IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Scale-Aware Transformers for Diagnosing Melanocytic Lesions

Wenjun Wu[1], Sachin Mehta[1], Shima Nofallah[1], Stevan Knezevich[2], Caitlin J. May[3], Oliver H. Chang[1], Joann G. Elmore[4] and Linda G. Shapiro[1]

[1]University of Washington, Seattle, WA, 98195, USA (e-mail: wenjunw@uw.edu; sacmehta@uw.edu; shimz@uw.edu; ochang@uw.edu; shapiro@cs.washington.edu)
[2]Pathology Associates, Clovis, CA, 93611 (e-mail: stevanrk@gmail.com)
[3]Dermatopathology Northwest, Bellevue, WA 98005 (e-mail: campbell.cait@gmail.com)
[4]David Geffen School of Medicine, UCLA, Los Angeles CA 90024, USA (e-mail: JElmore@mednet.ucla.edu)

Corresponding author: Wenjun Wu (e-mail: wenjunw@uw.edu).

Research reported in this study was supported by grants R01CA200690 and U01CA231782 from the National Cancer Institute of the National Institutes of Health, 622600 from the Melanoma Research Alliance, and W81XWH-20-1-0798 from the United States Department of Defense. The funders had no role in the design and conduct of the study, collection, management, analysis, and interpretation of the data, preparation, review, or approval of the manuscript, nor decision to submit the manuscript for publication. (Wenjun Wu and Sachin Mehta contributed equally to this work.)

ABSTRACT Diagnosing melanocytic lesions is one of the most challenging areas of pathology with extensive intra- and inter-observer variability. The gold standard for a diagnosis of invasive melanoma is the examination of histopathological whole slide skin biopsy images by an experienced dermatopathologist. Digitized whole slide images offer novel opportunities for computer programs to improve the diagnostic performance of pathologists. In order to automatically classify such images, representations that reflect the content and context of the input images are needed. In this paper, we introduce a novel self-attention-based network to learn representations from digital whole slide images of melanocytic skin lesions at multiple scales. Our model softly weighs representations from multiple scales, allowing it to discriminate between diagnosis-relevant and -irrelevant information automatically. Our experiments show that our method outperforms five other state-of-the-art whole slide image classification methods by a significant margin. Our method also achieves comparable performance to 187 practicing U.S. pathologists who interpreted the same cases in an independent study. To facilitate relevant research, full training and inference code is made publicly available at https://github.com/meredith-wenjunwu/ScAtNet.

INDEX TERMS Convolutional Neural Network, Histopathological Images, Melanocytic Risk Lesions, Melanoma, Multi-scale, Transformers, Skin Cancer Diagnosis, Whole-slide Image Classification

I. INTRODUCTION

INVASIVE melanoma, with more than 100,000 estimated new cases in 2021, is one of the most commonly diagnosed cancers in the U.S [1]. The "gold standard" for diagnosis of skin biopsy specimens relies on the visual assessments of pathologists. Unfortunately, diagnostic errors are common, and even expert pathologists may not reach consensus on diagnostically challenging cases in many areas within pathology [2]–[5]. For instance, pathologists disagree in up to 60% of melanoma in situ and stage T1a invasive cases [6]. Variability in diagnostic decisions is a serious problem and can cause substantial patient harm. A computer-aided diagnostic system can act as a *second reader* and help pathologists reduce classification uncertainties.

For a reliable diagnostic system, it is important to obtain representations that reflect both the content and context of the input biopsy image. This paper introduces a self-attention-based deep neural network called the Scale-Aware Transformer Network (`ScAtNet`) for classifying melanocytic skin lesions in digital whole slide images (WSIs). `ScAtNet`, shown in Figure 1, extends the standard transformer model of Vaswani *et al.* (2017) to learn representations from biopsy images at multiple input scales. The key idea is to learn patch-wise representations independently for each input scale using a convolutional neural network (CNN), and then learn inter-patch and inter-scale representations from concatenated multi-scale contextualized patch embeddings using transformers. This allows our system to learn diagnostic class-

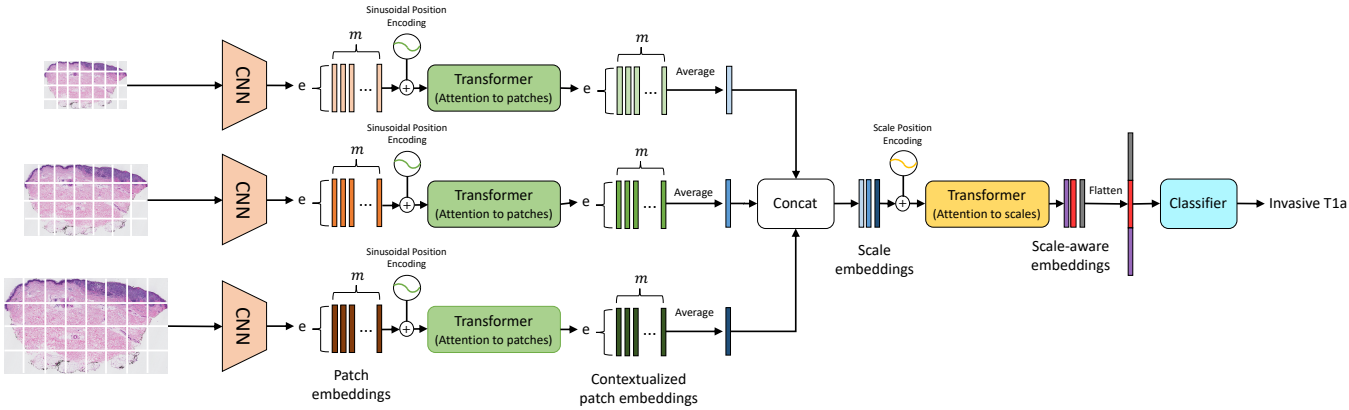


FIGURE 1: Overview of ScAtNet for classifying skin biopsy images. To learn representations from these large WSIs at multiple input scales in an end-to-end fashion, ScAtNet factorizes the classification pipeline into three steps. The first step involves learning local patch-wise embeddings using an off-the-shelf 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.

specific representations at different scales and helps improve the performance. Also, each WSI contains multiple tissue slices, while usually only one or two tissue slices help pathologists in diagnosis. We introduce a soft-label assignment method to (1) reduce the ambiguity between different tissue slices in a WSI and (2) improve the diagnostic classification performance.

We demonstrate the effectiveness of ScAtNet on a skin biopsy image dataset [6]. Experimental results show that ScAtNet outperforms state-of-the-art methods by a significant margin. For example, ScAtNet is 8% more accurate than the method proposed by Chikontwe *et al.* [7] and 6% more accurate than the method proposed by Hashimoto *et al.* [8]. Importantly, ScAtNet delivers comparable performance to 187 practicing pathologists who interpreted the same test set cases in an independent study.

To summarize, the main contributions of this paper are: (1) a novel self-attention-based end-to-end framework for classifying WSIs at multiple input scales (Section III-B), (2) a soft label assignment method to reduce ambiguities that arise by assigning the same label to all tissue slices in a WSI (Section III-C), and (3) experimental results, along with comparisons with state-of-the-art methods and practicing U.S. pathologists, demonstrating ScAtNet's competitive performance (Section IV).

II. RELATED WORK

ScAtNet was inspired by the success of several works in the area of WSI image classification and transformers. We briefly discuss these approaches in the following sub-sections.

Multiple instance learning (MIL). Convolutional neural networks (CNNs) are the de facto machine learning-based method for image classification, including WSIs [9]–[11]. Unlike the images in standard datasets (e.g., ImageNet [12]), 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 [11], kidney [13], and breast [14]. In general, 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 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. This work extends the MIL framework with the transformers of Vaswani *et al.* (2017) to learn global representations in an end-to-end fashion. In our experiments, we compared our method to the MIL methods of Chikontwe et al. [7] and Hashimoto et al. [8]. In addition, we compared our system to a standard patch-based CNN classification framework. Details of these methods are described in section IV-D.

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 [11], [15] 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. In our experiments, we compared our method

with a CNN-based deep-feature-aggregation framework developed by C. Mercan et al [15]. Details of this method are described in section IV-D.

Segmentation-based methods. These approaches use semantic information about tissues in a WSI to produce an image-level decision [16]–[20]. 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, this work introduces 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 [21], [22]. 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 [9] and gradient check-pointing [23], 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, Mehta *et al.* [21] uses the transformers of Vaswani *et al.* (2017) to aggregate the information from all instances in a breast biopsy image, while Pinckaers *et al.* [22] stitches the instance-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 $10\times$) 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. [22], details of this which are described in section IV-D.

Vision Transformers. The transformers of Vaswani *et al.* [24], initially introduced for the task of machine translation (e.g., [25], [26]), are being explored for modeling images and computer vision tasks (e.g., [27], [28]). 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.* [29] extended the standard encoder-decoder network of Vaswani *et al.* [24] for the task of object detection. Recent work has extended transformers using a patch-based approach to image recognition at a large scale [27], [28]. Concurrent work has also utilized transformers and self-attention to medical image segmentation [30]–[33] and classification [34].

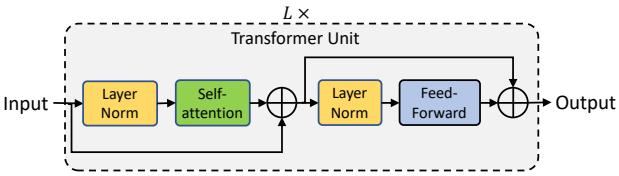


FIGURE 2: The transformer network stacks L transformer units sequentially. Each transformer unit consists of self-attention and feed-forward modules.

Motivated by (1) the success of transformers in vision, (2) the methods for learning representations from different input scales [35]–[37], and (3) the importance of input scales for diagnosis in clinical settings [38], [39], we propose a scale-aware transformer model that adapts to the information from different input scales using self-attention and predicts the classification label.

III. METHOD

This section first reviews the architecture of transformers and then elaborates on the details of the proposed method, scale-aware transformers (Section III-B), that allows our system to learn representations from histopathological images at multiple scales in an end-to-end fashion. In Section III-C, a soft-labeling method is discussed that reduces the ambiguity in instance-level (patches) labels and improves the learning of representations from skin-biopsy images. The software associated with this work will be made available.

A. TRANSFORMERS

The transformer unit, shown in Figure 2, is comprised of two modules: (1) self-attention and (2) feed-forward. The self-attention module allows the inputs to interact with each other and learn contextual relationships. This layer applies three projections, with each projection branch having multiple linear layers to the input $\mathbf{I} \in \mathbb{R}^{n\times e}$ to produce query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) embeddings, where n is the number of inputs and e is the input dimensionality. A dot-product between query ($\mathbf{Q}$) and key ($\mathbf{K}$) is computed to produce an $n \times n$ matrix to which a row-wise softmax is applied to encode relationships between the n inputs. Finally, a weighted sum is computed between the resultant $n \times n$ matrix and V.

$$\text{Self-attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\mathbf{Q} \cdot \mathbf{K}^T) \cdot \mathbf{V} \qquad (1)$$

The feed-forward module stacks two linear layers, and is responsible for learning wider representations. The first linear layer projects the input to a high-dimensional space, while the second linear layer projects from the high-dimensional space to the same dimensionality as that of the

input. This work extends the transformers model to learn scale-aware representations from skin biopsy images.

B. SCALE-AWARE TRANSFORMERS

Patch-based CNNs are state-of-the-art WSI classification methods that allow computer systems to learn representations from gigapixel size images (e.g. [11], [13], [14], [16], [40]). 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 [41] (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 1 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:

$$\mathbf{PE}_i^{sc} = \text{CNN}\left(\mathbf{x}_i^{sc}\right) \tag{2}$$

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 relationships. Similar to vision transformers [27], 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 $\mathbf{CPE}^{sc} \in \mathbb{R}^{m \times e}$.

$$\mathbf{CPE}^{sc} = \text{Transformer}\left(\mathbf{PE}^{sc} + \mathbf{PPE}^{sc}\right) \tag{3}$$

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 $\mathbf{CSE} \in \mathbb{R}^{sc \times e}$.

$$\mathbf{CSE} = \text{Transformer}\left(\mathbf{SE} + \mathbf{PSE}\right) \tag{4}$$

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.

C. SOFT-LABELS FOR SKIN BIOPSY IMAGES

Skin biopsy images often contain multiple tissue slices on a single WSI, as shown in Figure 4a. In general, the representative regions-of-interest (ROIs; shown in red in Figure 4a) 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 3.

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 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:

$$\bar{\mathbf{s}}_i = \frac{1}{n} \sum_{j=1}^{n} \mathbf{s}_i^j \tag{5}$$

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

[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%.

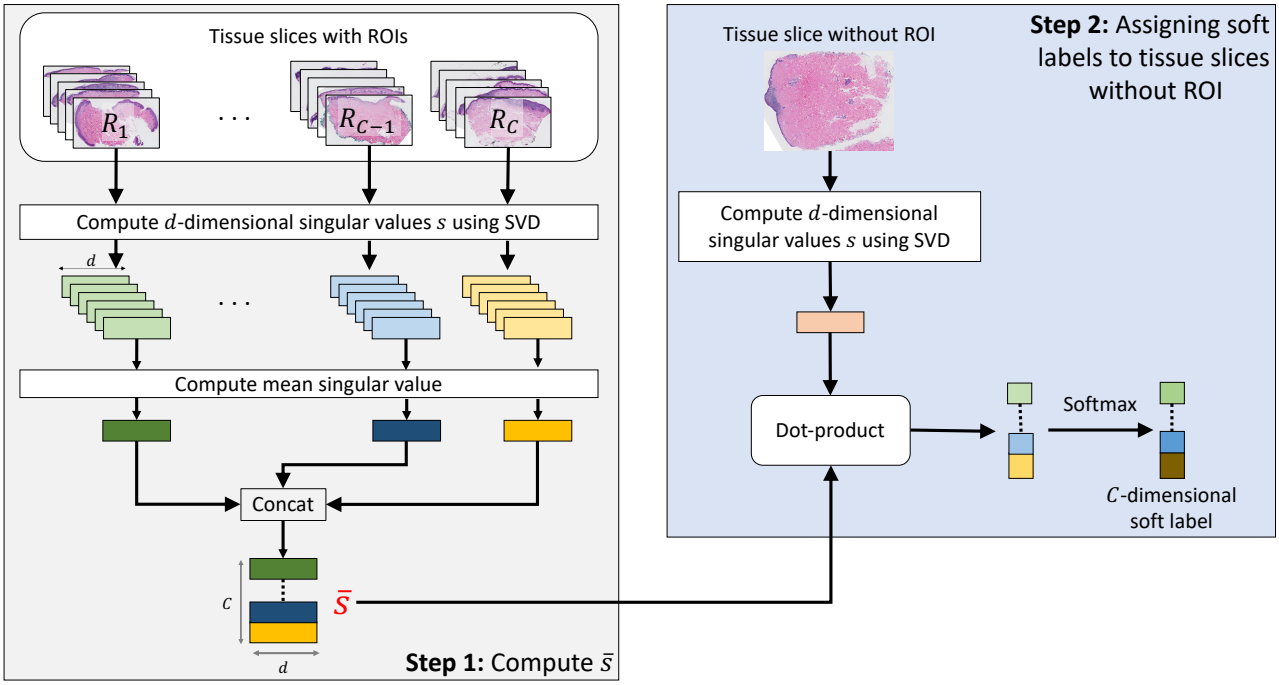


FIGURE 3: Overview of Soft labels calculation . Diagnostically constrained soft labels are calculated for tissue slices without an ROI using singular value decomposition (see Section III-C).

used singular values because of their uniqueness and robustness properties [42]–[45]. 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:

$$\hat{y}^j = \text{softmax}\left(\bar{\mathbf{s}} \cdot \hat{\mathbf{s}}^j\right) \tag{6}$$

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 4a. 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 $\tilde{y}^j = \{\tilde{y}^j_1, ..., \tilde{y}^j_C\}$ is computed as:

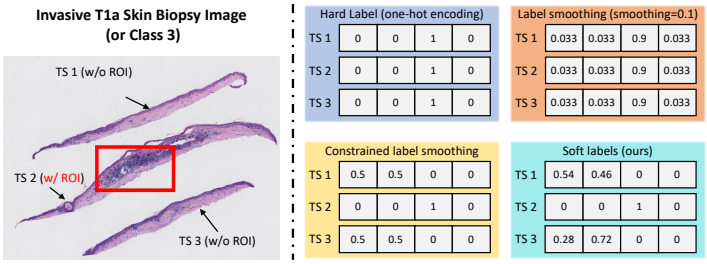
$$\begin{aligned} \tilde{y}^j_c &= \frac{\hat{y}^j_c}{\sum_{c=0}^{k} \hat{y}^j_c}, \quad if\ c < k \\ \tilde{y}^j_c &= 0 \quad if\ c \geq k \end{aligned} \tag{7}$$

Figure 4a 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.

IV. EXPERIMENTAL RESULTS

A. DATASET AND EVALUATION

Skin biopsy dataset and ground truth consensus. The data used for this study was acquired as a part of the MPATH study (R01CA151306) and consists of 240 skin biopsy images with hematoxylin and eosin (H&E) staining [6]. The 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 [46]. The consensus panel assessments were grouped into five different MPATH-Dx (Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis) [47] simplified categories based on perceived risk for progression. These five classes were regrouped to four diagnostic classes for the classification task in this paper 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*



(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 III-C)	**0.60**	**0.87**	**0.77**

(b)

FIGURE 4: (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.

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 1: Statistics of skin biopsy whole slide image (WSI) dataset. The average WSI size is computed at a magnification factor of $10\times$. 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**).

melanoma stage $\geq$ *pT1b* (**pT1b**) which is the greatest risk for regional and/or distant metastases. We randomly split 240 WSIs into 102 training, 23 validation and 115 test WSIs (see Table 1). Additionally, the consensus panel of three experienced dermatopathologists marked in total 240 regions of interest (ROIs) that best defined the diagnostic classification of each case during the review process. Information about these ROIs was used to produce soft labels for the training set (Section III-C).

Outcome metrics. The performance of `ScAtNet` is evaluated in terms of the following standard quantitative metrics: (1) classification (or Top-1) accuracy, (2) F1 score, (3) sensitivity, (4) specificity, and (5) area under receiver operating characteristic curves (ROC-AUC). The values of these metrics range between zero and one, and higher values of these metrics mean better performance. Multi-class F1 and specificity have the same value as accuracy.

Accuracy data from U.S. pathologists. 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 [6]. 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.

B. IMPLEMENTATION DETAILS

Extracting tissue slices from WSIs. The original WSIs were collected at a zoom level of $40\times$. Because WSIs at $40\times$ require extensive computational resources, we extracted WSIs at lower zoom levels of $7.5\times$ (average size 8348 x 7202), $10\times$ (average size 11130 x 9603), and $12.5\times$ (average size 13913 x12003). These zoom levels were selected based on previous work on histopathological image classification for different tissues [11], [16], [40], 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 4a). Therefore, individual tissue slices were extracted using a histogram-based segmentation method of Otsu [48] 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 [49] pretrained on the ImageNet dataset [41] 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 [50] and ResNet [10] may also be suitable for extracting patch-wise embeddings.

MobileNetv2 outputs 1280-dimensional patch-wise em-

beddings 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.

C. 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 [51] 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*.

D. 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. [11] and that of E. Mercan et al. [39], (R1 and R2 in Table 2). 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. [15] that builds slide-level feature representations via weighted aggregation of the patch representations (R3 and R4 in Table 2). 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 2) [7] 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. [8] (R7 and R8 in Table 2) 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. [22] (R4 in Table 2). 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.

E. RESULTS

Hard vs. soft labels. The performance of our soft labeling method (Section III-C) 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 4a) 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.* [52] 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 [53]. For the 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 4a contrasts our soft labeling method with these methods while quantitative comparison between these methods is given in Figure 4b. 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 5 compares the performance of single scale `ScAtNet` with different numbers of crops m at three different input resolutions ($7.5\times$, $10\times$, and $12.5\times$). Using fewer crops at larger resolution (e.g., 25 crops at a resolution of $12.5\times$) and more crops at smaller resolutions (e.g., 81 crops at a resolution of $7.5\times$) 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 [54], may help improve the performance.

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

Single vs. multiple input scales. Figure 6a compares the overall performance of `ScAtNet` across different metrics on single- and multi-scale inputs, while class-wise accuracy is given in Figure 6b. 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

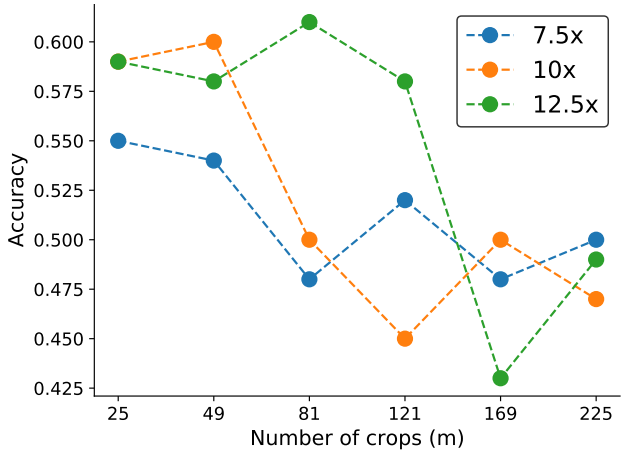
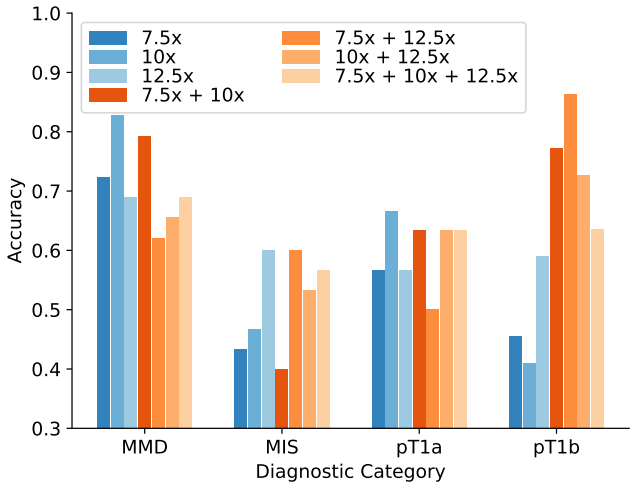


FIGURE 5: Effect of number of crops (m) on the performance of `ScAtNet` (single scale) for inputs at three different scale levels ($7.5\times$, $10\times$, and $12.5\times$).

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 6: 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*).

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 6b) is much more evenly distributed, which is not seen in all other combinations.

Comparison with baseline methods. Figure 2 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 [7] and the MS-DA-MIL method [8] 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 7), except the pT1b category. This is likely because the ChikonMIL method samples more relevant patches

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 2: 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 ($10\times$). MSC denotes multiple input scales ($7.5\times$, $10\times$, $12.5\times$). MSC* denotes multiple input scales ($10\times$, $20\times$)

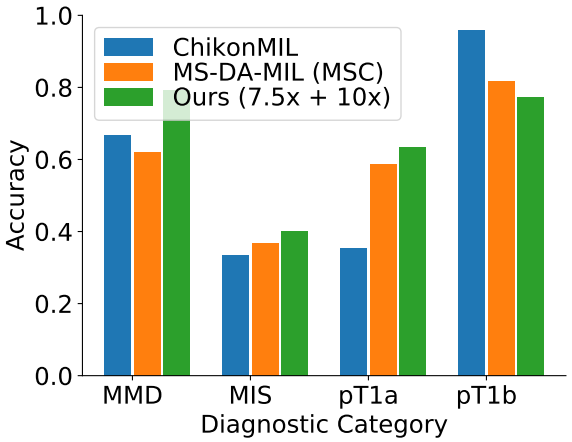


FIGURE 7: Comparison of class-wise accuracy with state-of-the-art WSI classification methods on the test set. Diagnostic terms are defined as the following: *mild and moderate dysplastic nevi* (*MMD*), *melanoma in situ* (*MIS*), *invasive melanoma stage pT1a* (*pT1a*), *invasive melanoma stage* $\geq$ *pT1b* (*pT1b*). Overall, `ScAtNet` delivered better performance across all diagnostic categories except the pT1b category.

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 3 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.

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 3: Comparison of `ScAtNet` with pathologists' (PG) performance. Pathologists' performance data is from a prior *independent* clinical study of 187 pathologists [6] who interpreted these same 115 cases in our test set (Table 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*).

V. DISCUSSION

Previous studies on computer-aided skin lesion analysis have been mainly focused on using dermoscopic images due to its inexpensiveness and availability [55]–[57]. Although dermoscopic images showed improvement for diagnosis of skin cancer compared to bare visual inspection, the gold standard for the diagnosis of melanocytic lesions is the interpretation of histopathology slides. There has been limited application of deep learning techniques in whole slide skin biopsy images due to their gigapixel size and the lack of large public datasets. Earlier studies analyzing whole slide skin biopsy images using deep learning have focused on dermis and epidermis segmentation, as well as two- or three-class classification problems. For example, Phillips *et al.* [58] explored segmentation of dermis and epidermis as well as tumor segmentation using convolutional neural network with a dataset of 50 WSIs (Training/validation/test: 36/7/7). Hekler *et al.* [59], [60] studied the binary classification of *nevi* vs. *melanoma* with a dataset of 695 WSIs (Training/Test: 595/100). Similarly, Lu and Mandal [61] and Xu *et al.* [17] performed a three-way classification task (17 normal skin, 17 melanocytic nevi, and 32 superficial spreading melanoma) using 66 WSIs. Note that the dataset used by Lu and Mandal et al. [61] and Xu et al. [17] is much smaller than ours and limited to only two of our classes, making direct comparison

impossible.

Unlike these studies, this work classifies the full spectrum of melanocytic skin biopsy lesions ranging from mildly atypical nevi and more advanced atypical pre-cursor lesions, to melanoma in situ to invasive melanoma. Our dataset consists of 240 WSIs, including 115 WSIs in an independent test set (Table 1). An independent test set allows us to demonstrate the generalization ability of `ScAtNet`. A key strength of our work is that we were able compare the diagnostic classification of `ScAtNet` with the performance of actively practicing U.S. pathologists who interpreted the same cases (test set) in an independent study.

Although the proposed method has shown great potential for automated melanocytic lesion classification, limitations are recognized. Our study is only relevant to melanocytic lesions, while only about one in four skin biopsies have melanocytic cells [62]. Moreover, despite having an independent test set, `ScAtNet` was evaluated on only 115 WSIs. In order to demonstrate its application in clinical settings, `ScAtNet` should be tested on a larger test set. Also, in this paper, we only studied skin biopsies. However, we believe that `ScAtNet` is generic and can be extended to other types of biopsy images, such as breast and lung.

VI. CONCLUSION

Diagnosis of melanocytic lesions is among the most challenging areas of pathology. Previous studies indicate that diagnostic errors occur frequently [3]–[5]. False positive readings for suspected melanoma range from 6% to 17% [63], [64]. 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 this paper, 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. The implementations of the models we use and algorithms we introduce are available at https://github.com/meredith-wenjunwu/ScAtNet.

VII. APPENDIX

a: Outcome metrics

The following metrics were used to evaluate the performance of `ScAtNet` [65]:

- 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:

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

  where TP, FP, TN, and FN denotes the true positive, false positive, true negative, 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 measures 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 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 [66].

The values of these metrics range between zero and one, and higher values of these metrics mean better performance.

A. SALIENCY ANALYSIS

Saliency analysis using gradients helps identify relevant areas in an input image that contributed to the prediction [67]. Figure 8 shows that both $7.5\times$ and $10\times$ contributed to the decision in the cases of *MMD* and *pT1a*, while $12.5\times$ contributes more in the cases of *MIS* and *pT1b*. This pattern illustrates that depending on the input whole slide image, diagnosis-specific features exist at different input scales and `ScAtNet` learns to weigh these features automatically.

B. ROC CURVES

In Figure 9, we compared the Receiver Operating Characteristic (ROC) curves of the proposed method with different numbers of input scales. With a single scale, the overall area under the curve (AUC) score as well as the class-wise AUC score of invasive cancer categories (*pT1a* and *pT1b*) improve with larger input scale. With two scales, we observed the best performance in the combination of the smallest and the largest scale ($7.5\times$ and $12.5\times$).

a: Comparison of baseline methods

In Figure 10, we compared ROC curves of the baseline methods. The MS-DA-MIL method of Hashimoto et al. [8] delivered the best AUC score, compared to the weighted feature aggregation method by C. Mercan et al. [15], ChikonMIL method by Chikontwe et al. [7], the patch-based classification method [11], [39] and the Streaming CNN method [22]. With multiple input scales, the patch-based method did not show significant improvement in AUC score, but the performance across all classes is more evenly distributed.

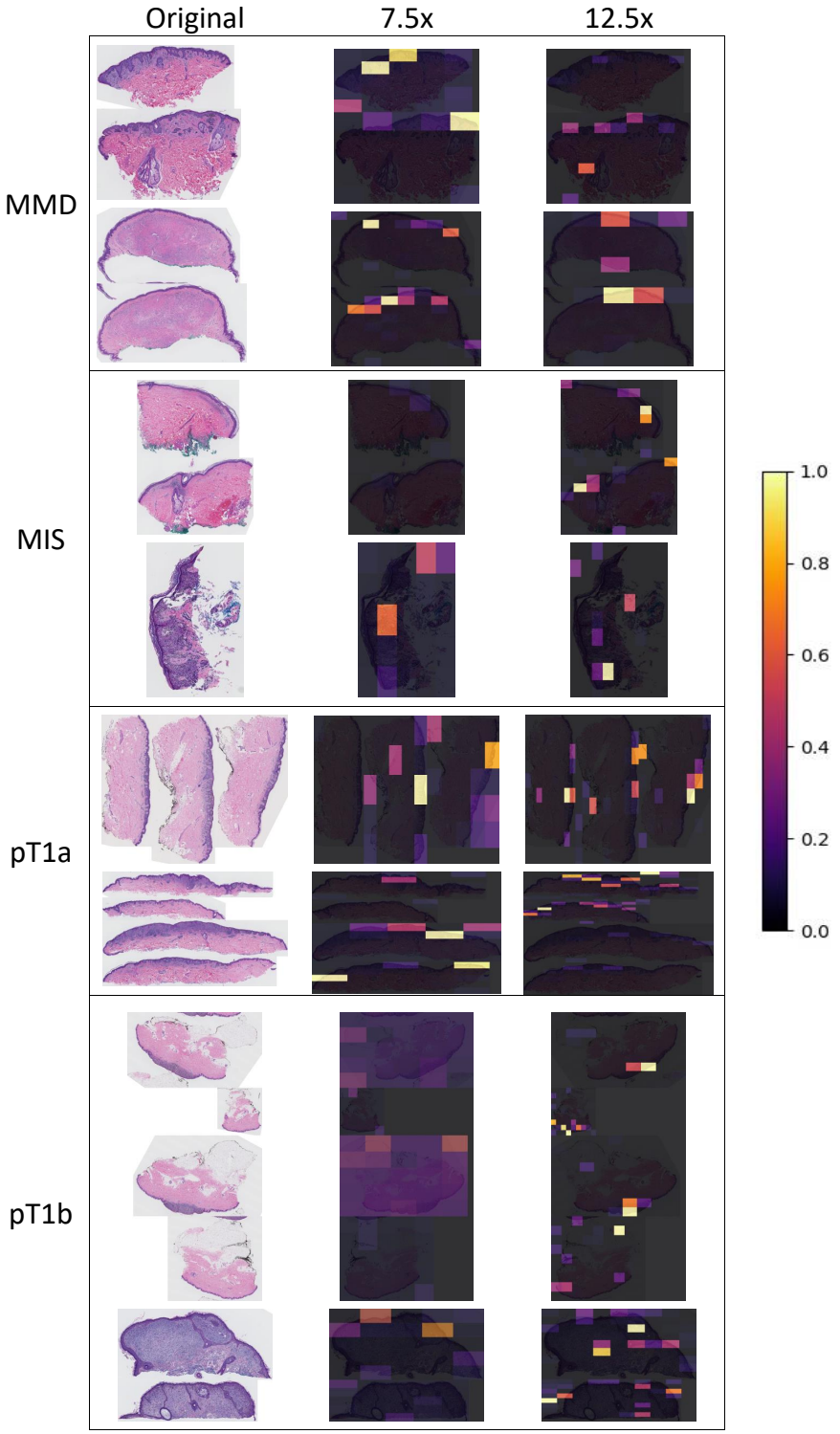
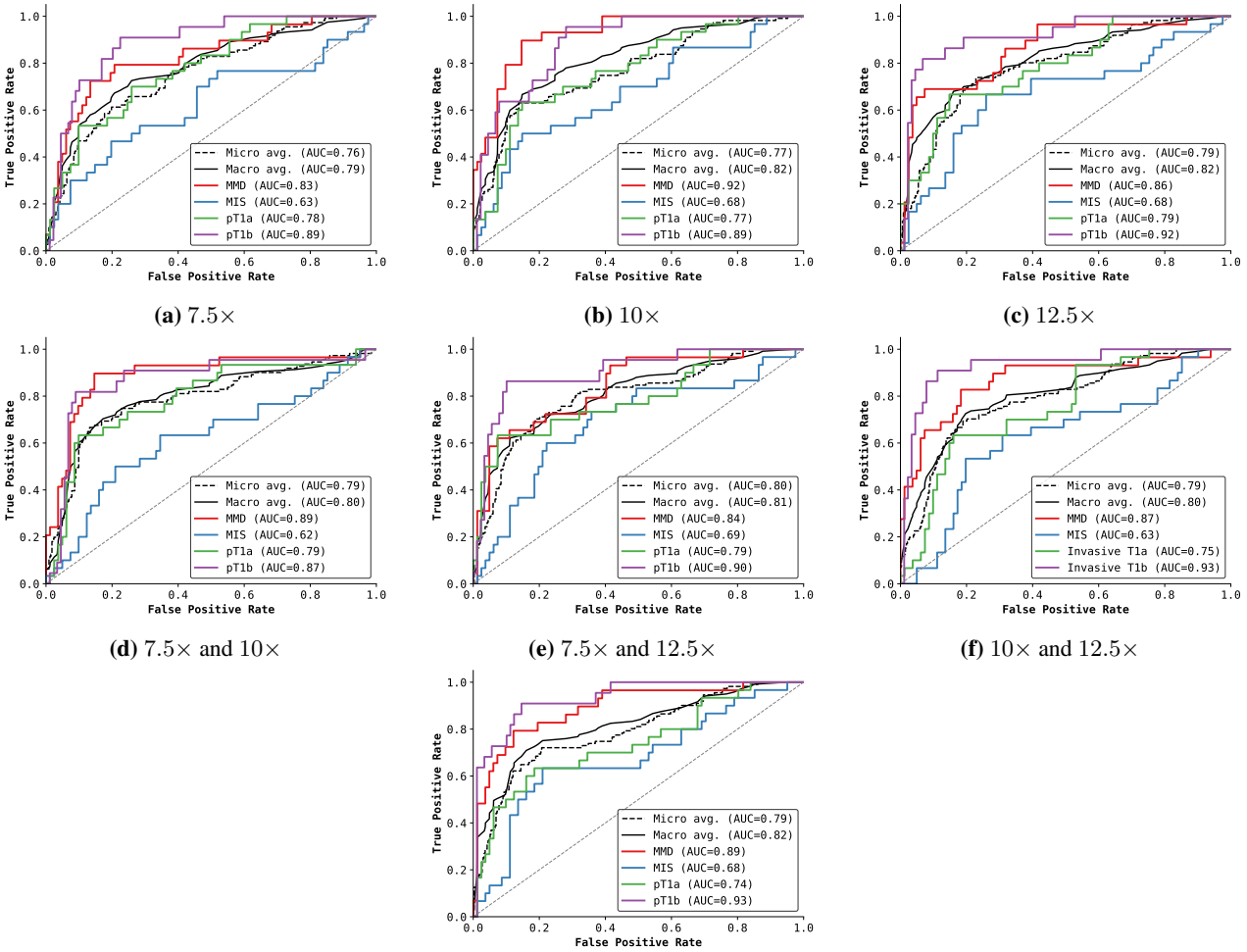


FIGURE 8: Visualization of gradient in `ScAtNet`. The left column shows original whole slide images in all diagnostic categories: *mild and moderate dysplastic nevi* (*MMD*), *melanoma in situ* (*MIS*), *invasive melanoma stage pT1a* (*pT1a*), *invasive melanoma stage* $\geq$ *pT1b* (*pT1b*). The right two columns are the corresponding gradient maps calculated from $7.5\times$ and $12.5\times$ input scales. All examples shown were correctly classified into their diagnostic categories. Colors from purple to yellow are assigned to values between 0 and 1.

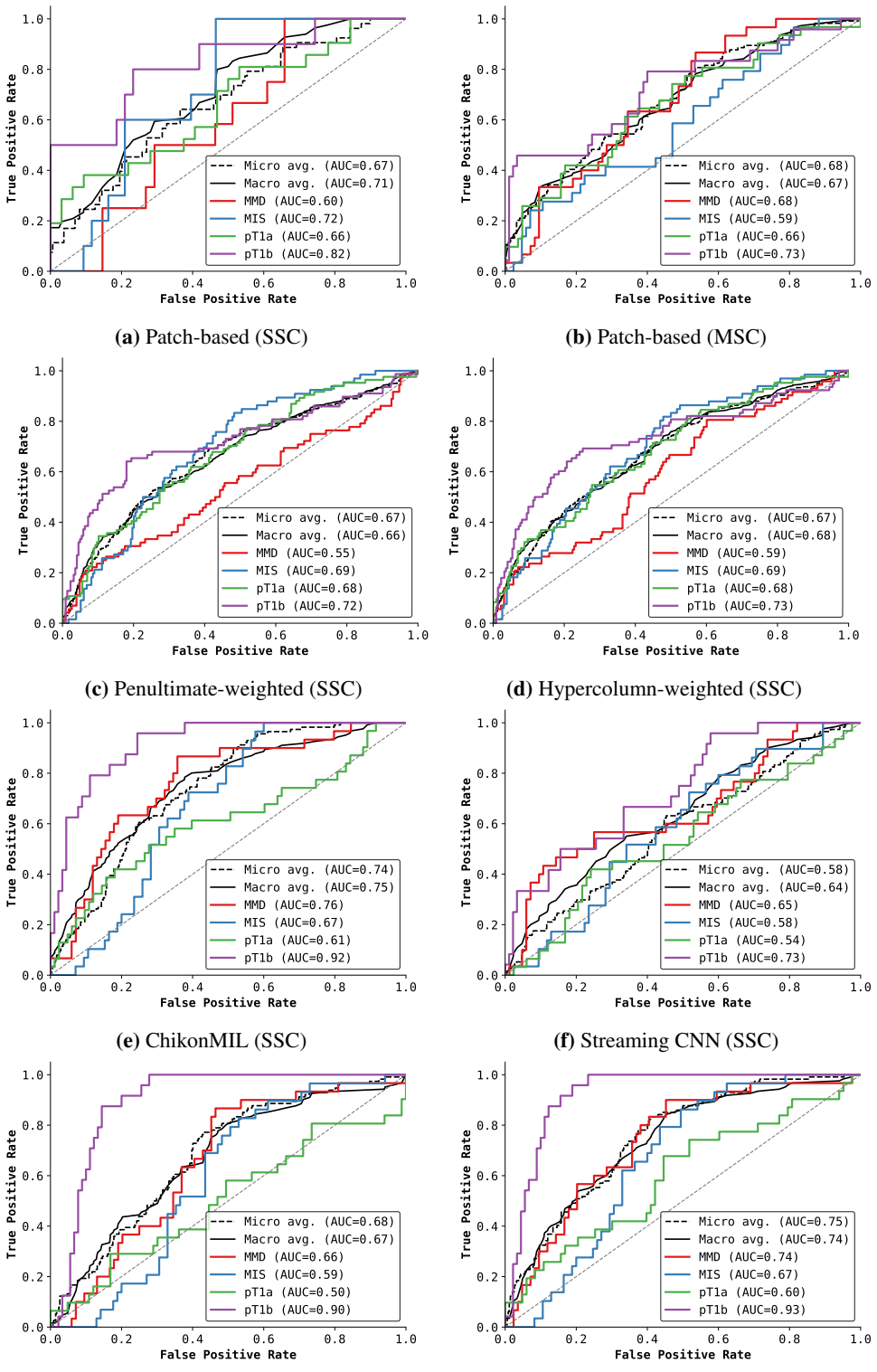


(a) 7.5× **(b)** 10× **(c)** 12.5×

(d) 7.5× and 10× **(e)** 7.5× and 12.5× **(f)** 10× and 12.5×

(g) 7.5× and 10× and 12.5×

FIGURE 9: Receiver operating characteristic (ROC) curves of `ScAtNet` with different numbers of input scales. For a single scale (a-c), the performance improves with the input scale, especially for invasive cancers. For two scale combinations (d-f), we do not observe significant gains. However, a combination of smaller and larger input scales (7.5× and 12.5×) delivered good performance across all diagnostic classes. 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*).



(a) Patch-based (SSC) **(b)** Patch-based (MSC)

(c) Penultimate-weighted (SSC) **(d)** Hypercolumn-weighted (SSC)

(e) ChikonMIL (SSC) **(f)** Streaming CNN (SSC)

(g) MS-DA-MIL (SSC) **(h)** MS-DA-MIL (MSC*)

FIGURE 10: Comparison of ROC curves with state-of-the-art WSI classification methods on the test set. Here, SSC denotes single input scale ($10\times$). MSC denotes multiple input scales ($7.5\times$, $10\times$, $12.5\times$), while MSC* denotes $10\times$, $20\times$. Overall, the MS-DA-MIL method of Hashimoto et al. [8] delivers the best performance of all other existing methods. 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*).

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021." *CA: a Cancer Journal for Clinicians*, vol. 71, no. 1, pp. 7–33, 2021.

[2] W. A. Wells, P. A. Carney, M. S. Eliassen, A. N. Tosteson, and E. R. Greenberg, "Statewide study of diagnostic agreement in breast pathology," *JNCI: Journal of the National Cancer Institute*, vol. 90, no. 2, pp. 142–145, 1998.

[3] V. Della Mea, F. Puglisi, M. Bonzanini, S. Forti, V. Amoroso, R. Visentin, P. Dalla Palma, and C. 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*, vol. 10, no. 6, pp. 636–641, 1997.

[4] K. H. Allison, L. M. Reisch, P. A. Carney, D. L. Weaver, S. J. Schnitt, F. P. O'Malley, B. M. Geller, and J. G. Elmore, "Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel," *Histopathology*, vol. 65, no. 2, pp. 240–251, 2014.

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

[6] J. G. Elmore, R. L. Barnhill, D. E. Elder, G. M. Longton, M. S. Pepe, L. M. Reisch, P. A. Carney, L. J. Titus, H. D. Nelson, T. Onega *et al.*, "Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study," *Bmj*, vol. 357, 2017.

[7] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 519–528.

[8] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, and I. 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*, 2020, pp. 3852–3861.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. 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*, 2016, pp. 2424–2433.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[13] J. N. Marsh, T.-C. Liu, P. C. Wilson, S. J. Swamidass, and J. P. Gaut, "Development and validation of a deep learning model to quantify glomerulosclerosis in kidney biopsy specimens," *JAMA network open*, vol. 4, no. 1, pp. e2 030 939–e2 030 939, 2021.

[14] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images," *IEEE transactions on medical imaging*, vol. 37, no. 1, pp. 316–325, 2017.

[15] C. Mercan, B. Aygunes, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Deep feature representations for variable-sized regions of interest in breast histopathology," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[16] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, "Y-net: joint segmentation and classification for diagnosis of breast biopsy images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 893–901.

[17] H. Xu, C. Lu, R. Berendt, N. Jha, and M. Mandal, "Automated analysis and classification of melanocytic tumor on skin whole slide images," *Computerized medical imaging and graphics*, vol. 66, pp. 124–134, 2018.

[18] E. Mercan, S. Mehta, J. Bartlett, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions," *JAMA network open*, vol. 2, no. 8, pp. e198 777–e198 777, 2019.

[19] H. Ni, H. Liu, K. Wang, X. Wang, X. Zhou, and Y. 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*. Springer, 2019, pp. 36–44.

[20] M. Van Zon, N. Stathonikos, W. A. Blokx, S. Komina, S. L. Maas, J. P. Pluim, P. J. Van Diest, and M. Veta, "Segmentation and classification of melanoma and nevus in whole slide images," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 263–266.

[21] S. Mehta, X. Lu, D. Weaver, J. G. Elmore, H. Hajishirzi, and L. Shapiro, "Hatnet: An end-to-end holistic attention network for diagnosis of breast biopsy images," *arXiv preprint arXiv:2007.13007*, 2020.

[22] H. Pinckaers, W. Bulten, J. Van der Laak, and G. Litjens, "Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels," *IEEE transactions on medical imaging*, vol. PP, March 2021.

[23] T. Chen, B. Xu, C. Zhang, and C. Guestrin, "Training deep nets with sublinear memory cost," *arXiv preprint arXiv:1604.06174*, 2016.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[26] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[27] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[28] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv preprint arXiv:2012.12877*, 2020.

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[30] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," *arXiv preprint arXiv:2102.10662*, 2021.

[31] Z. Zhang, B. Sun, and W. Zhang, "Pyramid medical transformer for medical image segmentation," *arXiv preprint arXiv:2104.14702*, 2021.

[32] Y. Zhang, R. Higashita, H. Fu, Y. Xu, Y. Zhang, H. Liu, J. Zhang, and J. Liu, "A multi-branch hybrid transformer networkfor corneal endothelial cell segmentation," *arXiv preprint arXiv:2106.07557*, 2021.

[33] O. Petit, N. Thome, C. Rambour, and L. Soler, "U-net transformer: Self and cross attention for medical image segmentation," *arXiv preprint arXiv:2103.06104*, 2021.

[34] C. Sitaula and M. B. Hossain, "Attention-based vgg-16 model for covid-19 chest x-ray image classification," *Applied Intelligence*, vol. 51, no. 5, pp. 2850–2863, 2021.

[35] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[36] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[37] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. Elmore, and L. Shapiro, "Learning to segment breast biopsy whole slide images," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 663–672.

[38] T. T. Brunyé, E. Mercan, D. L. Weaver, and J. G. Elmore, "Accuracy is in the eyes of the pathologist: the visual interpretive process and diagnostic accuracy with digital whole slide images," *Journal of biomedical informatics*, vol. 66, pp. 171–179, 2017.

[39] E. Mercan, L. G. Shapiro, T. T. Brunyé, D. L. Weaver, and J. G. Elmore, "Characterizing diagnostic search patterns in digital breast pathology: scanners and drillers," *Journal of digital imaging*, vol. 31, no. 1, pp. 32–41, 2018.

[40] B. Li, Y. Li, and K. 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.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[42] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of dimensionality reduction in recommender system-a case study," Minnesota Univ Minneapolis Dept of Computer Science, Tech. Rep., 2000.

[43] R. Liu and T. Tan, "An svd-based watermarking scheme for protecting rightful ownership," *IEEE transactions on multimedia*, vol. 4, no. 1, pp. 121–128, 2002.

[44] C.-C. Chang, P. Tsai, and C.-C. Lin, "Svd-based digital image watermarking scheme," *Pattern Recognition Letters*, vol. 26, no. 10, pp. 1577–1586, 2005.

[45] M. Nilashi, O. Ibrahim, and K. Bagherifard, "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques," *Expert Systems with Applications*, vol. 92, pp. 507–520, 2018.

[46] P. A. Carney, L. M. Reisch, M. W. Piepkorn, R. L. Barnhill, D. E. Elder, S. Knezevich, B. M. Geller, G. Longton, and J. G. Elmore, "Achieving consensus for the histopathologic diagnosis of melanocytic lesions: use of the modified delphi method," *Journal of cutaneous pathology*, vol. 43, no. 10, pp. 830–837, 2016.

[47] M. W. Piepkorn, R. L. Barnhill, D. E. Elder, S. R. Knezevich, P. A. Carney, L. M. Reisch, and J. G. Elmore, "The mpath-dx reporting schema for melanocytic proliferations and melanoma," *Journal of the American Academy of Dermatology*, vol. 70, no. 1, pp. 131–141, 2014.

[48] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[49] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[51] H. Chen, S. Lundberg, and S.-I. Lee, "Checkpoint ensembles: Ensemble methods from a single training process," *arXiv preprint arXiv:1710.03282*, 2017.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[53] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" *arXiv preprint arXiv:1906.02629*, 2019.

[54] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[55] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2092–2103, 2019.

[56] N. Nida, A. Irtaza, A. Javed, M. H. Yousaf, and M. T. Mahmood, "Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy c-means clustering," *International journal of medical informatics*, vol. 124, pp. 37–48, 2019.

[57] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific data*, vol. 8, no. 1, pp. 1–8, 2021.

[58] A. Phillips, I. Teo, and J. Lang, "Segmentation of prognostic tissue structures in cutaneous melanoma using whole slide images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[59] A. Hekler, J. S. Utikal, A. H. Enk, W. Solass, M. Schmitt, J. Klode, D. Schadendorf, W. Sondermann, C. Franklin, F. Bestvater *et al.*, "Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images," *European Journal of Cancer*, vol. 118, pp. 91–96, 2019.

[60] A. Hekler, J. S. Utikal, A. H. Enk, C. Berking, J. Klode, D. Schadendorf, P. Jansen, C. Franklin, T. Holland-Letz, D. Krahl *et al.*, "Pathologist-level classification of histopathological melanoma images with deep neural networks," *European Journal of Cancer*, vol. 115, pp. 79–83, 2019.

[61] C. Lu and M. Mandal, "Automated analysis and diagnosis of skin melanoma on whole slide histopathological images," *Pattern Recognition*, vol. 48, no. 8, pp. 2738–2750, 2015.

[62] J. P. Lott, D. M. Boudreau, R. L. Barnhill, M. A. Weinstock, E. Knopp, M. W. Piepkorn, D. E. Elder, S. R. Knezevich, A. Baer, A. N. Tosteson *et al.*, "Population-based analysis of histologically confirmed melanocytic proliferations using natural language processing," *JAMA dermatology*, vol. 154, no. 1, pp. 24–29, 2018.

[63] L. Brochez, E. Verhaeghe, E. Grosshans, E. Haneke, G. Piérard, D. Ruiter, and J.-M. 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*, vol. 196, no. 4, pp. 459–466, 2002.

[64] M. Cook, T. Clarke, S. Humphreys, A. Fletcher, K. McLaren, N. Smith, A. Stevens, J. 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*, vol. 28, no. 6, pp. 497–512, 1996.

[65] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, 2020.

[66] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[67] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

WENJUN WU received her B.S. in Biomedical Engineering from Georgia Institute of Technology, Atlanta, USA in 2017. She is currently pursuing her Ph.D. in Biomedical Informatics at University of Washington, Seattle, Washington, USA.

From 2018 till now, she is a research assistant, advised by Linda Shapiro at University of Washington. Her research interests lie at the intersection of biomedical image analysis, machine learning, and computer vision.

SACHIN MEHTA is an AI/ML Research Scientist at Apple, Inc. Prior to joining Apple, he received his Ph.D. from the University of Washington, Seattle, Washington, USA.

His research interests lies in the intersection of computer vision, NLP, and machine learning, especially in designing fast, light-weight, power efficient, and memory efficient neural architectures that can be used for modeling visual and textual data on resource-constrained devices across different domains, including computer vision for accessible technologies and health care.

SHIMA NOFALLAH received her B.Sc and M.Sc in Biomedical Engineering from Amirkabir University of Technology, Tehran, Iran. She is currently pursuing her Ph.D. in Electrical and Computer Engineering at University of Washington, Seattle, WA, USA.

Her research interests include Computer Vision, Machine Learning, and Medical Image processing.

STEVAN KNEZEVICH is a 2004 graduate of the University of Toronto Medical School, Ontario, Canada. Prior to this, he received his PhD in Pathology from the University of British Columbia in 1999 and then spent an additional year as a post-doctoral fellow in Lymphoma research. His Residency and Surgical Pathology Fellowship were completed at Washington University in St. Louis, MO and his Dermatopathology Fellowship at Stanford University.

Dr. Knezevich worked at the VA Medical Center and served as Assistant Professor at the University of Washington in Seattle prior to joining Pathology Associates in July 2014. He is Board Certified in Anatomic Pathology, Clinical Pathology, and Dermatopathology.

CAITLIN J. MAY received an MD degree from the University of Washington School of Medicine in 2013 and subsequently completed a Dermatology residency (2017) and Dermatopathology fellowship (2018) at the same institution. She currently works as a dermatopathologist at Dermatopathology Northwest (Bellevue, WA) and as a teledermatologist for the VA Seattle Medical Center.

She is a collaborator on Dr. Elmore's NIH funded grant, IMPACT (R01CA200690), which utilizes novel computational methods to analyze whole slide digital images to improve the diagnosis of melanoma and related skin lesions. She also has ongoing projects with her research team that involve evaluating trends in immunohistochemical and molecular testing among U.S. pathologists in their diagnoses of melanocytic lesions. Her main research interests include the diagnostic challenges associated with the histopathologic diagnoses of melanocytic lesions.

OLIVER H. CHANG received his B.A. at the University of Illinois Urbana-Champaign (2005) and his M.D. at the University of Illinois Chicago College of Medicine (2010). He completed a residency at University of Washington Medical Center (2015) and is board certified in Anatomic Pathology, Clinical Pathology, and Dermatopathology.

He is an assistant professor at the University of Washington in the Department of Laboratory Medicine where he serves as the Director of Medical Student Clerkships and Post-sophomore Fellowship. His clinical practice is at the VA Puget Sound Hospital in Seattle, WA. His research interests include medical education in pathology, melanocytic lesions, and AI/Machine learning.

LINDA G. SHAPIRO (M'74-SM'81-F'96) received the B.S. degree in mathematics from the University of Illinois, Urbana, in 1970, and the M.S. and Ph.D. degrees in computer science from the University of Iowa, Iowa City, in 1972 and 1974, respectively.

She was an Assistant Professor of Computer Science at Kansas State University, Manhattan, from 1974 to 1978 and was an Assistant Professor of Computer Science from 1979 to 1981 and Associate Professor of Computer Science from 1981 to 1984 at Virginia Polytechnic Institute and State University, Blacksburg. She was Director of Intelligent Systems at Machine Vision International in Ann Arbor from 1984 to 1986. She is currently Professor of Computer Science and Engineering and of Electrical Engineering at the University of Washington. Her research interests include computer vision, image database systems, artificial intelligence, pattern recognition, and robotics. She has co-authored three textbooks, one on data structures and two on computer vision.

Dr. Shapiro is a Fellow of the IEEE and a Fellow of the IAPR. She is a past Chair of the IEEE Computer Society Technical Committee on Pattern Analysis and Machine Intelligence and is currently an editorial board member of Computer Vision and Image Understanding and of Pattern Recognition. She has served as Editor-in-Chief of CVGIP: Image Understanding, Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence, co-Program Chair of the IEEE Conference on Computer Vision and Pattern Recognition in 1994, General Chair of the IEEE Workshop on Directions in Automated CAD-Based Vision in 1991, and General Chair of the IEEE Conference on Computer Vision and Pattern Recognition in 1986. She was the Co-Chair of the Medical and Multimedia Applications Track of the International Conference on Pattern Recognition for 2002 and a Co-Chair of CVPR 2008. She has also served on the program committees of numerous vision and AI workshops and conferences.

JOANN G. ELMORE received her medical degree from the Stanford University School of Medicine, residency training in internal medicine at Yale- New Haven Hospital, with advanced epidemiology training from the Yale School of Epidemiology and Public Health and the RWJF Clinical Scholars Program. In addition, she was a RWJF generalist physician faculty scholar. Dr. Elmore is board certified in internal medicine and serves on many national and international committees. She is Editor in Chief for Primary Care at UpToDate and enjoys seeing patients as a primary care internist and teaching clinical medicine to students and residents.

Dr. Elmore is The Rosalinde and Arthur Gilbert Foundation Endowed Chair in Health Care Delivery, professor of medicine at the David Geffen School of Medicine at UCLA, and Director of the UCLA National Clinician Scholars Program.

She conducts scientific research on diagnostic accuracy of screening and medical tests as well as AI/machine learning to develop computer aid tools for the early detection of high-risk cancers. She previously held faculty and leadership positions at the University of Washington, Fred Hutchinson Cancer Research Center, Group Health Research Institute, Yale University and has been an Associate Director and member of the National Advisory Committee for of the Robert Wood Johnson Clinical Scholars program at Yale and the University of Washington.