ABSTRACT Diagnosis of 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 weights 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-
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

![Diagram of ScAtNet](image-url)

**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.
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 model that adapts to the information from different input scales using self-attention and predicts the classification label.

**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].

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 $I \in \mathbb{R}^{n \times e}$ to produce query ($Q$), key ($K$), and value ($V$) embeddings, where $n$ is the number of inputs and $e$ is the input dimensionality. A dot-product between query ($Q$) and key ($K$) 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 $V$.

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

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 × 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 $X^{sc} \in \mathbb{R}^{W \times H}$ at scale $sc$ with width $W$ and height $H$ is divided into $m$ non-overlapping patches $X^{sc} = (x^{sc}_1, \ldots , x^{sc}_m)$, where $x^{sc}_i$ 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 $PE^{sc}_i \in \mathbb{R}^{e}$ for the $i$-th patch $x^{sc}_i$ is thus:

$$PE^{sc}_i = CNN(x^{sc}_i) \quad (2)$$

**Contextualized patch embeddings.** The patch embeddings $PE^{sc} \in \mathbb{R}^{m \times e}$ are produced independently for each patch. In other words, these embeddings $PE^{sc}$ do not encode inter-patch relationships. These embeddings $PE^{sc}$ are fed to a transformer to learn inter-patch relationships. Similar to vision transformers [27], patch-wise sinusoidal positional embeddings $PPE^{sc} \in \mathbb{R}^{m \times e}$ are added to $PE^{sc}$ to encode the position of input patches. The resultant embeddings are then fed to a transformer to produce contextualized patch embeddings $CPE^{sc} \in \mathbb{R}^{m \times e}$.

$$CPE^{sc} = Transformer(PE^{sc} + PPE^{sc}) \quad (3)$$

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

**Contextualized scale embeddings.** The embedding $CPE^{c}$ encodes the information in an image $X^{sc}$ at scale $sc$. Let us assume that we have $\mathcal{S}$ scales. For each scale $sc \in \{0, \ldots , \mathcal{S}\}$, we produce embedding vector $CPE^{sc}$ and concatenate them to produce scale-level embeddings $SE = Concat(CPE^{1}, \ldots , CPE^{\mathcal{S}})$. These embeddings $SE \in \mathbb{R}^{\mathcal{S} \times e}$ 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 $PSE \in \mathbb{R}^{sc \times e}$ are added to $SE^{sc \times e}$. The resultant embeddings are then fed to another transformer to produce contextualized scale embeddings $CSE \in \mathbb{R}^{sc \times e}$.

$$CSE = Transformer(SE + PSE) \quad (4)$$

For predicting the diagnostic class, ScAtNet first flattens the scale-aware embeddings $CSE \in \mathbb{R}^{sc \times e}$ 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 $D$ 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 $R$ with an ROI and (2) tissue slices $N \bar{R}$ without an ROI. Since each slice in $R$ has a representative ROI, we further split $R$ into $C$ subsets, $\bar{R} = \{R_1, \ldots , R_C\}$, 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 $\bar{s}_i$ for each subset $R_i$ as:

$$\bar{s}_i = \frac{1}{m} \sum_{j=1}^{m} s_{ij} \quad (5)$$

where $s_{ij}$ 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 $\mathcal{S}$ 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%.
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 \( \mathcal{N}R \), the \( C \)-dimensional soft label vector \( \hat{y}^j \) is computed as:

\[
\hat{y}^j = \text{softmax} \left( \bar{s} \cdot \hat{s}^j \right)
\]  

(6)

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

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 \( j \)-th 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 \( \tilde{y}^j = \{\tilde{y}_1^j, \ldots, \tilde{y}_C^j\} \) is computed as:

\[
\tilde{y}_c^j = \begin{cases} 
\frac{\hat{y}_c^j}{\sum_{c=0}^{k} \hat{y}_c^j}, & \text{if } c < k \\
0, & \text{if } c \geq k
\end{cases}
\]  

(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 (RO1CA151306) 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
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.

**TABLE 1:** Statistics of 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 ≥ pT1b (pT1b).

<table>
<thead>
<tr>
<th>Diagnostic Category</th>
<th>Number of WSIs</th>
<th>Average WSI size (in pixels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMD</td>
<td>26</td>
<td>6</td>
</tr>
<tr>
<td>MIS</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>pT1a</td>
<td>33</td>
<td>6</td>
</tr>
<tr>
<td>pT1b</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>102</td>
<td>23</td>
</tr>
</tbody>
</table>

**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 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×. 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 x 12003). 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 ($\varepsilon = 128$) 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 $5 \times 10^{-4}$ in 500 steps. In the next 50 epochs, the model is trained with a learning rate of $5 \times 10^{-4}$. 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 Szegedy 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 ($7.5 \times 7.5$, $10 \times 10$, and $12.5 \times 12.5$). Using fewer crops at larger resolution (e.g., 25 crops at a resolution of $12.5 \times 12.5$) and more crops at smaller resolutions (e.g., 81 crops at a resolution of $7.5 \times 7.5$) 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 $224 \times 224$. 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 $7.5 \times 7.5$ input resolution, $m = 49$ for $10 \times 10$ input resolution, and $m = 81$ for $12.5 \times 12.5$ 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 cancer category. Compared to two scales, the overall performance with three scales remains the same. However, with three scales, the performance across all diagnostic categories (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...
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.

**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 performed a three-way classification task (17 normal skin, 595 melanocytic nevi, and 32 superficial spreading melanoma) [64] using 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 ≥ pT1b (pT1b). Overall, ScAtNet delivered better performance across all diagnostic categories except the pT1b category.

**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×). MSC denotes multiple input scales (7.5×, 10×, 12.5×). MSC* denotes multiple input scales (10×, 20×).

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

**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 ≥ pT1b (pT1b).

<table>
<thead>
<tr>
<th>Diagnostic Category</th>
<th>Accuracy</th>
<th>F1</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIS</td>
<td>0.46</td>
<td>0.40</td>
<td>0.49</td>
<td>0.40</td>
</tr>
<tr>
<td>pT1a</td>
<td>0.51</td>
<td>0.65</td>
<td>0.62</td>
<td>0.63</td>
</tr>
<tr>
<td>pT1b</td>
<td>0.72</td>
<td>0.77</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td>Overall</td>
<td>0.65</td>
<td>0.64</td>
<td>0.65</td>
<td>0.64</td>
</tr>
</tbody>
</table>

**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 ≥ pT1b (pT1b). Overall, ScAtNet achieved 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.
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]:

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

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

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

- Sensitivity measures proportion of the positive cases that are correctly classified and is defined as:

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

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

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

- 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× and 10× contributed to the decision in the cases of MMD and pT1a, while 12.5× 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× and 12.5×).

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 ≥ pT1b (pT1b). The right two columns are the corresponding gradient maps calculated from 7.5× and 12.5× 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.
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 ≥ pT1b (pT1b).
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×). MSC denotes multiple input scales (7.5×, 10×, 12.5×), while MSC* denotes 10×, 20×. 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 ≥ pT1b (pT1b).
REFERENCES


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.