0.57	0.56	0.85	
R10	Y-Net	3.91 M	NA		0.62	0.62	0.62	0.87	
R11	HATNet (w/ ESPNetv2)	2.21 M	2.37 M		0.67	0.64	0.67	0.89	0.89
R12	HATNet (w/ MobileNetv2)	2.22 M	2.37 M		0.66	0.65	0.66	0.89	0.88
R13	HATNet (w/ MNASNet)	3.10 M	2.37 M		**0.70**	**0.70**	**0.70**	**0.90**	**0.90**
R14	HATNet (Ensemble)	NA	NA		**0.71**	**0.70**	**0.71**	**0.90**	**0.90**

Figure 4.2: Comparison with state-of-the-art networks. `HATNet` outperformed existing methods by a significant margin. Network parameters are reported for single models only. We use majority voting for ensembling the models. These works split the dataset (240 slides) into training (180 slides) and validation (60 slides) sets, and reports the performance on validation set. For completeness, we only report the accuracy of these methods. Note that the performance of networks in R8-R14 is on the same independent test set of 119 slides.

effectively. This capability not only enhances diagnostic accuracy but also facilitates the development of tools that can automatically annotate and explain decisions based on clinically relevant features within the images.

Dataset The main dataset used in this study is the Digipath dataset, described in Chapter 2.3. Regions of interests (ROIs) are randomly split into 164 training, 42 validation, and 216 test ROIs.

4.2.3 Main Results

Performance Comparison In a comprehensive evaluation, `HATNet` outstripped existing models across several key performance metrics. Notably, it exhibited an 8% increase in F1-score over the previously leading segmentation-based method and, when combined with ensemble techniques, it further improved accuracy and sensitivity by an additional 1%. The

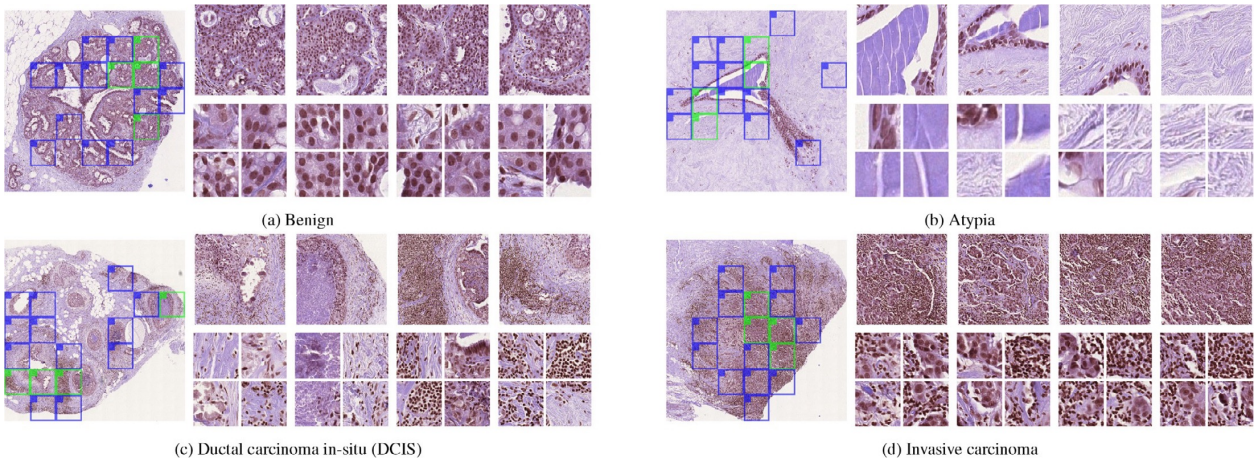


Figure 4.3: Example results of bags and words identified using `HATNet` across different diagnostic categories. `HATNet` aggregates information from different parts of the image and different textures. Here, each sub-figure of the breast biopsy image is shown on the left of each panel with the top-30% bags (top-4 in green, the rest in blue) identified using `HATNet` overlayed on the image. The upper right in each panel shows the top-4 bags, and the bottom right in each panel shows the top-4 words in each bag.

comparative analysis detailed in Table 4.2 underscores `HATNet`'s superior diagnostic accuracy and robustness, positioning it as a significant advancement in histopathological image analysis.

Saliency and Annotation Concordance One of the standout features of `HATNet` is its ability to identify and focus on clinically significant regions within histopathological slides. The model's attention mechanism is adept at highlighting areas like ductal and stromal tissues, which are crucial for accurate breast cancer diagnosis. Quantitative analysis using dice scores shows high agreement between `HATNet`'s saliency maps and expert annotations, reinforcing the model's relevance and applicability in clinical settings.

4.2.4 Summary

The results presented in this section not only validate the effectiveness of `HATNet` but also highlight its potential to transform histopathological diagnosis. By automating complex diagnostic tasks with high accuracy and speed, `HATNet` could significantly reduce the cognitive load on pathologists and decrease the likelihood of diagnostic errors. Furthermore, its ability to generate interpretable saliency maps aligns well with the increasing demand for transparency in AI-driven medical applications.

4.3 Scale-Aware Transformers for Diagnosing Melanocytic Lesions (`ScATNet`)

4.3.1 Introduction

Invasive melanoma continues to pose significant diagnostic challenges within the field of oncology, with over 100,000 new cases estimated in 2021 alone [59]. Despite advancements in diagnostic techniques, the interpretation of skin biopsy specimens remains prone to significant variability, leading to potential misdiagnoses and patient harm, especially in ambiguous cases [5–9]. The need for enhanced diagnostic accuracy and consistency is imperative.

Building on the foundational success of `HATNet` (described in Chapter 4.2), which introduced holistic attention mechanisms for breast cancer classification, r 4.2), which leverages self-attention for breast cancer classification, we introduces the Scale-Aware Transformer Network (`ScATNet`) [4]. `ScATNet` extends the innovative principles of transformer technology to address the unique challenges of melanoma diagnosis in whole slide images (WSIs). Unlike `HATNet`, which focuses on a single-scale analysis, `ScATNet` is engineered to interpret and integrate multi-scale histopathological features effectively, recognizing that critical diagnostic information can vary significantly across different magnifications and tissue structures.

Illustrated in Figure 4.4, `ScATNet` uniquely combines a convolutional neural network (CNN) with transformer technology to analyze patch-wise representations independently at multiple scales. This approach allows the model to capture a broad spectrum of diagnostic features, from very fine details to overarching tissue patterns. As a result, `ScATNet` offers im-

proved precision in discriminating between benign and malignant lesions, addressing complex diagnostic challenges and capturing intricate details critical for accurate medical analysis.

Furthermore, `ScATNet` introduces a soft-label assignment technique to alleviate the challenges associated with ambiguous tissue slice classifications within a single WSI, thereby refining the diagnostic process. This approach not only clarifies the diagnostic pathway but also optimizes the performance by focusing on the most informative tissue slices, reducing the ambiguity inherent in complex cases.

The effectiveness of `ScATNet` is demonstrated through rigorous testing against current state-of-the-art methods, where it shows significant improvements in diagnostic accuracy, outperforming existing models by substantial margins [60, 61]. These results underscore `ScATNet` 's potential as a powerful tool for pathologists, offering a level of precision comparable to expert human analysis.

The primary contributions of this work are threefold: the development of a multi-scale, self-attention-based framework for classifying WSIs (`ScATNet`), a novel soft-label assignment method to enhance diagnostic accuracy, and a comprehensive evaluation that highlights `ScATNet` 's superior performance in clinical settings (Sections 4.3.3, 4.3.4, and 4.3.7).

4.3.2 Related Work

`ScATNet` builds on the foundational successes of various methodologies within the realm of whole slide image (WSI) classification and transformer technologies. These approaches are discussed briefly below.

Multiple instance learning (MIL) Unlike images in standard datasets (e.g., ImageNet [62]), WSIs are orders of magnitude larger and cannot be processed in an end-to-end fashion using CNNs. The MIL framework has been widely studied for classifying different types of WSIs, such as lung [63], kidney [64], and breast [31]. In short, the input WSI is divided into instances (or patches) and the same classification label is assigned to all instances during training. During evaluation, methods such as averaging and majority voting are used

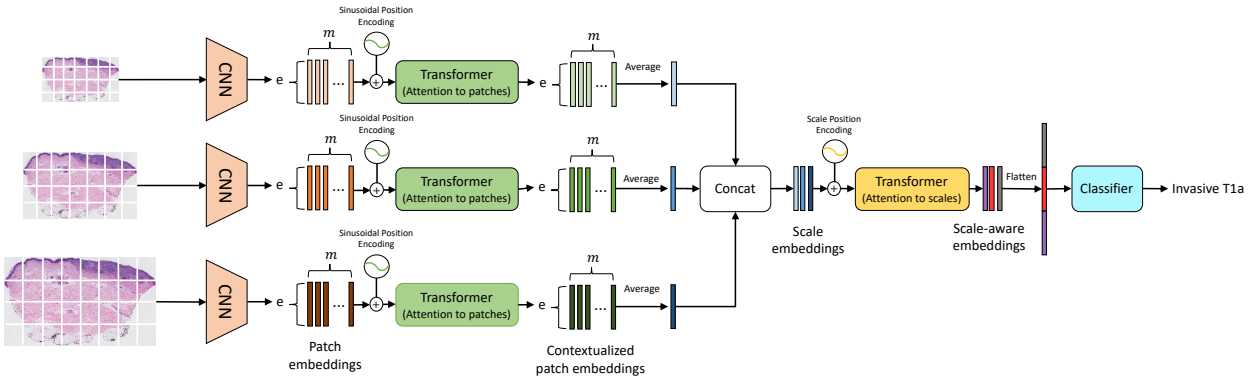


Figure 4.4: Overview of `ScATNet` for classifying skin biopsy images. To learn representations from these large WSIs at multiple input scales in an end-to-end fashion, `ScATNet` factorizes the classification pipeline into three steps. The first step involves learning local patch-wise embeddings using an off-the-shelf CNN for each input scale independently. In the second step, `ScATNet` learns inter-patch representations using transformers and produces contextualized patch embeddings for each input scale. In the last step, `ScATNet` learns inter-scale representations from concatenated multi-scale contextualized patch embeddings using another transformer network and produces scale-aware embeddings, which are then classified linearly into diagnostic categories.

to aggregate the information from all instances in an image and produce an image-level classification label. Though these approaches are effective, they learn local instance-wise representations and do not allow global or long-range feature interactions. `ScATNet` extends the MIL framework with the transformers of Vaswani *et al.*[55] to learn global representations in an end-to-end fashion.

Patch-based Feature Aggregation Patch-based methods provide a solution to the gigapixel size of WSIs, while only requiring slide-level labels. However, learning robust instance representations is challenging due to the ambiguity in instance-level labels. To address this, many recent methods [37, 63] adopt a two-step approach that consists of (1) training an

instance encoder for obtaining a prediction score or low-dimensional features, and (2) learning a model that aggregates the features extracted by the learned instance encoder to form instance-level information for slide-level prediction. Although this approach has had some success, it often suffers from worse performance when noisy labels are present, causing the features to not be representative of their given labels.

Segmentation-based methods. These approaches use semantic information about tissues in a WSI to produce an image-level decision [33, 35, 65–67]. Typically, these approaches have three steps: (1) produce a tissue-level semantic segmentation mask using CNNs for an input WSI, (2) extract features, such as distribution of tissues, from these semantic masks, and (3) produce an image-level decision using the features extracted from the semantic masks. These approaches learn global representations (information from segmentation masks) and have been found to be more effective than plain patch- and MIL-based approaches. However, one key challenge with these approaches is that they require tissue-level segmentation masks whose collection is challenging, because (1) domain experts are required for annotations and (2) pixel-wise annotations on images of gigapixel order is very time consuming. In contrast, `ScATNet` is a method for learning global representations from histopathological WSIs without the need for tissue-level segmentation masks.

End-to-end learning. Recent attempts at WSI classification focus on designing a single neural network that aggregates information from the entire image in a single shot [25, 68]. These methods extend the MIL-based approach with gradient check-pointing and advanced feature-fusion methods, such as self-attention. Inspired by model-level parallelism [15] and gradient check-pointing [69], these approaches break down the WSI classification pipeline into multiple stages and cache the intermediate results of CNN layers during forward and backward passes, allowing the systems to learn representations in an end-to-end fashion. For example, `HATNet` uses the transformers of Vaswani *et al.* [55] to aggregate the information from all instances in a breast biopsy image, while Pinckaers *et al.*[68] stitches the instance-

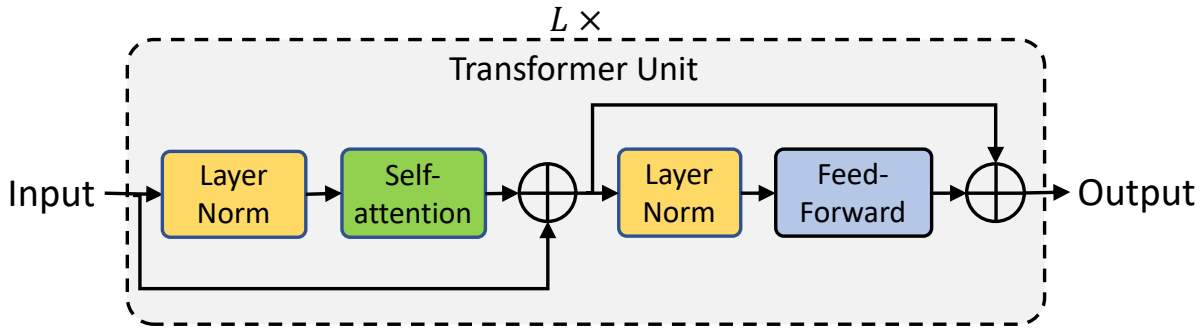


Figure 4.5: The transformer network stacks L transformer units sequentially. Each transformer unit consists of self-attention and feed-forward modules.

wise feature maps of a prostate cancer image at a very low-spatial resolution obtained from a CNN to produce an image-level feature map. `ScATNet` extends these approaches for classifying skin biopsies. Unlike these approaches that use WSIs at a single scale (typically at a zoom-level of 10×) for classification, this work proposes a scale-aware transformer that adapts to and uses the representations from multiple input scales to achieve higher classification performance. In our experiments, we compared our method with a CNN-based end-to-end WSI classification framework developed by Pinckaers *et al.*[68], details of this which are described in section 4.3.6.

Vision Transformers The transformers of Vaswani *et al.* [55], initially introduced for the task of machine translation (e.g., [70, 71]), are being explored for modeling images and computer vision tasks (e.g., [72, 73]). Transformers use self-attention, which allows the inputs (e.g., words in a sentence) to interact with each other and learn global representations. Carion *et al.*extended the standard encoder-decoder network of Vaswani *et al.*for the task of object detection. Recent work has extended transformers using a patch-based approach to image recognition at a large scale [72, 73]. Concurrent work has also utilized transformers and self-attention to medical image segmentation [74–77] and classification [78].

4.3.3 `ScATNet`: Scale-aware Transformer Network

Patch-based CNNs are state-of-the-art WSI classification methods that allow computer systems to learn representations from gigapixel size images (e.g. [31, 33, 63, 64, 79]). One of the main limitations of such systems is that they learn local representations, since the context capturing ability of such systems is limited to the patch-level. Another challenge is learning representations from multiple input scales. Because of limited GPU memory and the sheer size of these images, training multi-scale classification systems is computationally intractable. For example, the average size of a WSI (11K $\times$ 9.5K) in our dataset is 2000 times larger than the standard image classification dataset: the ImageNet [18] (224×224).

Motivated by the recent advancements in computer vision, especially vision transformers and the importance of input scales in clinical settings, this paper introduces scale-aware transformers in `ScATNet`, which allows our system to learn local and global representations from multiple input scales in an end-to-end fashion. Figure 4.4 shows the overview of `ScATNet`, which has three main steps: (1) learn local patch-wise embeddings using a CNN for each input scale, (2) learn contextualized patch-embeddings for each input scale using transformers, and (3) learn scale-aware embeddings across multiple input scales using transformers. These steps are described below.

Patch embeddings. The input WSI image $\mathbf{X}^{sc} \in \mathbb{R}^{W \times H}$ at scale sc with width W and height H is divided into m non-overlapping patches $\mathbf{X}^{sc} = (\mathbf{x}_1^{sc}, \cdots, \mathbf{x}_m^{sc})$, where $\mathbf{x}_i^{sc}$ is the i-th patch with width $\frac{W}{\sqrt{m}}$ and height $\frac{H}{\sqrt{m}}$. Patch-wise feature representations, referred to as patch embeddings, are obtained using an off-the-shelf CNN. The patch embedding $\mathbf{PE}_i^{sc} \in \mathbb{R}^e$ for the i-th patch $\mathbf{x}_i^{sc}$ is thus:

$$\mathbf{PE}_i^{sc} = \text{CNN}\left(\mathbf{x}_i^{sc}\right) \tag{4.1}$$

Contextualized patch embeddings. The patch embeddings $\mathbf{PE}^{sc} \in \mathbb{R}^{m \times e}$ are produced *independently* for each patch. In other words, these embeddings $\mathbf{PE}^{sc}$ do not encode inter-patch relationships. These embeddings $\mathbf{PE}^{sc}$ are fed to a transformer to learn inter-patch

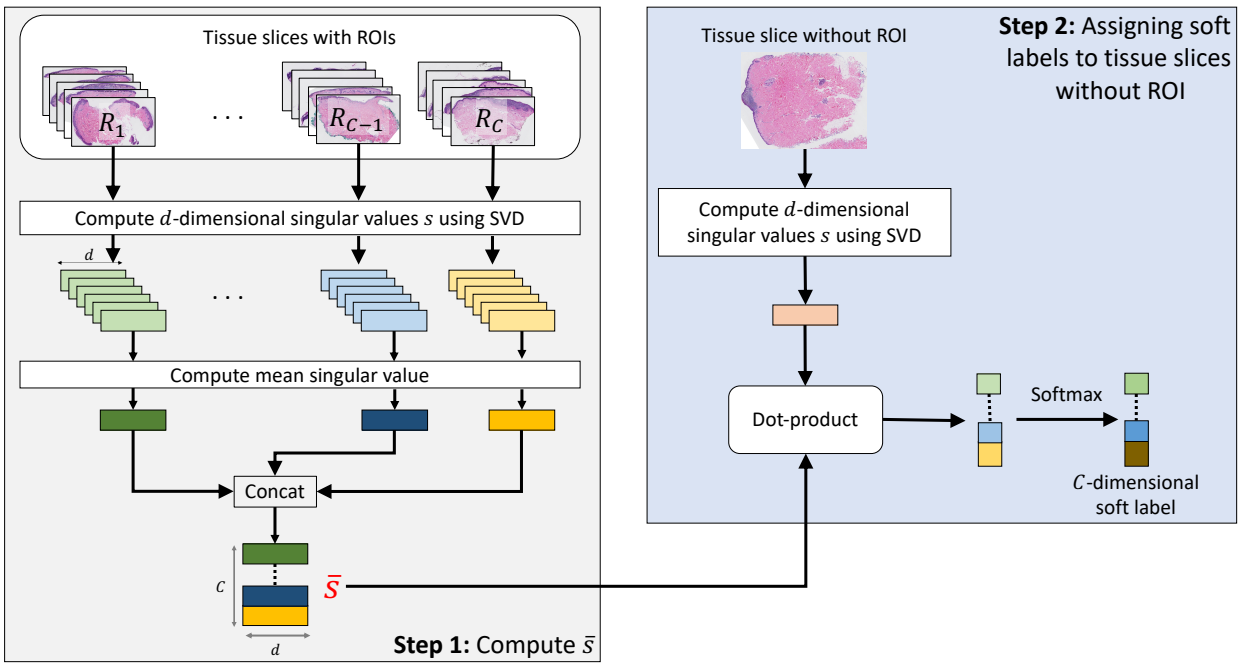


Figure 4.6: Overview of Soft labels calculation . Diagnostically constrained soft labels are calculated for tissue slices without an ROI using singular value decomposition (see Section 4.3.4).

relationships. Similar to vision transformers [72], patch-wise sinusoidal positional embeddings $\mathbf{PPE}^{sc} \in \mathbb{R}^{m \times e}$ are added to $\mathbf{PE}^{sc}$ to encode the position of input patches. The resultant embeddings are then fed to a transformer to produce contextualized patch embeddings $\mathbf{CPE}^{sc} \in \mathbb{R}^{m \times e}$.

$$\mathbf{CPE}^{sc} = \text{Transformer}\left(\mathbf{PE}^{sc} + \mathbf{PPE}^{sc}\right) \tag{4.2}$$

These contextualized embeddings $\mathbf{CPE}^{sc} \in \mathbb{R}^{m \times e}$ are then averaged along the m-dimension to produce an e-dimensional embedding vector $\overline{\mathbf{CPE}}^{sc} \in \mathbb{R}^{e}$. $\overline{\mathbf{CPE}}^{sc}$ encodes the local (from CNN) and global (from Transformer) information in an image $\mathbf{X}^{sc}$.

Contextualized scale embeddings. The embedding $\overline{\mathbf{CPE}}^{sc}$ encodes the information in an image $\mathbf{X}^{sc}$ at scale sc. Let us assume that we have $\mathscr{S}$ scales. For each scale $sc \in$

$[0, ..., \mathscr{S}]$, we produce embedding vector $\overline{\mathbf{CPE}}^{sc}$ and concatenate them to produce scale-level embeddings $\mathbf{SE} = \text{Concat}(\overline{\mathbf{CPE}}^{1}, \cdots, \overline{\mathbf{CPE}}^{\mathscr{S}})$. These embeddings $\mathbf{SE} \in \mathbb{R}^{\mathscr{S} \times e}$ do not encode information about the relationships between the different scales. To learn scale-aware representations while retaining positional information about each scale, scale-level learnable positional embeddings $\mathbf{PSE} \in \mathbb{R}^{sc \times e}$ are added[1]to $\mathbf{SE}^{sc \times e}$. The resultant embeddings are then fed to another transformer to produce contextualized scale embeddings $\mathbf{CSE} \in \mathbb{R}^{sc \times e}$.

$$\mathbf{CSE} = \text{Transformer}\left(\mathbf{SE} + \mathbf{PSE}\right) \tag{4.3}$$

For predicting the diagnostic class, `ScATNet` first flattens the scale-aware embeddings $\mathbf{CSE} \in \mathbb{R}^{sc \times e}$ to produce a $(sc \cdot e)$-dimensional vector and then classifies it using a linear classifier into C diagnostic categories.

4.3.4 Soft-labels for Skin Biopsy Images

Skin biopsy images often contain multiple tissue slices on a single WSI, as shown in Figure 4.7a. In general, the representative regions-of-interest (ROIs; shown in red in Figure 4.7a) that helped pathologists in diagnosis belong to one or two tissue slices, while the other tissue slices may correspond to other diagnosis categories. Assigning the same diagnostic label to all tissue slices (similar to MIL-based approaches) results in more false tissue-label pairs and hinders learning representations. To address this, we propose a soft labeling method, as illustrated in Figure 4.6.

Given a dataset $\mathcal{D}$ with N training WSIs along with representative ROIs for each WSI (each WSI contains multiple slices) that helped in diagnosis, we aim to assign soft labels to tissue slices that do not have ROIs. Tissue slices from each WSI are extracted and then categorized into one of the two sets: (1) tissue slices $\mathcal{R}$ with an ROI and (2) tissue slices $\mathcal{NR}$ without an ROI. Since each slice in $\mathcal{R}$ has a representative ROI, we further split $\mathcal{R}$ into C subsets, $\mathcal{R} = \{R_1, \cdots, R_C\}$, based on the diagnostic category, where R_i represents the

[1]Unlike the number of patches m, the number of scales $\mathscr{S}$ is fixed. Therefore, we learned the positional embeddings for each scale using torch.nn.Embedding in PyTorch. Compared to sinusoidal positional embeddings, learned embeddings improves the performance by about 0.5-1.0%.

subset for diagnostic category i and C denotes the number of diagnostic categories. Next, we compute the mean singular value vector $\bar{\mathbf{s}}_i$ for each subset R_i as:

$$\bar{\mathbf{s}}_i = \frac{1}{n}\sum_{j=1}^{n} \mathbf{s}_i^j \tag{4.4}$$

where $\mathbf{s}_i^j$ is the d-dimensional singular-value vector obtained after applying singular-value decomposition (SVD) to the j-th tissue slice in R_i. The idea is to use these vectors to represent the appearance of the diagnostic categories. We used singular values because of their uniqueness and robustness properties [80–83]. However, other dimensionality reduction methods could also be used.

For the j-th slice in $\mathcal{NR}$, the C-dimensional soft label vector $\hat{y}^j$ is computed as:

$$\hat{y}^j = \text{softmax}\left(\bar{\mathbf{s}} \cdot \hat{\mathbf{s}}^j\right) \tag{4.5}$$

where $\hat{\mathbf{s}}^j$ is a d-dimensional singular value vector obtained after applying SVD to the j-th tissue slice in $\mathcal{NR}$ and $\bar{s} = \{\bar{s_1}, \ldots, \bar{s_C}\}$.

Tissue slices without an ROI do not help in diagnosis decisions. Clinically, such slices can often belong to lower diagnostic categories than the category assigned to the WSI they are part of. We incorporate this diagnostic constraint in our soft labeling method. For a four-class dataset (1: *MMD*, 2: *MIS*, 3: *pT1a*, and 4: *pT1b*), suppose that a WSI corresponding to class k has m tissue slices and one of the tissue slices has an ROI, as shown in Figure 4.7a. Soft label vectors $\hat{y}^j$ for the jth slices without ROI ($j \in [0, m-1]$) can be obtained from equation 6. Then, to take one step further, *diagnostically constrained* soft label vector $\tilde{y}^j = \{\tilde{y}_1^j, ..., \tilde{y}_C^j\}$ is computed as:

$$\tilde{y}_c^j = \frac{\hat{y}_c^j}{\sum_{c=0}^{k} \hat{y}_c^j}, \quad if\ c < k$$
$$\tilde{y}_c^j = 0 \quad if\ c \geq k \tag{4.6}$$

Figure 4.7a illustrated an example WSI corresponding to class 3 (pT1a), which has three tissue slices, and one of the tissue slices has an ROI. If the soft label vectors $\hat{y}^j$ for these

two slices without ROI are $[0.46, 0.39, 0.08, 0.07]$, $[0.21, 0.54, 0.1, 0.15]$, the resulting soft label vectors with the diagnostic constraint $\tilde{y}^j$ are $[0.54, 0.46, 0, 0]$, and $[0.28, 0.72, 0, 0]$ respectively.

4.3.5 Dataset and Implementation details

Dataset and Accuracy Data from Pathologists The dataset used in this study is the Skin Biopsy Dataset described in Chapter 2.2. To compare the results from `ScATNet` with the interpretations of practicing U.S. pathologists, we used data from a prior clinical study in which 187 pathologists interpreted the same WSIs [9]. Each pathologist interpreted a random subset of 36 cases, and their diagnoses were classified into the same four diagnostic categories. This resulted in 10 independent diagnostic labels (on an average) per slide and provided a way to compare the classifications performed by human pathologist to `ScATNet`. These interpretations are only used for independent evaluation. The ground truth diagnosis of each slide is the consensus diagnosis of three experienced dermatopathologists.

Extracting tissue slices from WSIs. The original WSIs were collected at a zoom level of 40×. Because WSIs at 40× require extensive computational resources, we extracted WSIs at lower zoom levels of 7.5× (average size 8348 x 7202), 10× (average size 11130 x 9603), and 12.5× (average size 13913 x12003). These zoom levels were selected based on previous work on histopathological image classification for different tissues [33, 63, 79], since they provide a good tradeoff for 1) capturing sufficient local context without including irrelevant details and 2) providing variable local information without losing similar correlation. We refer to different zoom levels as "input scales" in this work. Each WSI has multiple tissue slices with a background region between the slices that does not aid in diagnosis (Figure 4.7a). Therefore, individual tissue slices were extracted using a histogram-based segmentation method of Otsu [84] followed by morphological operations (opening-closing and hole filling) and contour-related operations available in OpenCV.

Soft-labels. To assign soft labels for tissue slices without an ROI, SVD is applied to obtain d-dimensional singular-value vectors as described in the Methods section. In this study, d is set to 50.

Architecture. We use MobileNetv2 [85] pretrained on the ImageNet dataset [18] as our CNN for extracting patch-wise embeddings. MobileNetv2 was chosen, because it is light-weight, fast, and delivers state-of-the-art performance across different machine vision tasks, such as classification, detection, and segmentation. `ScATNet` is not limited to a particular CNN and other CNNs, such as VGG [86] and ResNet [13] may also be suitable for extracting patch-wise embeddings.

MobileNetv2 outputs 1280-dimensional patch-wise embeddings after global average pooling. `ScATNet` projects these patch-wise embeddings linearly to a 128-dimensional space ($e = 128$) and then learns contextualized patch-wise and scale-wise embeddings using transformers. For learning contextualized patch-wise and scale-wise representations, a stack of two transformer units is used. Also, in each transformer unit, the number of heads in the self-attention layer is set to 4, and the feed forward network dimension is set to 512.

Training Details `ScATNet` is trained for 200 epochs in an end-to-end fashion using the ADAM optimizer with a linear learning rate warm-up strategy and step learning rate decay. The learning rate is first warmed up from 10^{-6} to 5×10^{-4} in 500 steps. In the next 50 epochs, the model is trained with a learning rate of 5×10^{-4}. After that, the learning rate is reduced by half at the 100-th and 150-th epochs. Because of the large size of these images, extensive computational resources are required. To learn representations with limited computational resources, we freeze the convolutional layers in a CNN and train only the transformer networks. Our models are trained on a single NVIDIA GeForce 2080 GPU with 10 GB GPU memory. Similar to other medical imaging datasets, our dataset is small. Therefore, to improve its robustness against stochastic noise, we average best 3 and best 5 model checkpoints within a single training process [87] and select the one that performs best

on the validation set. We then evaluate it on the (unseen) test set. A WSI in a test set may contain multiple tissue slices. To predict the final diagnostic label, we use max-voting. This choice is inspired by pathologists' diagnosing behavior, i.e., if one of the tissue slices in a WSI is invasive melanoma, then the entire WSI corresponds to invasive melanoma and cannot be *MMD* or *MIS*.

4.3.6 Baseline Methods

`ScATNet`'s performance is compared with five recent whole slide image classification methods.

Patch-based classification. The first method is a standard patch-based CNN classification framework that was built following saliency-based methods, related to the work of Hou et. al. [63] and that of E. Mercan *et al.* [36], (R1 and R2 in Table 4.1). This method treats each patch independently and assigns the same diagnostic label to all patches in the WSI during training. During evaluation, majority-voting is used for predicting the slide-level diagnostic label. Similar to the use of `ScATNet`, Mobilenetv2, pretrained on the ImageNet dataset was used as the CNN model.

Weighted feature aggregation. The second method is a CNN-based deep feature extraction framework developed by C. Mercan *et al.* [37] that builds slide-level feature representations via weighted aggregation of the patch representations (R3 and R4 in Table 4.1). Under this framework, feature extraction is performed in three steps: (1) using a CNN (e.g. VGG16) to extract features on a patch-by-patch basis; (2) concatenating the weighted instances of the extracted feature activations using either penultimate layer features (penultimate-weighted) or hypercolumn features (hypercolumn-weighted) to form patch-level feature representations; and (3) fusing the patch-level representations via average pooling to form the slide-level representation.

ChikonMIL. The method of Chikontwe *et al.*(ChikonMIL) (R3 in Table 4.1) [60] first selects the top-k patches, and then uses these patches for instance- and bag-representation learning. This method also uses a center loss that reduces intra-class variability and a soft assignment to learned diagnostic centroid for final diagnosis.

MS-DA-MIL . Multi-scale Domain-adversarial Multiple-instance (MS-DA-MIL) CNN developed by Hashimoto *et al.* [61] (R7 and R8 in Table 4.1) is a framework that learns from groups of patches extracted different scales (x10 and x20) with attention mechanism. However, in contrast to the proposed end-to-end learning framework, MS-DA-MIL-CNN first trains a single-scale MIL network to classify for each scale. Then, a multi-scale network is trained using the features extracted using pre-trained single-scale MIL networks.

Streaming CNN. Streaming CNN is a work of Pinckaers *et al.* [68] (R4 in Table 4.1). This method uses a patch-based approach with gradient checkpointing and streaming, which allows it to classify whole slide images in an end-to-end fashion.

4.3.7 Results

Hard vs. soft labels. The performance of our soft labeling method (Section 4.3.4) is compared with three other labeling methods. For illustration, for the four classes in our dataset (1: *MMD*, 2: *MIS*, 3: *pT1a*, and 4: *pT1b*), we use a WSI corresponding to *pT1a* (class 3; shown in Figure 4.7a) with 3 slices, one having a ROI.

- *Hard labels:* Similar to MIL-based approaches, all tissue slices in the WSI are assigned the same diagnostic label. For the above example, each tissue slice will have a label of [0, 0, 1, 0] (one-hot vector encoding).

- *Label smoothing:* The label smoothing method of Szegedyet *et al.* [88] produces soft labels that are a weighted average of the hard labels and the uniform distribution over labels. It regularizes the network and helps improve the performance [89]. For the

Row #	Method	Accuracy	F1	Sensitivity	Specificity	AUC
R1	Patch-based (SSC)	0.35	0.35	0.35	0.79	0.67
R2	Patch-based (MSC)	0.40	0.40	0.40	0.80	0.68
R3	Penultimate-weighted (SSC)	0.44	0.44	0.44	0.81	0.67
R4	Hypercolumn-weighted (SSC)	0.43	0.43	0.43	0.43	0.67
R5	Streaming CNN (SSC)	0.32	0.32	0.32	0.77	0.58
R6	ChikonMIL (SSC)	0.56	0.56	0.56	0.85	0.74
R7	MS-DA-MIL (SSC)	0.49	0.49	0.49	0.83	0.68
R8	MS-DA-MIL (MSC*)	0.58	0.58	0.58	0.86	0.75
R9	`ScATNet` (SSC)	0.60	0.60	0.60	0.87	0.77
R10	`ScATNet` (MSC)	**0.64**	**0.64**	**0.64**	**0.88**	**0.79**

Table 4.1: Comparison of overall performance with state-of-the-art WSI classification methods across different metrics on the test set. Here, SSC denotes single input scale (10×). MSC denotes multiple input scales (7.5×, 10×, 12.5×). MSC* denotes multiple input scales (10×, 20×)

same example, the soft labels for each of these slices would be [0.033, 0.033, 0.9, 0.033] with a label smoothing value of 0.1. In other words, the label for class 3 is smoothed from 1 to 0.9 and the remaining mass of 0.1 is equally distributed among the remaining three classes.

- *Constrained label smoothing:* This extends the hard labels and label smoothing methods by incorporating the diagnostic constraint that tissue slices without a ROI should belong to lower diagnostic categories. For example, if the WSI has a hard label of $pT1a$ (i.e. class 3), then the tissue slices without a ROI can only belong to lower diagnostic categories (i.e., *MMD* and *MIS*). For the same example as above, the slice with an ROI

will have a label of [0, 0, 1, 0] while the slices without an ROI will have constrained labels of [0.5, 0.5, 0, 0].

Figure 4.7b contrasts our soft labeling method with these methods while quantitative comparison between these methods is given in Figure 4.7b. These experiments demonstrated that our soft labeling method is more effective as compared to these existing methods. In subsequent experiments, we use our soft labeling method.

Impact of number of patches m. Figure 4.8 compares the performance of single scale `ScATNet` with different numbers of crops m at three different input resolutions (7.5×, 10×, and 12.5×). Using fewer crops at larger resolution (e.g., 25 crops at a resolution of 12.5×) and more crops at smaller resolutions (e.g., 81 crops at a resolution of 7.5×) hurts the performance. This is likely because MobileNetv2, the CNN used in this work, is pre-trained on the ImageNet dataset at a fixed image size of 224×224. With very large (fewer number of crops at larger image resolution) or very small (larger number of crops at smaller image resolution) patch sizes, the CNNs may have difficulty in capturing representative features and yield poor patch embeddings, which hurts the performance. We note that scaling patch size alone may not be an optimal solution and future studies, especially compound model scaling in EfficientNet [90], may help improve the performance.

In the rest of the experiments, we used $m = 25$ for 7.5× input resolution, $m = 49$ for 10× input resolution, and $m = 81$ for 12.5× input resolution, as these had the best performance.

Single vs. multiple input scales. Figure 4.10a compares the overall performance of `ScATNet` across different metrics on single- and multi-scale inputs, while class-wise accuracy is given in Figure 4.10b. With inputs at multiple scales, we observe improvements in overall as well as class-wise performance. Notably, we observe significant improvement with multiple scales (two and three scales) in the *pT1b* invasive melanoma cancer category. Compared to two scales, the overall performance with three scales remains the same. However, with three

scales, the performance across all diagnostic classes (Figure 4.10b) is much more evenly distributed, which is not seen in all other combinations.

Comparison with baseline methods. Figure 4.1 compares the classification performance of `ScATNet` with existing methods on the test set. `ScATNet` outperforms all five existing methods to which it was compared by a significant margin across different metrics. Furthermore, compared to the ChikonMIL method [60] and the MS-DA-MIL method [61] with multi-scale input, which delivered the two best performances among the five baseline methods, `ScATNet` delivered better performance across all diagnostic categories (see Figure 4.9), except the pT1b category. This is likely because the ChikonMIL method samples more relevant patches corresponding to the pT1b category as compared to other diagnostic categories, while the MS-DA-MIL method uses an input at higher resolution (x20), which might yield more information at the cellular level that helped to distinguish the pT1b category. We believe that complementing the proposed method with the patch sampling method of Chikontwe *et al.* (2020) would further improve the performance. We will investigate such methods in the future.

Comparison with U.S. pathologists. Table 4.2 shows that `ScATNet` achieves similar performance to practicing U.S. pathologists who interpreted these same cases in overall accuracy (pathologists vs. `ScATNet`: 0.65 vs. 0.64), suggesting its potential as a second reader to help pathologists in clinical settings for reducing classification uncertainties.

4.3.8 Conclusion

Diagnosis of melanocytic lesions is among the most challenging areas of pathology. Previous studies indicate that diagnostic errors occur frequently [6–8]. False positive readings for suspected melanoma range from 6% to 17% [91, 92]. Diagnostic errors may lead to inappropriate treatment decisions and harm to patients. With FDA approval, digitized whole slide imaging systems show great potential for improving the diagnostic performance of pathologists. In

Diagnostic	Accuracy		F1		Sensitivity		Specificity	
Category	**PG**	**Ours**	**PG**	**Ours**	**PG**	**Ours**	**PG**	**Ours**
MMD	0.92	0.79	0.71	0.75	0.92	0.79	0.76	0.89
MIS	0.46	0.40	0.49	0.44	0.46	0.40	0.85	0.84
pT1a	0.51	0.65	0.62	0.63	0.51	0.65	0.95	0.84
pT1b	0.72	0.77	0.72	0.74	0.78	0.77	0.97	0.92
Overall	0.65	0.64	0.65	0.64	0.65	0.64	0.88	0.88

Table 4.2: Comparison of `ScATNet` with pathologists' (PG) performance. Pathologists' performance data is from a prior *independent* clinical study of 187 pathologists [9] who interpreted these same 115 cases in our test set (Table 2.1). Diagnostic terms are defined as the following: *mild and moderate dysplastic nevi* (*MMD*), *melanoma in situ* (*MIS*), *invasive melanoma stage pT1a* (*pT1a*), *invasive melanoma stage* $\geq$ *pT1b* (*pT1b*).

this section, we introduce the scale-aware transformer network `ScATNet` for learning representations from variably-sized whole slide skin biopsy images at multiple scales. Compared to existing methods, `ScATNet` delivered better performance. Importantly, `ScATNet` also delivered comparable performance to practicing U.S. pathologists who interpreted the same cases.

4.4 Summary

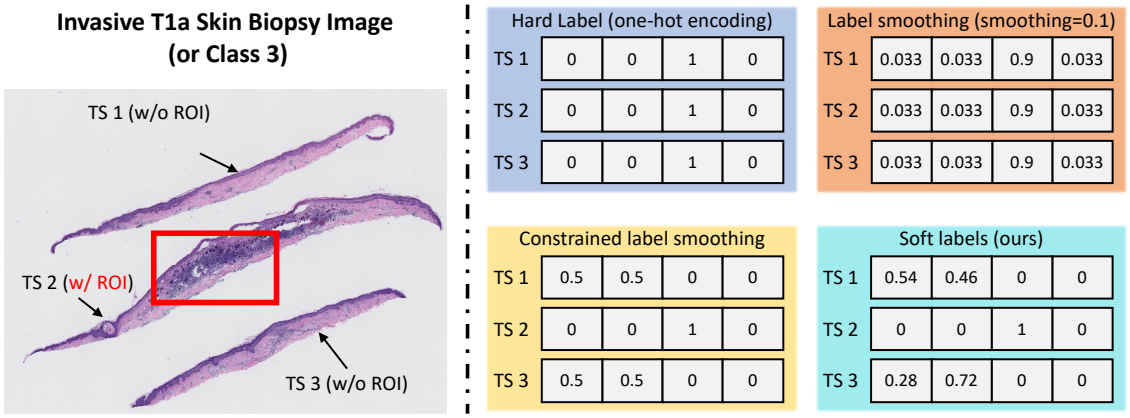
This chapter delineates the innovative application of transformer networks to the field of medical imaging, specifically addressing the complexities involved in diagnosing diseases from whole slide images (WSIs). Two projects were described in this chapter:

1. The first project, `HATNet` [25] extends the bag-of-words approach and uses self-attention

to encode global information, allowing it to learn representations from clinically relevant tissue structures without any explicit supervision. It outperforms a lot of state-of-the-art baseline methods. More importantly, our analysis reveals that `HATNet` learns representations from clinically relevant structures, and it matches the classification accuracy of 87 U.S. pathologists for this challenging test set

2. The second project, `ScATNet` [4] expands on `HATNet` to include learning a multi-scale representation. This enhancement allows `ScATNet` to not only detect minute, fine-grained details crucial for accurate diagnostics but also to comprehend broader contextual nuances, thereby significantly augmenting the model's diagnostic precision across diverse histopathological features and scales.

The chapter highlights the potential of transformer-based models to handle the immense scale and complexity inherent in WSIs, aiming to illustrate the transformative impact these models could have on medical diagnostics.



(a)

Method	Accuracy	Specificity	AUC
Hard labels	0.50	0.83	0.73
Label smoothing	0.50	0.83	0.71
Constrained label smoothing	0.56	0.85	0.77
Soft labels (Ours; Section 4.3.4)	**0.60**	**0.87**	**0.77**

(b)

Figure 4.7: **(a)** shows different labeling methods, including our soft label method, for an *pT1a* skin biopsy image with three tissue slices and one representative region of interest (red box) that helped expert pathologists in diagnosing the image. **(b)** compares the performance of different labeling methods. Our soft labeling method is simple and effective; it reduces the ambiguity that arises during training because of multiple tissue slices in a WSI that do not have a ROI and helps improve the performance. In (b), we do not report sensitivity and specificity, because their values are the same as accuracy.

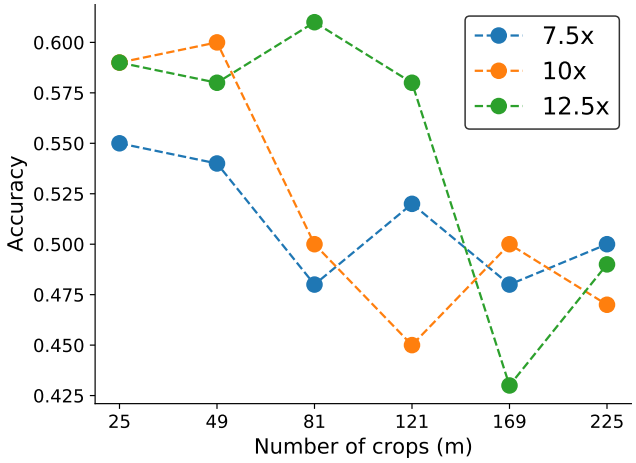


Figure 4.8: Effect of number of crops (m) on the performance of `ScATNet` (single scale) for inputs at three different scale levels (7.5×, 10×, and 12.5×).

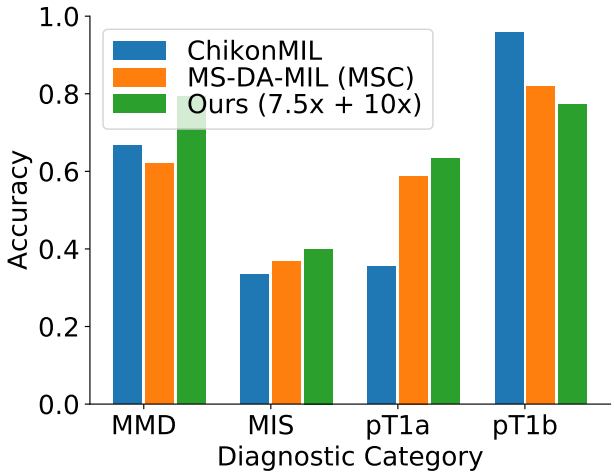
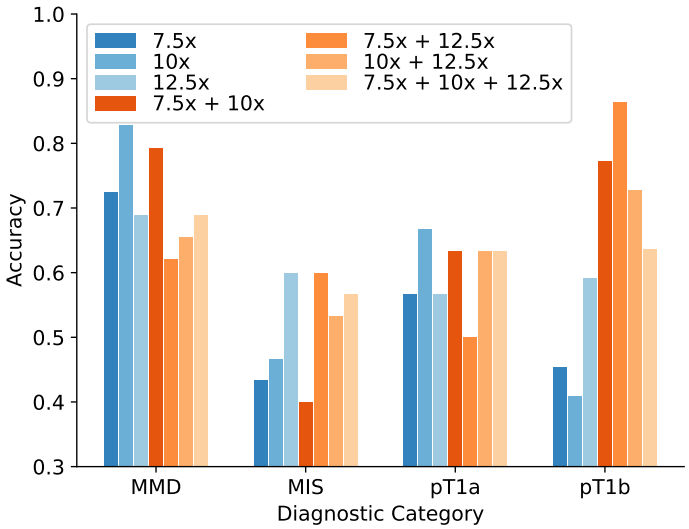


Figure 4.9: Comparison of class-wise accuracy with state-of-the-art WSI classification methods on the test set. Diagnostic terms are defined as the following: *mild and moderate dysplastic nevi* (*MMD*), *melanoma in situ* (*MIS*), *invasive melanoma stage pT1a* (*pT1a*), *invasive melanoma stage* $\geq$ *pT1b* (*pT1b*). Overall, `ScATNet` delivered better performance across all diagnostic categories except the pT1b category.

Input scales			Accuracy	F1	Sensitivity	Specificity	AUC
7.5×	10×	12.5×					
✓			0.55	0.55	0.55	0.85	0.75
	✓		0.60	0.60	0.60	0.87	0.77
		✓	0.61	0.61	0.61	0.87	0.78
✓	✓		0.64	0.64	0.64	0.88	0.79
✓		✓	0.63	0.63	0.63	0.88	0.80
	✓	✓	0.63	0.63	0.63	0.88	0.79
✓	✓	✓	0.63	0.63	0.63	0.88	0.79

(a) Overall performance of `ScATNet`



(b) Class-wise accuracy of `ScATNet`

Figure 4.10: Effect of single and multiple input scales. For single and multiple input scales, we compared the overall performance of `ScATNet` across different metrics in (a) while in (b), we compared the class-wise accuracy. With multiple input scales, overall and class-wise performance, especially in invasive cancer categories (pT1a and pT1b), of `ScATNet` improved across all evaluation metrics. Diagnostic terms are defined as the following: *mild and moderate dysplastic nevi* (*MMD*), *melanoma in situ* (*MIS*), *invasive melanoma stage pT1a* (*pT1a*), *invasive melanoma stage* $\geq$ *pT1b* (*pT1b*).

Chapter 5

SEMANTICS-AWARE ATTENTION GUIDANCE

5.1 Introduction

The integration of deep learning technologies, notably convolutional neural networks (CNNs) and transformers, has dramatically transformed the landscape of histopathological image analysis [93, 94]. However, learning from gigapixel whole slide images (WSIs) remains a challenging problem due to their size, making end-to-end learning extremely expensive. As a result, WSI classification methods often follow a bag-of-words (BoW) model, treating large patches as bags and smaller image patches as words or instances, as described in Chapter 4. However, as described in Chapter 4, human pathologists diagnose whole slide images by identifying suspicious regions at low magnification, then switching to high magnification to examine individual cells and structures, ultimately reaching a definitive diagnosis [95]. This crucial multi-scale assessment is often overlooked by MIL models (described in Section 4.3.2), which treat image patches independently, thereby missing the essential context provided by varying magnifications.

The limitations of MIL models in capturing long-range interactions between entities hinder their ability to effectively capture nuanced details critical for accurate diagnosis. To address this, transformer models have been adopted to capture interdependencies among patches and formulate comprehensive representations, notably advancing beyond MIL's limitations [4, 96–99]. In the past few years, our research has focused on developing WSI diagnosis pipelines using original WSIs and their diagnostic categories, including 1) an end-to-end holistic attention network for classifying breast biopsy images (`HATNet` discussed in Chapter 4) [25], and 2) an end-to-end scale-aware transformer network (`ScATNet`, also discussed in Chapter 4) for classifying skin biopsy images [4]. However, our analysis of the

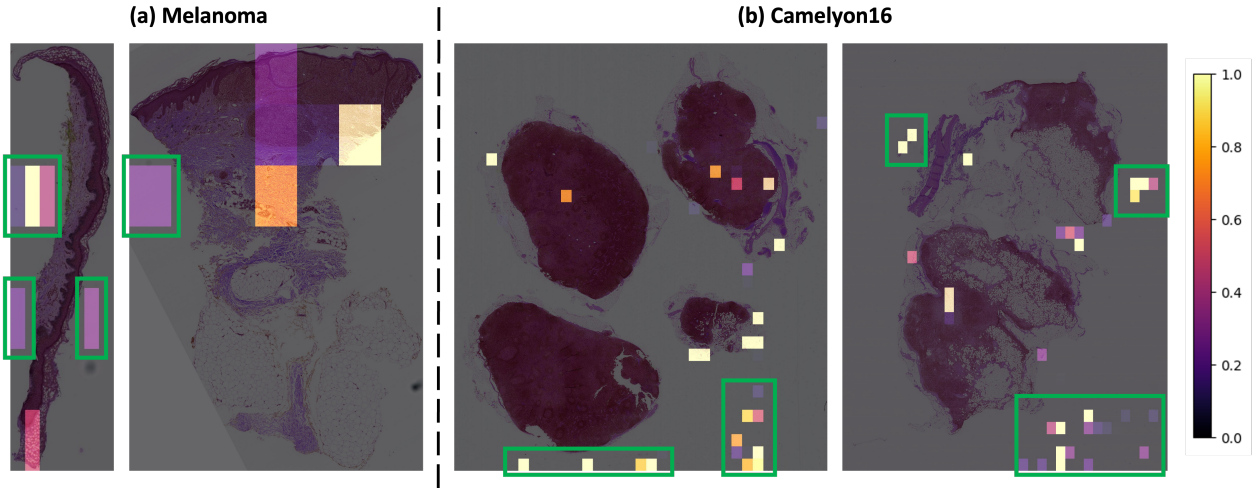


Figure 5.1: Visualization of the baseline model's (ScAtNet [4]) attention on (a) skin biopsy WSIs in the melanoma dataset and (b) breast biopsy WSIs in the Camelyon16 dataset. Green boxes show examples of the baseline model mistakenly focusing on background regions. The signal and attention values are normalized for visualization purposes.

attention regions produced by `ScATNet` revealed that these models often mistakenly focus on non-cancerous regions or empty spaces, as shown in Fig. 5.1. This raises concerns about the interpretability, reliability, and diagnostic accuracy of these models.

In addition to diagnosis pipelines, we have worked on detecting diagnostically important cellular entities (discussed Section 3.2) [24], and semantic segmentation of clinically important tissues in skin biopsy images using sparse and coarse annotations from pathologists (discussed in Section 3.3) [3]. After achieving satisfying preliminary results, we would like to create an end-to-end diagnosis framework that leverages the knowledge in the semantic segmentation to further improve diagnosis performance and interpretability.

In this chapter, we introduce the `SAG` (Semantics-Aware Attention Guidance) framework, an interpretable, multimodal learning framework that extends the methodologies discussed in previous chapters. The `SAG` framework is designed around two core components: 1) a technique for converting diagnostically relevant entities into attention signals, and 2) a

flexible attention loss that efficiently integrates various semantically significant information. The content of this chapter is based on Liu and **Wu** *et al.* [26].

5.2 Related Work

The integration of additional domain information into diagnostic models has transformed medical imaging, particularly by enhancing diagnostic accuracy where data is scarce. Miao *et al.* [100] introduced spatial prior attention with binary anatomy knowledge maps to infuse anatomical knowledge into whole slide image (WSI) diagnosis, highlighting the potential of integrating more complex prior knowledge. Similarly, Nofallah *et al.*[101] have incorporated tissue segmentation masks into the diagnostic process [101]. These masks, generated from sparse and coarse annotations of full skin biopsy WSIs, are used as additional channels in the ScATNet model to enrich the input data. While this approach has shown promise in improving model learning, especially in challenging diagnostic classes, a significant limitation arises from the fact that these segmentation masks are integrated without any supervisory feedback mechanism to ensure that the model effectively learns from this crucial signal. This oversight might limit the model's ability to fully utilize the segmented information for more accurate predictions.

Expanding on the concept of integrating diverse data types, Chen *et al.* [97] combines genomics data with whole slide images (WSIs) through their Multimodal Co-Attention Transformer (MCAT) framework to predict patient outcomes. This model utilizes a dense co-attention mapping to learn how histology patches attend to genes, enhancing interpretability and reducing the computational complexity associated with handling large WSI bags. Their method, inspired by techniques used in Visual Question Answering (VQA), consistently outperforms existing approaches across five different cancer datasets, comprising 4,730 WSIs and 67 million patches. Despite these advancements, the MCAT framework lacks the ability to adapt to other forms of data like tissue maps or additional images, underscoring the ongoing challenge of developing models that can seamlessly integrate various kinds of biomedical information.

Expanding on the concept of integrating diverse data types, Chen *et al.* [97] have broadened the scope by integrating genomics data with whole slide images (WSIs) to predict patient outcomes, demonstrating the benefits of using multiple types of data together. However, their model struggles to adapt to other forms of data like tissue maps or extra images, highlighting the need for models that can effortlessly combine various kinds of biomedical information.

In response to the shortcomings of existing approaches, we introduce the Semantics-Attention-Guiding (`SAG`) framework, offering several key advancements:

- An innovative attention guiding module that can be integrated with any attention-based multiple instance learning or Transformer models.
- A versatile attention-guiding loss designed to effectively utilize diverse semantic information, such as tissue and cancerous region masks.
- A heuristic method for transforming diagnostically important entities into heuristic-guidance signals.
- Demonstrating generalizability across different cancer types, showcasing robust performance on diverse datasets.

5.3 `SAG`

The Semantics-Aware Attention Guidance (`SAG`) framework, illustrated in Fig. 5.2, addresses a critical limitation in existing approaches by actively guiding the attention of deep learning models towards diagnostically relevant information within whole slide images (WSIs). This targeted attention improves diagnostic accuracy and interpretability. A key strength of SAG lies in its versatility. Unlike prior methods that are often limited to specific architectures, SAG is demonstrably compatible with various multi-instance learning (MIL) and transformer-based architectures, making it a more broadly applicable solution.

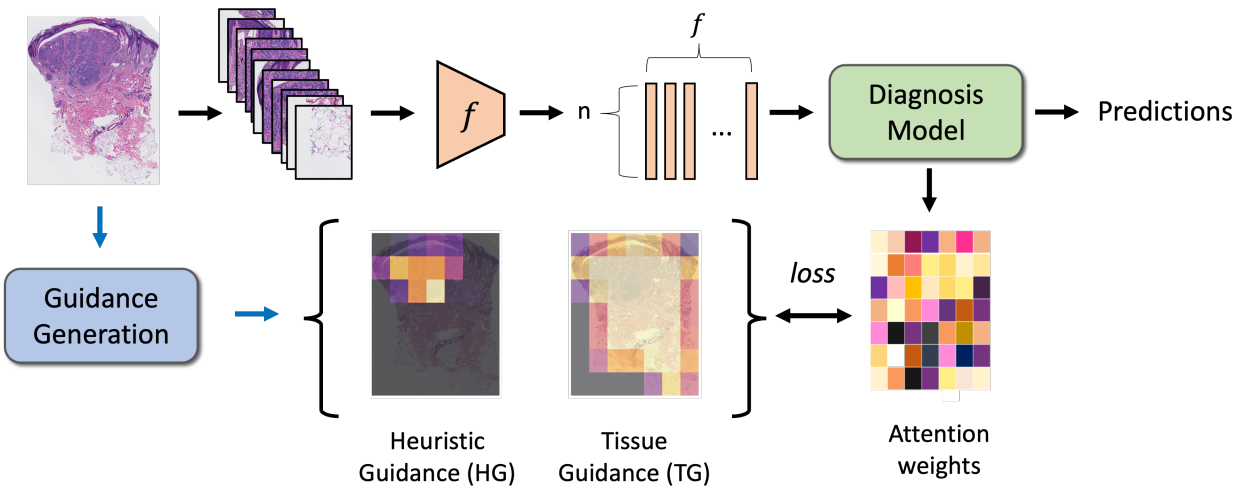


Figure 5.2: Overview of the SAG approach for improving WSIs diagnosis models. The process begins by dividing a high-resolution histopathological image into n non-overlapping patches, following the methodologies described in ScATNet. These patches are then processed to extract embeddings using an off-the-shelf feature extractor g, similar to the initial steps in ScATNet. A diagnostic network subsequently utilizes the $n \times f$-dimensional feature vector for classification into distinct categories. Throughout training, heuristic guidance (**HG**) and tissue guidance (**TG**) are employed to direct the model's attention towards areas of diagnostic relevance.

5.3.1 Diagnosis Models

Utilizing a pretrained feature extractor g for deriving patch embeddings, detailed in Chapter 5.4, our framework extends the versatility demonstrated by ScATNet. We apply the SAG framework to two state-of-the-art baseline models: a transformer-based model, ScATNet [4], and an MIL-based model, ABMIL [102], to showcase its generalizability.

5.3.2 Attention Mechanism

`ScATNet` As described in Section 4.3.3), `ScATNet` processes each scale independently, learning the relationships between image patches using a transformer architecture. It employs self-attention to analyze and understand the context within the patches. Initially, the input matrix $\mathbf{x} \in \mathbb{R}^{p_{sc} \times d}$, where p_{sc} represents the number of patches at scale sc and d is the feature dimension, is transformed into query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) vectors. The self-attention mechanism is computed as follows:

$$\text{Self-attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\mathbf{Q} \cdot \mathbf{K}^T) \cdot \mathbf{V} \tag{5.1}$$

This process generates the attention matrix $\mathbf{Attn}_{sc} \in \mathbb{R}^{p_{sc} \times p_{sc}}$:

$$\mathbf{Attn}_{sc} = \text{softmax}(\mathbf{Q}_{sc} \cdot \mathbf{K}_{sc}^T) \tag{5.2}$$

MIL methods The model attention weights ($\mathbf{Attn_m}$) of the MIL methods are formulated as the weighted aggregation of instance embeddings [102]:

$$\mathbf{Attn_m} = \sigma(x) \in \mathbb{R}^p, \tag{5.3}$$

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

5.3.3 Attention Guidance

To regularize the model's attention **MA**, we induce two types of semantic attention guidance: tissue guidance (**TG**) and heuristic guidance (**HG**) (Fig. 5.3), each represented as a vector $\in \mathbb{R}^p$. The generation of attention guidance is described in two steps: 1) Acquisition of tissue mask and diagnostic heuristics, and 2) Calculation of guidance weights.

For the generation of the tissue guidance mask **TG**, we apply Otsu's method [103] to segment tissue regions effectively. This technique converts the initial input image (Fig. 5.3a) into a binary tissue mask as displayed in Fig. 5.3b.

The heuristic guidance **HG** leverages specific information relevant to the dataset and disease, including various structures, tissues, and cell types. As depicted in Fig. 5.3, cell segmentation is first performed on a targeted cell type (Fig. 5.3d). These cells are then grouped using the density-based spatial clustering algorithm, DBSCAN [104], to form clusters. For each cluster, a convex hull is computed [105] (Fig. 5.3e), serving as a semantic signal to direct the model's focus during training (Fig. 5.3f).

The guidance weights $G \in \mathbb{R}^p$ are calculated to map these semantic signals to the model's attention mechanism. This mapping is detailed in the following equation (Fig. 5.3c):

$$G_i^k = \frac{N_i^k}{\sum_{j=1}^{p} N_j^k}, \quad k \in \{\mathbf{TG}, \mathbf{HG}\}, \tag{5.4}$$

where G_i^k represents the guidance weight for patch i, and N_i^k quantifies the area of the mask within that patch relative to the total mask area.

5.3.4 Loss Functions

In our `SAG` framework, we design two specific loss functions to optimize the model's attention mechanisms. The first is an inclusion-exclusion loss (L_{focus}, defined in Eqn. 5.6), which ensures that the model focuses on relevant tissue areas while ignoring background and artifacts. The second is an L2-loss for attention alignment (L_{mse}, defined in Eqn. 5.5), intended to align the model's attention with heuristic guidance signals indicating diagnostic relevance.

For the heuristic guidance (**HG**), which indicates areas of diagnostic importance, we employ the mean squared error (MSE) loss, L_{mse}, to fine-tune the model's attention outputs, **MA**:

$$L_{mse} = \frac{1}{p} \sum_{i=1}^{p} (V_i^{\mathbf{HG}} - \mathbf{MA}_i)^2. \tag{5.5}$$

Conversely, tissue guidance (**TG**) assists the model in distinguishing relevant tissue regions from background or artifacts. We use a specialized loss, L_{focus}, that penalizes misplaced attention and rewards correct focus:

$$L_{focus} = \frac{1}{p}(- \sum_{i, V_i^{\mathbf{TG}} > 0}^{p} \mathbf{MA}_i + \sum_{i, V_i^{\mathbf{TG}} = 0}^{p} \mathbf{MA}_i). \tag{5.6}$$

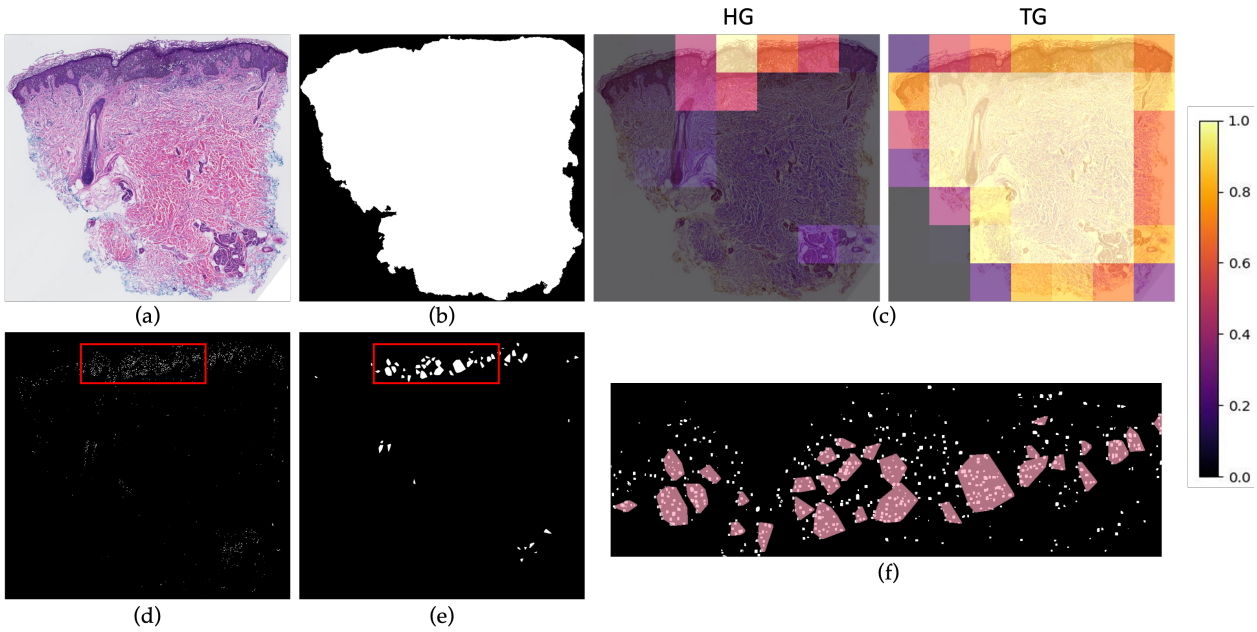


Figure 5.3: Generation of attention guidance: (a) H&E sample image. (b) Tissue segmentation mask. (c) **HG** and **TG**. The values are normalized for visualization purpose. (d) Cellular entities detected (**zoom-in for best view**). (e) Convex hull of cellular clusters. (f) A zoomed-in view of the red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

To integrate these losses effectively, we apply an uncertainty weighting strategy, $\mathcal{UW}$ [106], which adapts the impact of each loss function based on each task's specific uncertainties. The comprehensive loss formula is:

$$L = \mathcal{UW} \otimes \{L_{cls}, L_{mse}, L_{focus}\}, \tag{5.7}$$

where L_{cls} is the cross-entropy loss used for the classification component.

5.4 Implementation Details

To study the generalization ability of SAG across public and in-house datasets, the Camelyon16 Dataset (Chapter 2.3) and the Skin Biopsy Dataset (Chapter 2.2) were used to demon-

strate the generalizability of `SAG`.

5.4.1 Skin Biopsy Dataset

Feature Extraction Similar to `ScATNet` (Chapter 4.3.3), an ImageNet pre-trained MobileNetV2 model [85] extracts a 1280-dimensional feature vector for each patch within the whole slide images (WSIs).

Semantic Guidance `VSGDNet` (described in Chapter 3.2) [24] to generate a melanocyte map. This map is further processed as detailed in Chapter 5.3.3 to contribute to the final heuric guidance **HG**. Additionally, DBSCAN clustering (scikit-learn package [107]) with pre-defined parameters (`eps=20` and `min_samples=5`) is used to group cell entities. Finally, tissue guidance **TG** is generated using Otsu's thresholding method [103].

5.4.2 Camelyon16 Dataset

Feature Extraction A SimCLR model pre-trained by DSMIL [108] extracts a 512-dimensional feature vector for each patch within the lymph node WSIs.

Semantic Guidance Unlike the melanoma dataset, the Camelyon16 dataset provides pre-annotated information directly relevant to the diagnostic task. We leverage the existing metastasis mask and tissue mask within the dataset to construct both**HG**and TG.

5.4.3 Diagnosis Models and Training Details

The proposed Semantics-Aware Attention Guidance (SAG) framework is applied to two models: a transformer model, ScAtNet (Chapter 4.3.3) [4], and a multi-instance learning (MIL) model, ABMIL [102].

ScAtNet We incorporate **TG** across all attention heads and **HG** on half of the attention heads. This approach maintains the model's adaptability while mitigating potential noise in

HG.

ABMIL For the melanoma dataset, we apply both **HG** and **TG**. For Camelyon16, we only apply **HG** as the dataset already excludes background patches.

5.5 Results

This section evaluates the effectiveness of the Semantics-Aware Attention Guidance (SAG) framework in enhancing diagnostic performance across different datasets and model architectures.

Quantitative Evaluation Table 5.1 summarizes the overall performance of SAG on both the melanoma and Camelyon16 datasets, employing ScAtNet [4] and ABMIL [102] as backbone models. For each configuration, we conduct 15 experiment runs with randomly chosen seeds and report the average accuracy.

The results demonstrate consistent improvements in diagnostic accuracy when incorporating SAG. For the melanoma dataset, significant gains are observed when using SAG with both single-scale and multi-scale ScAtNet models. The multi-scale configuration achieves the most notable improvement, with a 4.55% increase in accuracy (Table 5.1). Similar trends are evident on Camelyon16, where SAG boosts accuracy across all ScAtNet configurations (reaching a 3.81% improvement for multi-scale inputs) and also increases ABMIL's accuracy by 1.71% (Table 5.1). These findings highlight SAG's effectiveness in refining the model's focus on diagnostically relevant information, ultimately leading to enhanced performance.

Our analysis reveals an interesting observation regarding model performance on the two datasets. ABMIL exhibits superior diagnostic accuracy on Camelyon16 (94.73% vs. 71.60% for `ScATNet`) (Table 5.1). Conversely, ScAtNet outperforms ABMIL on the melanoma dataset (62.71% vs. 45.52%). This distinction can be attributed to the inherent characteristics of the datasets and the strengths of each model architecture. The skin biopsy dataset, presenting a four-class classification problem, benefits from a comprehensive understanding of the en-

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)†[4]			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†[102]			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**

tire image at various scales. This aligns well with ScATNet's transformer-based architecture, which excels at capturing long-range dependencies and aggregating multi-scale information through attention mechanisms [4]. In contrast, Camelyon16 focuses on a binary classification task, prioritizing local feature identification for accurate diagnosis. This characteristic aligns better with ABMIL's MIL-based approach, explaining its superior performance in this context. Furthermore, ScATNet's complexity and multi-scale processing might introduce overfitting risks on the smaller Camelyon16 dataset, potentially limiting its benefit. These observations emphasize the importance of selecting an appropriate model based on the specific characteristics of the data at hand.

Qualitative Evaluation Visualization of attention maps (Figure 5.1) illustrates SAG's impact: through tissue attention guidance, the model focuses on diagnostically relevant areas,

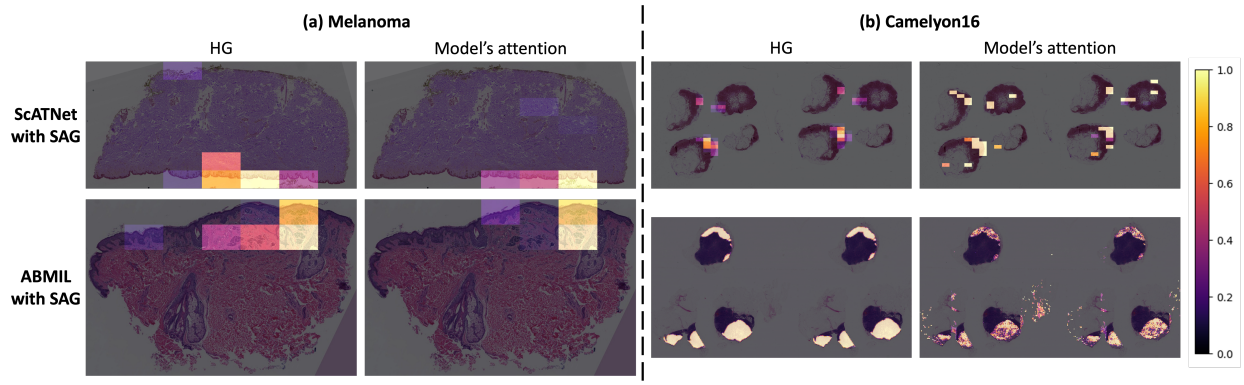


Figure 5.4: Comparative visualizations of **HG** and the models' attention under `SAG`'s training on the melanoma and Camelyon16 datasets. The images are sampled from test set. The **HG** and attention values are normalized for visualization purpose.

avoiding background patches. This process demonstrates `SAG` enhance model performance by ensuring that attention is selectively concentrated on regions essential for diagnosing, thereby refining the model's focus and improving its diagnostic capability.

5.6 Conclusion

In conclusion, the `SAG` network significantly enhances the performance of ScAtNet models by improving attention guidance, regardless of input scales. Its application leads to a more precise and reliable histopathological image analysis, bridging the gap between computational models and the nuanced diagnostic processes of pathologists. By facilitating a more accurate identification of cancerous regions, `SAG` emerges as a pivotal tool in the advancement of automated diagnostic systems.

Chapter 6

CONCLUSION

This dissertation makes significant contributions to the field of digital pathology by advancing the analysis and diagnosis of skin and breast cancer through whole slide images. It introduces segmentation techniques that enhance the detection of crucial diagnostic entities, paving the way for more accurate and efficient analysis with minimal annotation requirements. Building on this foundation, the thesis explores the application of transformer-based models, which harness the power of deep learning to manage the vast data inherent in WSIs, capturing essential details at multiple scales. The culmination of this work is the development of the Semantics-Aware Attention Guidance (`SAG`) framework, which innovatively integrates segmentation insights into an attention-driven diagnostic model. This integration significantly improves the model's focus and diagnostic accuracy, demonstrating the potential of machine learning to revolutionize medical diagnostics by making them more precise, adaptable, and efficient.

Segmentation of Diagnostic Entities in Whole Slide Images Chapter 3 introduced two deep learning approaches to address segmentation challenges in WSI analysis. VSGD-Net [24] tackles melanocyte identification in skin cancer images by leveraging virtual staining and knowledge transfer. The *Skin Biopsy Segmentation* project [101] tackles limited annotation data by employing a two-stage segmentation approach, achieving promising results with minimal supervision.

Transformer Network for Diagnosis Chapter 4 explored the application of transformer networks for breast cancer diagnosis. `HATNet` [25] utilizes a holistic attention mechanism to directly learn image representations from WSIs, capturing global information and clinically

relevant structures. Building upon this success, `ScATNet` [4] was developed to learn multi-scale representations directly from WSIs, capturing both fine-grained details and broader contextual information.

Semantics-Aware Attention Guidance Chapter 5 introduced a novel framework, `SAG` [26], that improves upon existing methods. `SAG` incorporates two key elements: Entity-to-Attention Conversion, which transforms diagnostically relevant entities into attention signals, and a Flexible Attention Loss Function, which optimizes model learning by leveraging the relationships between these entities and the overall diagnosis task. `SAG` demonstrably improves accuracy, precision, and recall compared to state-of-the-art models on independent cancer datasets.

These separate projects highlight the effectiveness of deep learning in WSI analysis and pave the way for further advancements in automated cancer diagnosis.

6.1 Limitations

While the projects presented in this dissertation demonstrate promising results, several limitations must be acknowledged, highlighting areas for further research and improvement:

Limited Datasets Access to larger and more comprehensive datasets could significantly enhance the generalizability and robustness of the proposed methods. The datasets employed, while diverse, still pose limitations in terms of size and scope. Particularly, `ScATNet` was studied using a specific subset of skin biopsies—melanocytic lesions, which constitute only about one in four skin biopsies [109]. Moreover, the test set, although independent, comprises only 115 WSIs, which may not adequately represent the wider clinical variety found in routine diagnostics. `HATNet` was also studied using a private breast cancer dataset with 240 samples. The generalizability of discussed methods in this dissertation to other biopsy types, such as breast and lung, while theoretically feasible, remains to be rigorously tested across more extensive and varied datasets.

Interpretability The Semantics-Aware Attention Guidance (SAG) framework enhances the interpretability of diagnostic models by directing attention to clinically significant regions. However, there is a need for further development in making these models fully transparent. Enhancing interpretability involves creating mechanisms within deep learning models that can provide clear, understandable explanations for their diagnostic decisions, thereby increasing their acceptance and trust among healthcare professionals.

Model Generalization and Bias The Semantics-Aware Attention Guidance (SAG) framework demonstrates promising capabilities in learning to diagnose various cancers. However, its performance on unseen, highly heterogeneous datasets or rare cancer types has not been extensively validated. Future work should focus on testing and refining these models across a broader range of pathological conditions to ensure their reliability and effectiveness in real-world clinical settings.

While computer-aided diagnosis systems have the potential to significantly improve patient outcomes by providing more accurate and consistent diagnoses, there are also potential risks that must be considered. The reliability of the training data is another critical factor in the performance of diagnostic models. Although generally reliable, models trained on consensus diagnoses can still be subject to biases and inconsistencies. Studies have shown that pathologists can disagree on diagnoses in up to 60% of cases [8], indicating the potential noisiness of the diagnostic labels. Variability and bias in human annotations, often used as ground truth in model training, compound this issue. These factors raise important questions about the validity of the learned models, as there is a significant risk they may inherit these biases, leading to skewed results and errors over time. Moreover, over-reliance on automated systems might result in decreased vigilance among pathologists, potentially exacerbating the consequences of any errors. Furthermore, unlike humans who understand the severe consequences of a misdiagnosis, holding the model accountable for the actual diagnosis is not a trivial problem itself. Ultimately, while CAD offers significant potential, addressing these challenges is paramount to ensure it enhances patient care without introducing new risks

through biased algorithms or decreased human vigilance.

Computational Demands The computational intensity of the methods discussed, particularly those involving deep learning for processing gigapixel WSIs, requires significant computational resources. This demand may restrict the accessibility of these advanced models to well-resourced institutions, potentially limiting their adoption in lower-resource settings. Efforts to optimize computational efficiency or develop more streamlined models could help mitigate this issue, making advanced diagnostic tools more accessible across a broader range of clinical environments.

6.2 Future Directions

This dissertation opens several avenues for extending the research on deep learning applications in digital pathology. By incorporating broader data sets and integrating these models into clinical workflows, the potential for these technologies to enhance diagnostic processes and patient care could be significantly realized.

Expanding Data Modalities Future research could greatly benefit from incorporating a wider array of biomedical data across multiple modalities. This includes demographic details such as patient age, gender, and race, along with genetic information which could provide crucial insights into the personalization of treatment plans. Integrating such diverse data sources may enhance the models' diagnostic accuracy and offer a more holistic view of patient health, aiding in personalized medicine.

Model Training with Diverse Datasets Continuing to expand the datasets used in training these models is essential. Utilizing datasets of varying sizes and from different cancer types can enhance the robustness and generalizability of the models. This diversification in training data would help address one of the significant challenges in computational pathology—working efficiently with small datasets and noisy labels. By improving mod-

els' capabilities to learn from limited or imperfect data, the dependency on large annotated datasets could be reduced, simplifying the data collection process.

Integration with Clinical Workflows For deep learning models to be practically applicable in medical settings, they must be seamlessly integrated into existing clinical workflows. This involves not only ensuring computational efficiency and developing user-friendly interfaces but also navigating regulatory hurdles. Effective integration requires collaboration between technologists, clinicians, and regulatory bodies to ensure that the models are both effective and compliant with medical standards.

Multi-Modal Learning Enhancement Exploring the integration of additional data modalities should go beyond initial patient demographics to include clinical history, pathology reports, and genetic markers. By synthesizing information from these varied sources, deep learning models could offer more precise diagnostic insights and tailor treatment plans to individual patient profiles, thus driving forward the personalized medicine initiative.

Foundation Models and Generalizability To address the limitations in generalizability, future research could explore the use of foundation models. These models, pre-trained on large and diverse datasets, have the potential to capture a wide range of patterns, thereby mitigating bias and noise in the training data. By fine-tuning these models for specific tasks, such as diagnosing certain types of cancer, we could further enhance their performance and generalizability.

However, several challenges must be carefully considered, including the substantial computational resources required, issues with model interpretability, and the risk of transferring biases from the pre-training data. Despite these challenges, the potential of foundation models to improve the generalizability and robustness of diagnostic models presents an exciting avenue for future research.

Longitudinal Data Analysis Developing models that leverage longitudinal studies, which monitor patient outcomes and disease progression over time, could transform patient management and follow-up care. These models would provide healthcare professionals with tools to predict disease trajectories, evaluate treatment efficacy, and adjust management plans in real-time, enhancing both immediate and long-term patient outcomes.

Computational Efficiency and Accessibility Improving the computational efficiency of these models is crucial for their adoption in diverse clinical settings, including those with limited resources. Research aimed at reducing the computational demands of processing WSIs could broaden the accessibility of advanced diagnostic tools, ensuring that benefits of digital pathology can be experienced more universally.

By addressing these directions, the field of digital pathology can leverage deep learning to not only enhance diagnostic accuracy and efficiency but also significantly impact patient care by enabling early detection and personalized treatment strategies.

BIBLIOGRAPHY

[1] 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.

[2] 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.

[3] Shima Nofallah, Mojgan Mokhtari, Wenjun Wu, Sachin Mehta, Stevan Knezevich, Caitlin J May, Oliver H Chang, Annie C Lee, Joann G Elmore, and Linda G Shapiro. Segmenting skin biopsy images with coarse and sparse annotations using u-net. *Journal of Digital Imaging*, pages 1–12, 2022.

[4] 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.

[5] Wendy A Wells, Patricia A Carney, M Scottie Eliassen, Anna N Tosteson, and E Robert Greenberg. Statewide study of diagnostic agreement in breast pathology. *JNCI: Journal of the National Cancer Institute*, 90(2):142–145, 1998.

[6] 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.

[7] 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.

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

[9] 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.

[10] Lourdes Duran-Lopez, Juan P Dominguez-Morales, Antonio Felix Conde-Martin, Saturnino Vicente-Diaz, and Alejandro Linares-Barranco. Prometeo: A cnn-based computer-aided diagnosis system for wsi prostate cancer detection. *IEEE Access*, 8:128613–128628, 2020.

[11] Ilhame Ait Lbachir, Imane Daoudi, and Saadia Tallal. Automatic computer-aided diagnosis system for mass detection and classification in mammography. *Multimedia Tools and Applications*, 80(6):9493–9525, 2021.

[12] Heang-Ping Chan, Lubomir M Hadjiiski, and Ravi K Samala. Computer-aided diagnosis in the era of deep learning. *Medical physics*, 47(5):e218–e227, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] 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.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[17] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[20] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann Elmore, and Linda Shapiro. Learning to segment breast biopsy whole slide images. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 663–672. IEEE, 2018.

[21] Xiaofan Zhang, Hai Su, Lin Yang, and Shaoting Zhang. Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5361–5368, 2015.

[22] Beibin Li, Ezgi Mercan, Sachin Mehta, Stevan Knezevich, Corey W Arnold, Donald L Weaver, Joann G Elmore, and Linda G Shapiro. Classifying breast histopathology images with a ductal instance-oriented pipeline. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8727–8734. IEEE, 2021.

[23] Lyndon Chan, Mahdi S Hosseini, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10662–10671, 2019.

[24] 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.

[25] Sachin Mehta, Ximing Lu, Wenjun Wu, Donald Weaver, Hannaneh Hajishirzi, Joann G Elmore, and Linda G Shapiro. End-to-end diagnosis of breast biopsy images with transformers. *Medical image analysis*, 79:102466, 2022.

[26] Kechun Liu, Wenjun Wu, Joann G Elmore, and Linda G Shapiro. Semantics-aware attention guidance for diagnosing whole slide images. *arXiv preprint arXiv:2404.10894*, 2024.

[27] Patricia A Carney, Lisa M Reisch, Michael W Piepkorn, Raymond L Barnhill, David E Elder, Stevan Knezevich, Berta M Geller, Gary Longton, and Joann G Elmore. Achieving consensus for the histopathologic diagnosis of melanocytic lesions: use of the modified delphi method. *Journal of cutaneous pathology*, 43(10):830–837, 2016.

[28] 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):131–141, 2014.

[29] 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 learn-

ing algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[30] Rodney L Custer, Joseph A Scarcella, and Bob R Stewart. The modified delphi technique-a rotational modification. 1999.

[31] Caner Mercan, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. *IEEE transactions on medical imaging*, 37(1):316–325, 2017.

[32] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann Elmore, and Linda Shapiro. Learning to segment breast biopsy whole slide images. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 663–672, 2018.

[33] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann G. Elmore, and Linda Shapiro. Y-net: Joint segmentation and classification for diagnosis of breast biopsy images. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 893–901, Cham, 2018. Springer International Publishing.

[34] Baris Gecer, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern recognition*, 84:345–356, 2018.

[35] Ezgi Mercan, Sachin Mehta, Jamen Bartlett, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions. *JAMA network open*, 2(8):e198777–e198777, 2019.

[36] Ezgi Mercan, Linda G Shapiro, Tad T Brunyé, Donald L Weaver, and Joann G Elmore. Characterizing diagnostic search patterns in digital breast pathology: scanners and drillers. *Journal of digital imaging*, 31(1):32–41, 2018.

[37] 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*, 2020.

[38] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[39] 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.

[40] 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.

[41] 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.

[42] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image trans-

lation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[43] 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.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[45] 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.

[46] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[47] 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.

[48] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 265–273. Springer, 2018.

[49] 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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 132–142. Springer, 2021.

[50] 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.

[51] 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.

[52] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When unseen domain generalization is unnecessary? rethinking data augmentation. *arXiv preprint arXiv:1906.03347*, 2019.

[53] Jamshid Sourati, Ali Gholipour, Jennifer G Dy, Sila Kurugol, and Simon K Warfield. Active deep learning with fisher information for patch-wise semantic segmentation. In *International Workshop on Deep Learning in Medical Image Analysis*, pages 83–91. Springer, 2018.

[54] Jwan Saeed and Subhi Zeebaree. Skin lesion classification based on deep convolutional neural networks architectures. *Journal of Applied Science and Technology Trends*, 2(01):41–51, 2021.

[55] Ashish Vaswanica, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett,

editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[56] Jaafar Makki. Diversity of breast carcinoma: histological subtypes and clinical relevance. *Clinical medicine insights: Pathology*, 8:CPath–S31563, 2015.

[57] Kenneth A Kern. The delayed diagnosis of breast cancer: medicolegal implications and risk prevention for surgeons. *Breast disease*, 12(1):145–158, 2001.

[58] Lisa M Reisch, Patricia A Carney, Natalia V Oster, Donald L Weaver, Heidi D Nelson, Paul D Frederick, and Joann G Elmore. Medical malpractice concerns and defensive medicine: a nationwide survey of breast pathologists. *American journal of clinical pathology*, 144(6):916–922, 2015.

[59] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. *CA: a Cancer Journal for Clinicians*, 71(1):7–33, 2021.

[60] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528. Springer, 2020.

[61] 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.

[62] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[63] 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.

[64] Jon N Marsh, Ta-Chiang Liu, Parker C Wilson, S Joshua Swamidass, and Joseph P Gaut. Development and validation of a deep learning model to quantify glomerulosclerosis in kidney biopsy specimens. *JAMA network open*, 4(1):e2030939–e2030939, 2021.

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

[66] Haomiao Ni, Hong Liu, Kuansong Wang, Xiangdong Wang, Xunjian Zhou, and Yueliang Qian. Wsi-net: Branch-based and hierarchy-aware network for segmentation and classification of breast histopathological whole-slide images. In *International Workshop on Machine Learning in Medical Imaging*, pages 36–44. Springer, 2019.

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

[68] Hans Pinckaers, Wouter Bulten, Jeroen Van der Laak, and Geert Litjens. Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels. *IEEE transactions on medical imaging*, PP, March 2021.

[69] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

[70] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[71] Tom Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[72] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[73] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[74] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv preprint arXiv:2102.10662*, 2021.

[75] Zhuangzhuang Zhang, Baozhou Sun, and Weixiong Zhang. Pyramid medical transformer for medical image segmentation. *arXiv preprint arXiv:2104.14702*, 2021.

[76] Yinglin Zhang, Risa Higashita, Huazhu Fu, Yanwu Xu, Yang Zhang, Haofeng Liu, Jian Zhang, and Jiang Liu. A multi-branch hybrid transformer networkfor corneal endothelial cell segmentation. *arXiv preprint arXiv:2106.07557*, 2021.

[77] Olivier Petit, Nicolas Thome, Clément Rambour, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. *arXiv preprint arXiv:2103.06104*, 2021.

[78] Chiranjibi Sitaula and Mohammad Belayet Hossain. Attention-based vgg-16 model for covid-19 chest x-ray image classification. *Applied Intelligence*, 51(5):2850–2863, 2021.

[79] 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 conference on computer vision and pattern recognition*, 2021.

[80] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Application of dimensionality reduction in recommender system-a case study. Technical report, Minnesota Univ Minneapolis Dept of Computer Science, 2000.

[81] Ruizhen Liu and Tieniu Tan. An svd-based watermarking scheme for protecting rightful ownership. *IEEE transactions on multimedia*, 4(1):121–128, 2002.

[82] Chin-Chen Chang, Piyu Tsai, and Chia-Chen Lin. Svd-based digital image watermarking scheme. *Pattern Recognition Letters*, 26(10):1577–1586, 2005.

[83] Mehrbakhsh Nilashi, Othman Ibrahim, and Karamollah Bagherifard. A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. *Expert Systems with Applications*, 92:507–520, 2018.

[84] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[85] 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.

[86] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[87] Hugh Chen, Scott Lundberg, and Su-In Lee. Checkpoint ensembles: Ensemble methods from a single training process. *arXiv preprint arXiv:1710.03282*, 2017.

[88] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna.

Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[89] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.

[90] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[91] Lieve Brochez, Evelien Verhaeghe, Edouard Grosshans, Eckhart Haneke, Gérald Piérard, Dirk Ruiter, and Jean-Marie Naeyaert. Inter-observer variation in the histopathological diagnosis of clinically suspicious pigmented skin lesions. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 196(4):459–466, 2002.

[92] MG Cook, TJ Clarke, S Humphreys, A Fletcher, KM McLaren, NP Smith, A Stevens, JM Theaker, and J Melia. The evaluation of diagnostic and prognostic criteria and the terminology of thin cutaneous malignant melanoma by the crc melanoma pathology panel. *Histopathology*, 28(6):497–512, 1996.

[93] 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.

[94] 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.

[95] Claudia Mello-Thoms, Carlos AB Mello, Olga Medvedeva, Melissa Castine, Elizabeth Legowski, Gregory Gardner, Eugene Tseytlin, and Rebecca Crowley. Perceptual anal-

ysis of the reading of dermatopathology virtual slides by pathology residents. *Archives of pathology & laboratory medicine*, 136(5):551–562, 2012.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] Shima Nofallah, Beibin Li, Mojgan Mokhtari, Wenjun Wu, Stevan Knezevich, Caitlin J May, Oliver H Chang, Joann G Elmore, and Linda G Shapiro. Improving the diagnosis of skin biopsies using tissue segmentation. *Diagnostics*, 12(7):1713, 2022.

[102] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple

instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] Jason P Lott, Denise M Boudreau, Ray L Barnhill, Martin A Weinstock, Eleanor Knopp, Michael W Piepkorn, David E Elder, Steven R Knezevich, Andrew Baer, Anna NA Tosteson, et al. Population-based analysis of histologically confirmed melanocytic proliferations using natural language processing. *JAMA dermatology*, 154(1):24–29, 2018.

[110] NationalCancerInstitute. What is cancer? - nci. `https://www.cancer.gov/about-cancer/understanding/what-is-cancer#types`, May 2021. (Accessed on 05/05/2022).

[111] Kechun Liu, Wenjun Wu, Joann Elmore, and Linda Shapiro. Semantics-aware attention guidance for diagnosing whole slide images. Submitted for publication to *International Conference on Medical Image Computing and Computer-Assisted Intervention* March 2024.

[112] António Polónia, Sofia Campelos, Ana Ribeiro, Ierece Aymore, Daniel Pinto, Magdalena Biskup-Fruzynska, Ricardo Santana Veiga, Rita Canas-Marques, Guilherme Aresta, Teresa Araújo, et al. Artificial intelligence improves the accuracy in histologic classification of breast lesions. *American journal of clinical pathology*, 155(4):527–536, 2021.

[113] 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 (CVPR) Workshops*, pages 3766–3775, June 2021.

[114] 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.

[115] 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.

[116] 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.

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

[118] 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.

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

[120] Zhaoyang Xu, Xingru Huang, Carlos Fernández Moro, Béla Bozóky, and Qianni Zhang. Gan-based virtual re-staining: a promising solution for whole slide image analysis. *arXiv preprint arXiv:1901.04059*, 2019.

[121] Dan C. Ciresan et al. Mitosis detection in breast cancer histology images with deep neural networks. *MICCAI*, 2013.

[122] Angel Cruz-Roa et al. Automatic annotation of histopathological slides. In *MICCAI*, 2014.

[123] Jun Xu et al. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE Transactions on Medical Imaging*, 2015.

[124] Le Hou et al. Patch-based convolutional neural network for whole slide tissue image classification. *CVPR*, 2016.

[125] Caner Mercan et al. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. *IEEE Transactions on Medical Imaging*, 2017.

[126] Gabriele Campanella et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 2019.

[127] Ming Lu et al. A novel framework for medical image classification using multi-instance multi-label learning. *Medical Image Analysis*, 2021.

[128] Ozan Oktay et al. Attention u-net: Learning where to look for the pancreas. In *MIDL*, 2018.

[129] Leonardo Rundo et al. Use of deep learning to develop continuous-learning, precision medicine approaches to cancer treatment. *Precision Oncology*, 2019.

[130] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[131] Hugo Touvron et al. Training data-efficient image transformers & distillation through attention. *ICML*, 2021.

[132] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[133] Michal Heker and Hayit Greenspan. Joint liver lesion segmentation and classification via transfer learning. *arXiv preprint arXiv:2004.12352*, 2020.

[134] 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.