

Differentiable Zooming for Multiple Instance Learning on Whole-Slide Images

Kevin Thandiackal^{1,2*}, Boqi Chen^{1,2*}, Pushpak Pati¹, Guillaume Jaume^{3†}, Drew F. K. Williamson³, Maria Gabrani¹, and Orcun Goksel^{2,4}

¹ IBM Research Europe, Zurich, Switzerland
{kth,che,pus,mga}@zurich.ibm.com

² ETH Zurich, Zurich, Switzerland

³ Brigham and Women’s Hospital, Harvard Medical School, Boston, USA
{gjaume,dwilliamson}@bwh.harvard.edu

⁴ Uppsala University, Uppsala, Sweden
orcun.goksel@it.uu.se

Abstract. Multiple Instance Learning (MIL) methods have become increasingly popular for classifying gigapixel-sized Whole-Slide Images (WSIs) in digital pathology. Most MIL methods operate at a *single* WSI magnification, by processing *all* the tissue patches. Such a formulation induces high computational requirements and constrains the contextualization of the WSI-level representation to a single scale. Certain MIL methods extend to multiple scales, but they are computationally more demanding. In this paper, inspired by the pathological diagnostic process, we propose ZOOMMIL, a method that *learns* to perform multi-level zooming in an end-to-end manner. ZOOMMIL builds WSI representations by aggregating tissue-context information from multiple magnifications. The proposed method outperforms the state-of-the-art MIL methods in WSI classification on two large datasets, while significantly reducing computational demands with regard to Floating-Point Operations (FLOPs) and processing time by 40–50×. Our code is available at: <https://github.com/histocartography/zoommil>.

Keywords: Whole-Slide Image Classification, Multiple Instance Learning, Multi-scale Zooming, Efficient Computational Pathology

1 Introduction

Histopathological diagnosis consists of examining tissue samples to characterize their phenotype, morphology, and the topological distribution of their constituents. With advancements in slide-scanning technologies, tissue specimens can now be digitized into Whole-Slide Images (WSIs) with high resolution, enabling the pathological assessment to be conducted on a computer rather than under a microscope. A WSI contains rich tissue information and can be up to

* Contributed equally.

† Work done while at IBM Research Europe

100 000 \times 100 000 pixels in size at 40 \times magnification (0.25 μ m/pixel). Due to the image size, complexity, and multi-scale nature of biological systems, a pathologist generally examines a WSI in a hierarchical manner, *i.e.*, detecting informative regions at a *low* magnification, and evaluating selected areas at a *high* magnification, as shown in Figure 1(a). However, such manual examination of a gigapixel-sized WSI can be cumbersome, time-consuming, and prone to inter- and intra-observer variability [16,14].

To alleviate the aforementioned challenges, Deep Learning (DL)-based diagnosis tools are being developed in digital pathology. However, these tools encounter additional challenges pertaining to the size of WSIs, and the difficulty of acquiring fine-grained annotations. To this end, DL methods have been proposed, in particular, using Multiple Instance Learning (MIL). Here, a WSI is decomposed into a bag of patches, which are individually encoded by a Convolutional Neural Network (CNN) backbone. A pooling operation then combines the patch embeddings into a slide-level representation that is finally mapped to the slide label. Although MIL methods have achieved remarkable performance on several pathology tasks, *e.g.*, tumor classification [10,32,42,39], tumor segmentation [22,30], and survival prediction [44], they pose the following drawbacks.

First, the performance of MIL methods relies on a carefully tuned context-resolution trade-off [6,41,36], *i.e.*, an optimal operating resolution that includes adequate context in a patch. As the dimensions of diagnostically relevant tissue vary significantly in histopathology, patches of different sizes across magnifications convey different context information about the tissue microenvironment. Thus, identifying an optimal resolution and patch size involves several tailored and tedious steps. Typical MIL methods use patches at a *single* magnification (Figure 1(b)) and disregard the spatial distribution of patches, causing the above problem. Although [33,40] address this via visual self-attention, they are constrained by expensive computations of attention scores on a large number of patches in a WSI. Differently, [26] addresses the issue via random patch sampling and sparse convolutions, consequently preventing deterministic inference. In the literature, other methods [41,17,19] are extracting concentric patches across multiple magnifications (Figure 1(c)) to acquire richer context per patch. However, they are computationally more expensive as they need to encode all patches at high magnification and the corresponding patches across lower magnifications.

Second, most MIL methods process *all* tissue patches at high magnification, thus processing a large number of uninformative patches, which increases computational cost, inference time, and memory requirements. For instance, inference on a WSI of 50 000 \times 50 000 pixels using CLAM [32], an MIL method, requires \approx 150 Tera Floating-Point Operations (FLOPs), which is 37 500 \times the processing of an ImageNet [12] sample by ResNet34 [18]. Further, the high memory footprint of MIL methods inhibits their scalability to large histopathology images, *e.g.*, prostatectomy slides which can be 300 000 \times 400 000 pixels at 40 \times magnification. Such computational requirements can in turn hinder the clinical deployment of these methods. Their adoption becomes even prohibitive when computational resources are scarce due to limited access to GPUs or cloud ser-

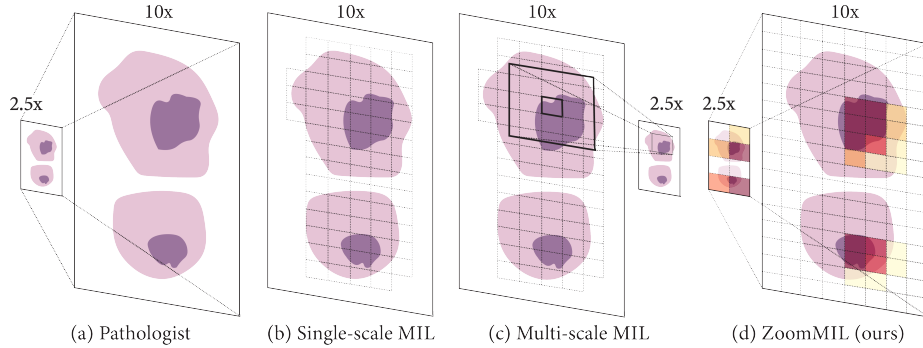


Fig. 1. Comparison of different methods for the diagnosis of WSIs

vices. In view of the above challenges, a multi-scale context-aware MIL method with high computational efficiency is desired.

In this paper, we propose ZOOMMIL, a novel method inspired by the hierarchical diagnostic process of pathologists. We first select Regions-of-Interest (ROIs) at low magnification, and zoom in on them at high magnification for finer analysis, as in Figure 1(d). The RoI selection is performed through a gated-attention and a differentiable top-K (DIFF-TOPK) module, which learns where to zoom, in an end-to-end manner, while moderating computational requirements at high magnifications. The process can be repeated across an arbitrary number of magnifications, *e.g.*, $5\times \rightarrow 10\times \rightarrow 20\times$, as per the task at hand. Finally, we aggregate the information acquired across multiple scales to obtain a context-aware WSI representation for downstream pathology tasks, as shown in Figure 2. In summary, our contributions are:

1. A novel multi-scale context-aware MIL method that learns to perform multi-level zooming in an end-to-end manner for WSI classification.
2. A computationally more efficient method compared to the state of the art in MIL, *e.g.*, $40\times$ faster inference on a WSI of size $26\,009 \times 18\,234$ pixels at $10\times$ magnification, while achieving better (2/3 datasets) or comparable (1/3 datasets) WSI classification performance.
3. Comprehensive benchmarking of the method with regard to WSI classification performance and computational requirements (on GPU and CPU) on multiple datasets across multiple organs and pathology tasks, *i.e.*, tumor subtyping, grading, and metastasis detection.

2 Related Work

2.1 Multiple Instance Learning in Histopathology

MIL in histopathology was introduced in [20] to classify breast and colon ROIs. The experiments established the superiority of *attention*-based pooling over *max*

and *mean* pooling. Concurrently, [10] scaled MIL to WSIs for grading prostate biopsies. They proposed Recurrent Neural Network (RNN)-based pooling for end-to-end training. Later, several works [44,32,31] consolidated attention-based MIL across several organs and pathology tasks. Recently, transformer-based MIL [33,40] has been proposed to consider inter-patch dependencies, with the downside of computing a quadratic number of interactions, which introduces memory constraints. Further, all the above MIL methods are limited to operate on *all* patches in a WSI at a single magnification. In view of the benefits of multi-scale information in histopathology [4,15,43,28,19,36], a few recent methods [17,27] have extended MIL to combine information across multiple magnifications. However, similar to single-scale methods, these multi-scale versions also require the processing of *all* patches in a WSI, which is computationally more expensive. In contrast, our proposed ZOOMMIL *learns* to identify informative regions at low magnification and subsequently zooms in on these regions at high magnification for efficient and comprehensive analysis. Differently, several other approaches aim to learn the inter-instance relations in histopathology via Graph Neural Networks (GNNs) [3,36,45,38,29,1,2] or CNNs [42,39,26].

2.2 Instance Selection Strategies in Histopathology

Most MIL methods encode all patches in a WSI irrespective of their functional types. This compels MIL to be computationally expensive for large WSIs. To reduce the computational memory requirements, [26] randomly sampled a subset of instances, with the consequence of potentially missing vital information, especially when the informative set is small, *e.g.*, in metastasis detection. Differently, reinforcement learning-based methods [13,37] have also been developed to this end. [37] proposed to sequentially identify some of the diagnostically relevant RoIs in a WSI by following a parameterized policy. However, the method leverages a very coarse context for the RoI identification and is limited to utilizing only single-scale information for the diagnosis. Additionally, the reinforcement learning method [13] and the recurrent visual attention-based model [7] aim to select patches, which mimics pathological diagnosis. However, these methods require pixel-level annotations to learn discriminative regions, which is expensive to acquire on large WSIs. In contrast to the above methods, ZOOMMIL requires only WSI-level supervision. Our method is flexible to attend to several magnifications, while efficiently classifying WSIs with high performance.

The attention-score-based iterative sampling strategy proposed in [25,23] closely relates to our work. For the final classification, the selected patch embeddings are simply concatenated, analogous to average pooling. Instead, ZOOMMIL incorporates a dual gated-attention module between two consecutive magnifications to simultaneously learn to select the relevant instances to be zoomed in on, and learn an improved WSI-level representation for the lower magnification.

The patch selection module employed in our work is inspired by the perturbed optimizer-based [8] differentiable Top-K algorithm proposed in [11]. ZOOMMIL advances upon [11] by extending to several magnifications, *i.e.*, multi-level zooming, and scaling the applications to gigapixel-sized WSIs.

3 MIL with Differentiable Zooming

In this section, we present ZOOMMIL, which identifies informative patches at low magnification and zooms in on them for fine-grained analysis. In Sec. 3.1, we introduce the gated-attention mechanism determining the informative patches at a given magnification. In Sec. 3.2, we describe how to enable the attention-based patch selection to be differentiable while employing multiple magnifications. Finally, we present in Sec. 3.3 our overall architecture, in particular our proposed Dual Gated Attention and multi-scale information aggregation.

3.1 Attention-based MIL

In MIL, an input X is considered as a bag of instances $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. Given a classification task with C labels, there exists an *unknown* label $\mathbf{y}_i \in C$ for each instance and a *known* label $\mathbf{y} \in C$ for the bag. In our context, the input is a WSI and the instances denote the extracted patches. We follow the embedding-based MIL approaches [20,32,40], where a patch-level feature extractor h maps each patch \mathbf{x}_i to a feature vector $\mathbf{h}_i = h(\mathbf{x}_i) \in \mathbb{R}^D$. Afterwards, a pooling operator $g(\cdot)$ aggregates the feature vectors $\mathbf{h}_{i=1:N}$ to a single WSI-level feature representation. Finally, a classifier $f(\cdot)$ uses the WSI representation to predict the WSI-level label $\hat{\mathbf{y}} \in C$. The end-to-end process can be summarized as:

$$\hat{\mathbf{y}} = f\left(g\left(\{h(\mathbf{x}_1), \dots, h(\mathbf{x}_N)\}\right)\right). \quad (1)$$

To aggregate the patch features, we use attention-pooling, specifically, Gated Attention (GA) from [20]. Let $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_N]^\top \in \mathbb{R}^{N \times D}$ be the patch-level feature matrix, then the WSI-level representation \mathbf{g} is computed as:

$$\mathbf{g} = \sum_{i=1}^N a_i \mathbf{h}_i, \quad a_i = \frac{\exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_i) \odot \eta(\mathbf{U}\mathbf{h}_i))\}}{\sum_{j=1}^N \exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_j) \odot \eta(\mathbf{U}\mathbf{h}_j))\}}, \quad (2)$$

where $\mathbf{w} \in \mathbb{R}^{L \times 1}$, $\mathbf{V} \in \mathbb{R}^{L \times D}$, $\mathbf{U} \in \mathbb{R}^{L \times D}$ are learnable parameters with hidden dimension L , \odot is element-wise multiplication, and $\eta(\cdot)$ is the sigmoid function. While previous attention-based MIL methods [20,32] were designed to operate at a *single* magnification, we propose an efficient and flexible framework that can be extended to arbitrarily many magnifications while being fully differentiable.

3.2 Attending to Multiple Magnifications

We assume the WSI is accessible at magnifications indexed by $m \in \{1, \dots, M\}$, where the highest magnification is at M and the magnification at $m+1$ is twice that at m , consistent with the pyramidal format of WSIs. To efficiently extend MIL to multiple magnifications, we hierarchically identify informative patches from low-to-high magnifications and aggregate their features to get the WSI representation. To identify the patches at m , we first compute $\mathbf{a}_m \in \mathbb{R}^N$,

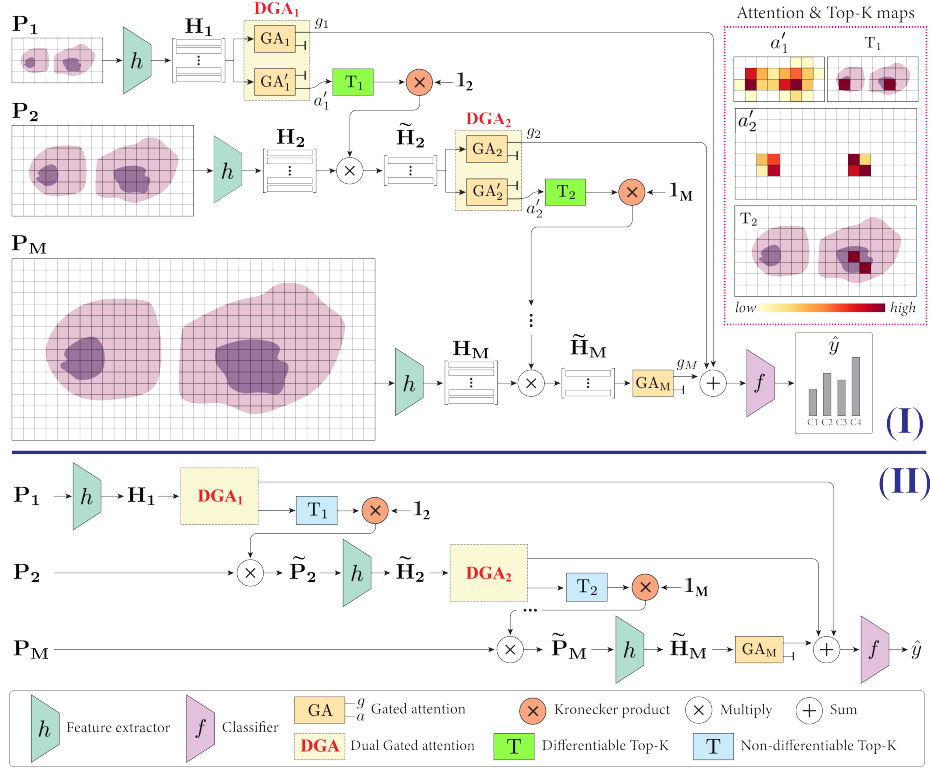


Fig. 2. Overview of the proposed ZOOMMIL. **(I)** and **(II)** present the distinct training and inference modes, generically exemplified for M magnifications.

which includes an attention score per patch. Then, the top K patches with the highest scores are selected for further processing at a higher magnification. The corresponding selected patch feature matrix is denoted by

$$\tilde{\mathbf{H}}_m = \mathbf{T}_m^\top \mathbf{H}_m, \quad (3)$$

where $\mathbf{T}_m \in \{0, 1\}^{N \times K}$ is an indicator matrix and $\mathbf{H}_m \in \mathbb{R}^{N \times D}$ is the patch feature matrix at m .

Instead of a handcrafted approach, we propose to drive the patch selection at m directly by the prediction output of $f(\cdot)$. This could be achieved via a backpropagation path from the output of $f(\cdot)$ to the attention module at m , without introducing any additional loss or associated hyperparameters. However, this naive formulation is non-differentiable as it involves a Top-K operation. To address this problem, we build on the perturbed maximum method [8] to make the Top-K selection differentiable, inspired by [11], and apply it to the attention weights \mathbf{a}_m at magnification m . Specifically, \mathbf{a}_m is first perturbed by adding uniform Gaussian noise $\mathbf{Z} \in \mathbb{R}^N$. Then, a linear program is solved for each of the perturbed attention weights, and their results are averaged. The forward pass of

the differentiable Top-K module can thus be written as:

$$\mathbf{T} = \mathbb{E}_{\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\arg \max_{\hat{\mathbf{T}}} \langle \hat{\mathbf{T}}, (\mathbf{a}_m + \sigma \mathbf{Z}) \mathbf{1}^\top \rangle \right], \quad (4)$$

where $\mathbf{1}^\top = [1 \cdots 1] \in \mathbb{R}^{1 \times K}$ and $(\mathbf{a}_m + \sigma \mathbf{Z}) \mathbf{1}^\top \in \mathbb{R}^{N \times K}$ denotes the perturbed attention weights repeated K times, and $\langle \cdot \rangle$ is a scalar product preceded by a vectorization of the matrices. The corresponding Jacobian is defined as:

$$J_{\mathbf{a}_m} \mathbf{T} = \mathbb{E}_{\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\arg \max_{\hat{\mathbf{T}}} \langle \hat{\mathbf{T}}, (\mathbf{a}_m + \sigma \mathbf{Z}) \mathbf{1}^\top \rangle \mathbf{Z}^\top / \sigma \right]. \quad (5)$$

More details on the derivation are provided in the supplemental material. The differentiable Top-K operator enables to *learn* the parameters of the attention module that weighs the patches at specific magnifications. Unlike [11], where patch sizes are scaled proportionally to the magnifications, we maintain a constant patch size across magnifications. This renders the number of patches proportional to the magnifications. It also provides different fields-of-view of the tissue microenvironment and enables us to capture a variety of contexts. This is crucial for analyzing WSIs as they contain diagnostically relevant constituents of various sizes. To achieve the zooming objective, we expand the indicator matrix \mathbf{T}_m to select from the patch features $\mathbf{H}_{m'} \in \mathbb{R}^{N \cdot 4^{(m'-m)} \times D}$, where $m' > m$. Specifically, we compute the Kronecker product between \mathbf{T}_m and the identity matrix $\mathbb{1}_{m'} = \text{diag}(1, \dots, 1) \in \mathbb{R}^{4^{(m'-m)} \times 4^{(m'-m)}}$ to obtain the expanded indicator matrix $\mathbf{T}_{m'} \in \{0, 1\}^{N \cdot 4^{(m'-m)} \times K \cdot 4^{(m'-m)}}$. Analogously to Eq. (3), patch selection at m' using the attention weights from m can be performed using

$$\tilde{\mathbf{H}}_{m'} = (\mathbf{T}_m \otimes \mathbb{1}_{m'})^\top \mathbf{H}_{m'}, \quad (6)$$

where $\mathbf{H}_{m'}$ is the feature matrix at m' and $\tilde{\mathbf{H}}_{m'}$ is the selected feature matrix.

3.3 Dual Gated Attention and Multi-Scale Aggregation

Figure 2 shows ZOOMMIL in its training (I) and inference (II) mode.

Training mode: The feature matrix \mathbf{H}_1 at $m=1$ passes through a Dual Gated Attention (DGA) block. DGA consists of two gated-attention modules GA_1 and GA'_1 . GA_1 is trained to obtain an optimal attention-pooled WSI-level representation \mathbf{g}_1 at low magnification. GA'_1 calculates attention weights \mathbf{a}'_1 that are used to identify important patches to zoom in. Alternatively, a single attention module could be used for both tasks. However, this would prevent optimal zooming, as the selected low-magnification patches would aim to optimize the classification performance only with information from the low magnification. Employing separate attention modules decouples the optimization tasks, and in turn, enables to obtain *complementary* information from both magnifications. Subsequently, the differentiable Top-K selection module, \mathbf{T}_1 , is employed to learn to select the most informative patches. The following selected higher-magnification patch feature matrix $\tilde{\mathbf{H}}_2$ is obtained via Eq. (6).

The process of selecting patch features for every subsequent higher magnification is repeated until the highest magnification M . The selected patch features $\tilde{\mathbf{H}}_M$ at M go through a last gated-attention block GA_M to produce \mathbf{g}_M . Finally, the attention-pooled features from all magnifications, $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_M$, are aggregated via sum-pooling to get a multi-scale, context-aware representation for the WSI. Inspired by residual learning [18], sum-pooling is used, as the features across different magnifications are closely related and the summation leverages their complementarity. The final classifier $f(\cdot)$ maps the WSI representation to the label $y \in C$ by producing the model prediction \hat{y} . The training phase can be regarded as extending Eq. (1) with sum-pooling over multiple magnifications:

$$\hat{y} = f\left(\mathbf{g}_1(\mathbf{H}_1) + \mathbf{g}_2(\tilde{\mathbf{H}}_2), \dots + \mathbf{g}_M(\tilde{\mathbf{H}}_M)\right). \quad (7)$$

Inference mode: The differentiable Top-K operator in our model learns to identify informative patches during training. However, this operator includes random perturbations to the attention weights, and thus makes the forward pass of the model non-deterministic. Therefore, we replace differentiable Top-K with conventional non-differentiable Top-K during inference, which is also faster as no perturbations have to be computed. As shown in Figure 2, another crucial difference to the training mode is that the patch selection directly operates on the WSI patches, $\mathbf{P}_{m'} \in \mathbb{R}^{N \cdot 4^{(m'-1)} \times p_h \times p_w \times p_c}$, instead of the pre-extracted patch features $\mathbf{H}_{m'}$. This avoids the extraction of features for uninformative patches during inference, unlike other MIL methods. It significantly reduces the computational requirements and speeds up model inference.

4 Experiments

4.1 Datasets

We benchmark ZOOMMIL on three H&E stained, public WSI datasets.

CRC [34] contains 1133 colorectal biopsy and polypectomy slides from *non-neoplastic*, *low-grade*, and *high-grade* lesions, accounting for 26.5%, 48.7%, 24.8% of the data. The slides were acquired at the IMP Diagnostics laboratory, Portugal, and were digitized by a Leica GT450 scanner at $40\times$. We split the data into 70%/10%/20% stratified sets for training, validation, and testing.

BRIGHT [9] consists of breast WSIs from *non-cancerous*, *precancerous*, and *cancerous* subtypes. The slides were acquired at the Fondazione G. Pascale, Italy, and scanned by an Aperio AT2 scanner at $40\times$. We used the BRIGHT challenge splits¹ containing 423, 80, and 200 WSIs for training, validation, and testing.

CAMELYON16 [5] includes 270 WSIs, 160 normal and 110 with metastases, for training, and 129 slides for testing. The slides were scanned by 3DHIS-TECH and Hamamatsu scanners at $40\times$ at the Radboud University Medical Center and the University Medical Center Utrecht, Netherlands. We split the 270 slides into 90%/10% stratified sets for training and validation.

¹ www.research.ibm.com/haifa/Workshops/BRIGHT

The average number of (pixels, patches), within the tissue area, at $20\times$ magnification for CRC, BRIGHT, and CAMELYON16 datasets are (227.28 Mpx, 3468), (1.04 Gpx, 15872), and (648.28 Mpx, 9892), respectively.

4.2 Implementation details

Preprocessing: For each WSI, we detect the tissue area using a Gaussian tissue detector [21] and divide the tissue into 256×256 patches at all considered magnifications. We ensure that each high-magnification patch is associated with the corresponding lower-magnification patch. We encode the patches with ResNet-50 [18] pre-trained on ImageNet [12] and apply adaptive average pooling after the third residual block to obtain 1024-dimensional embeddings.

ZoomMIL: The gated-attention module comprises three 2-layer Multi-Layer Perceptrons (MLPs), where the first two are followed by Hyperbolic Tangent and Sigmoid activations, respectively. The classifier is a 2-layer MLP with ReLU activation. We use a dropout probability of 0.25 in all fully-connected layers.

Implementation: All methods are implemented in PyTorch [35] and run on a single NVIDIA A100 GPU. ZOOMMIL uses $K = \{16, 12, 300\}$ on CRC, BRIGHT, and CAMELYON16, respectively, and our more efficient variant ZOOMMIL-EFF uses $K = \{12, 8\}$ on CRC and BRIGHT, respectively. We use the Adam optimizer [24] with 0.0001 learning rate and plateau scheduler (patience=5 epochs, decay rate=0.8). The experiments are run for 100 epochs with a batch size of one. For CRC & CAMELYON16, the models with the best validation loss are saved for testing. On BRIGHT, we observed that the baselines perform poorly compared to ZOOMMIL when using validation loss as the model selection criterion. We therefore employ best validation weighted-F1 for model selection on BRIGHT since it improves the baselines, giving them a better competitive chance against ZOOMMIL.

4.3 Results and Discussion

Baselines: We compare ZOOMMIL with state-of-the-art MIL methods. Specifically, we compare with ABMIL [20], which uses a gated-attention pooling, and its variant CLAM [32], which also includes an instance-level clustering loss. We further compare with two spatially-aware methods, namely, TRANSMIL [40] which models instance-level dependencies using transformer-based pooling, and SPARSECONVMIL [26] which selects random subsets of patches and employs sparse convolutions for pooling. In addition, we compare with multi-scale methods MSMIL [17] and DSMIL [27], which are computationally less efficient than ZOOMMIL as they encode all patches in a WSI across all considered magnifications. For completeness, we also include vanilla MIL methods based on max-pooling (MAXMIL) [26] and mean-pooling (MEANMIL) [26], following SPARSECONVMIL’s strategy of random patch selection. Additional implementation details and hyper-parameters are provided in the supplemental material. For a fair comparison, preprocessing including the extraction of patch embeddings is done consistently in the same manner, as described in Section 4.2.

Table 1. Performance and efficiency measurement on CRC [34]. The best and second-best classification results are in **bold** and underlined, respectively.

Methods	Classification		Computation	
	Weighted-F1(%)	Accuracy(%)	TFLOPs	Time(s)
MAXMIL [26] (20 \times)	82.2 \pm 0.9	82.2 \pm 1.2	0.96	0.13
MEANMIL [26] (20 \times)	84.3 \pm 0.8	84.1 \pm 1.2	0.96	0.12
SPARSECONVMIL [26] (20 \times)	89.6 \pm 1.3	89.6 \pm 0.9	0.96	0.13
ABMIL [20] (20 \times)	90.1 \pm 0.6	90.2 \pm 0.5	13.63	4.85
CLAM-SB [32] (20 \times)	90.9 \pm 0.6	90.9 \pm 0.5	13.63	4.85
TRANSMIL [40] (20 \times)	89.8 \pm 1.1	90.2 \pm 0.9	13.63	4.85
MSMIL [17] (5 \times + 10 \times + 20 \times)	84.6 \pm 0.1	84.9 \pm 0.2	17.88	6.37
DSMIL [27] (5 \times + 10 \times + 20 \times)	<u>91.1\pm1.1</u>	<u>91.2\pm1.1</u>	17.94	6.37
ZOOMMIL-EFF (5 \times \rightarrow 10 \times)	90.3 \pm 1.3	90.3 \pm 1.3	1.06	0.38
ZOOMMIL (5 \times \rightarrow 10 \times \rightarrow 20 \times)	92.0\pm0.6	92.1\pm0.7	1.40	0.50

Table 2. Performance and efficiency measurement on BRIGHT [9]. The best and second-best classification results are in **bold** and underlined, respectively.

Methods	Classification		Computation	
	Weighted-F1	Accuracy	TFLOPs	Time(s)
MAXMIL [26] (10 \times)	46.8 \pm 3.7	51.3 \pm 1.7	0.96	0.13
MEANMIL [26] (10 \times)	44.9 \pm 2.8	47.1 \pm 0.1	0.96	0.12
SPARSECONVMIL [26] (10 \times)	53.2 \pm 3.6	55.3 \pm 3.7	0.96	0.13
ABMIL [20] (10 \times)	63.5 \pm 2.7	65.5 \pm 1.9	16.45	5.86
CLAM-SB [32] (10 \times)	63.1 \pm 1.7	64.3 \pm 1.7	16.45	5.86
TRANSMIL [40] (10 \times)	65.5 \pm 2.8	66.0 \pm 2.7	16.46	5.86
MSMIL [17] (1.25 \times + 2.5 \times + 10 \times)	61.7 \pm 0.6	62.5 \pm 1.1	21.59	7.69
DSMIL [27] (1.25 \times + 2.5 \times + 10 \times)	63.1 \pm 1.6	64.0 \pm 1.1	21.66	7.69
ZOOMMIL-EFF (1.25 \times \rightarrow 2.5 \times)	66.0 \pm 1.9	<u>66.5\pm1.5</u>	0.40	0.14
ZOOMMIL (1.25 \times \rightarrow 2.5 \times \rightarrow 10 \times)	68.3\pm1.1	69.3\pm1.0	1.29	0.46

WSI classification performance: We present the classification results in terms of weighted F1-score and accuracy in Table 1, 2, and 3. Mean \pm standard deviation of the metrics is computed over three runs with different weight initializations. Corresponding magnifications of operation are shown alongside each method for each dataset. We include two versions of ZOOMMIL using either 2 or 3 magnifications, denoted as ZOOMMIL-EFF (efficient) and ZOOMMIL.

On CRC, ZOOMMIL outperforms CLAM-SB and TRANSMIL by 1.1% and 2.2% weighted F1-score, and ZOOMMIL-EFF achieves comparable performance. Furthermore, ZOOMMIL shows superior performance compared to the multi-scale methods MSMIL and DSMIL. For the individual classes, ZOOMMIL achieves 94.3%, 93.6%, and 86.4% average F1-scores in the one-vs-rest setting.

Table 3. Performance and efficiency measurement on CAMELYON16 [5]. The best and second-best classification results are in **bold** and underlined, respectively.

Methods	Classification		Computation	
	Weighted-F1(%)	Accuracy(%)	TFLOPs	Time(s)
MAXMIL[26] (20 \times)	64.0 \pm 3.0	67.1 \pm 0.9	0.96	0.13
MEANMIL [26] (20 \times)	63.5 \pm 1.1	65.9 \pm 1.6	0.96	0.12
SPARSECONVMIL [26] (20 \times)	67.7 \pm 0.6	68.7 \pm 0.1	0.96	0.13
ABMIL [20] (20 \times)	83.2 \pm 1.7	84.0 \pm 1.3	39.12	13.92
CLAM-SB [32] (20 \times)	83.3 \pm 1.5	84.0 \pm 1.3	39.12	13.92
TRANSMIL [40] (20 \times)	83.6\pm2.6	85.3\pm1.9	39.12	13.92
MSMIL [17] (10 \times + 20 \times)	81.4 \pm 1.1	82.4 \pm 1.0	48.87	17.41
DSMIL [27] (10 \times + 20 \times)	78.5 \pm 0.42	79.6 \pm 0.3	48.95	17.41
ZOOMMIL (10 \times \rightarrow 20 \times)	<u>83.3\pm0.3</u>	<u>84.2\pm0.4</u>	14.94	5.32

WSIs in BRIGHT are 4.5 \times larger than in CRC and thus provide a better evaluation ground for efficient scaling. ZOOMMIL achieves the best performance, outperforming MSMIL by 6.6%, CLAM-SB and DSMIL by 5.2%, and TRANSMIL by 2.8% in weighted F1-score. Notably, ZOOMMIL-EFF achieves the second-best results. For the individual classes, ZOOMMIL reaches average F1-scores of 70.4%, 56.5%, and 77.8%. The performance is lowest for the challenging pre-cancerous class, which often resembles the other two classes.

For CAMELYON16, we set the lowest magnification to 10 \times as the metastatic regions can be extremely small (see Figure 4). Nevertheless, it still adversely impacts the performance, resulting in 1.1% lower average accuracy than TRANSMIL. However, this translates to misclassifying only 1-2 test WSIs.

Overall, ZOOMMIL performs better on CRC and BRIGHT, while being comparable to the state of the art on CAMELYON16. It also consistently outperforms ZOOMMIL-EFF, highlighting the apparent performance-efficiency trade-off, *i.e.*, performance reduction in exchange for gains in computational efficiency.

Efficiency measurements: We analyze the efficiency in terms of FLOPs and average processing time for inference (see Table 1, 2, and 3). Note that the computational cost in the MIL modules is negligible compared to patch feature extraction, which is computationally the most expensive. The FLOPs and processing time for different methods can therefore appear to be equal as their difference only becomes visible several digits after the decimal point. On CRC, ZOOMMIL uses $\approx 10\times$ less FLOPs and time than CLAM-SB and TRANSMIL. Compared to MSMIL and DSMIL, this factor increases to $>12\times$. On BRIGHT, our efficient variant reduces computational requirements by $>50\times$ compared to MSMIL and DSMIL, and $>40\times$ compared to CLAM-SB and TRANSMIL while providing comparable performance. On CAMELYON16, ZOOMMIL uses $\approx 1/3$ FLOPs compared to MSMIL, DSMIL, CLAM-SB, and TRANSMIL. The relatively lower efficiency gain is due to the fact that metastatic regions occupy only

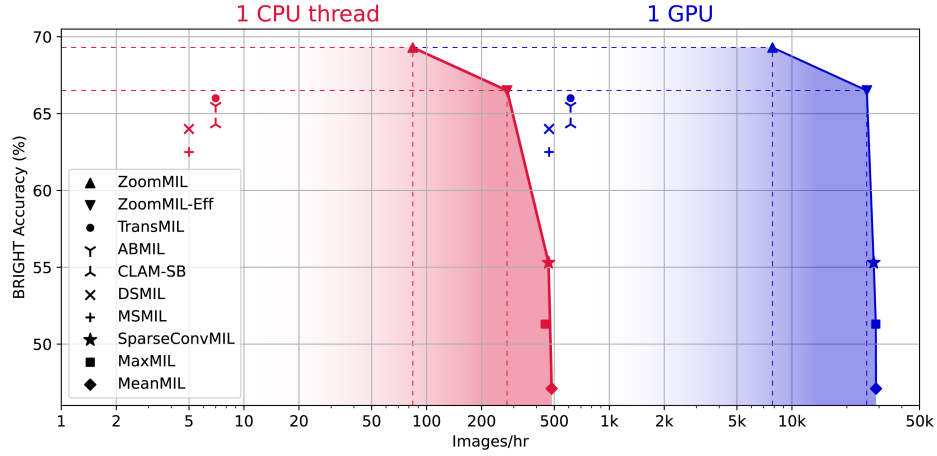


Fig. 3. Throughput vs classification accuracy for different MIL methods on BRIGHT, (left) on 1 single-core CPU, (right) on 1 NVIDIA A-100 GPU. Efficiency frontier curves are drawn in red and blue for CPU and GPU, respectively.

a small fraction of a WSI, and thus need to be analyzed at a finer magnification. Across all datasets, the methods adopting random patch selection (MAXMIL, MEANMIL, and SPARSECONVMIL) have similar computational requirements as ZOOMMIL but perform significantly worse.

To further highlight our efficiency gain, we show in Figure 3 the model throughput (images/hour) against the performance (accuracy) for all methods on BRIGHT. The marked efficiency frontier curves signify the best possible accuracies for different minimal throughput requirements. Noticeably, ZOOMMIL-EFF running on a single-core CPU processor (≈ 300 images/h) provides similar throughput to MSMIL, DSMIL, CLAM-SB, and TRANSMIL running on a cutting-edge NVIDIA A100 GPU (≈ 500 – 600 images/h). ZOOMMIL’s low computational requirements make it more practical and suitable for clinical deployment, where IT infrastructures are often under-developed and need large investments to establish and maintain a digital workflow.

Interpretability We interpret ZOOMMIL by qualitatively analyzing its patch-level attention maps. Figure 4(a,b) show the maps for two cancerous WSIs in BRIGHT at $1.25\times$, and Figure 4(c-f) show the maps for four metastatic WSIs in CAMELYON16 at $10\times$. We further include corresponding tumor regions annotated by an expert pathologist for comparison. Brighter regions in the maps mark higher attention scores, *i.e.*, more influential for model prediction.

For the BRIGHT WSIs, ZOOMMIL correctly attends to cancerous areas in (a,b), pays lower attention to the pre-cancerous area in (b), and least attention to the remaining non-cancerous areas that include non-cancerous epithelium, stroma, and adipose tissue. For the CAMELYON16 WSIs, (c,d) are correctly classified as ZOOMMIL gives high attention to the metastatic regions of differ-

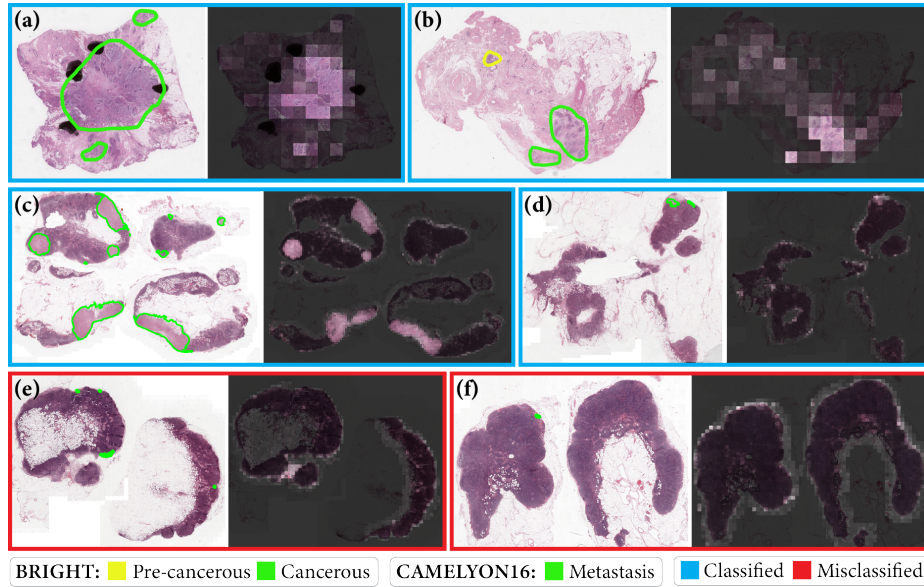


Fig. 4. Annotated tumor regions and attention maps from the lowest magnification of ZOOMMIL are presented for (a,b) BRIGHT and (c-f) CAMELYON16 WSIs.

ent sizes. However, the extremely small metastases in (e,f) get low attention and are disregarded by the Top-K module leading to misclassifying the WSIs. Notably, for cases with tiny metastases, relatively higher attention is imparted to the periphery of the tissues. This is consistent with the fact that metastases generally appear in the subcapsular zone of lymph nodes, as can be observed in (c-f). The presented visualizations are obtained from low magnifications, which signifies ZOOMMIL’s ability to learn to zoom in. More interpretability maps for other classes and fine-grained attention maps from higher attention modules in ZOOMMIL are provided in the supplemental material.

Ablation study: We ablated different modules in ZOOMMIL-EFF, due to its simple 2-magnification model. The results on BRIGHT are given in Table 4.

Differentiable patch selection: We compared our attention-based differentiable patch selection (DIFF-TOPK) against three alternatives: random selection at the lowest magnification (RANDOM K @ 1.25 \times), random selection at the highest magnification (RANDOM 4K @ 2.5 \times), and the non-differentiable Top-K selection (NONDIFF-TOPK) at the lowest magnification. The top rows in Table 4 show the superiority of DIFF-TOPK. Due to its differentiability, it learns to select patches via the gradient optimization of the model’s prediction.

Dual gated attention: We examined DGA consisting of two separate gated attention modules GA_1 and GA'_1 at low magnification, as discussed in Section 3.3. The former computes a slide-level representation and the latter learns

Table 4. Ablation study on BRIGHT, with the varied algorithmic component tabulated in the left-most column. All experiments use $K = 8$.

	Methods	Weighted F1(%)	Accuracy(%)
Patch sel.	RANDOM K @ $1.25\times$	61.0	61.0
	RANDOM 4K @ $2.5\times$	59.6	60.0
	NONDIFF-TOPK K @ $1.25\times$	59.9	60.0
	DIFF-TOPK K @ $1.25\times$ (Ours)	68.1	68.0
Attn.	Single GA @ $1.25\times$	59.6	61.0
	DGA @ $1.25\times$ (Ours)	68.1	68.0
Feat.	Features @ $2.5\times$	62.7	63.5
	Features @ $1.25\times$ @ $2.5\times$	64.9	65.0
	Features @ $1.25\times$ + @ $2.5\times$ (Ours)	68.1	68.0

to select patches at higher magnification. We can conclude from Table 4 that two separate attentions lead to better patch selection and improved slide representation for overall improved classification.

Feature aggregation: We aggregate slide-level representations across magnifications through sum-pooling, as shown in Eq. (7). Among several alternatives, we compared with: using the highest-magnification features (Features@ $2.5\times$) and fusing representations via concatenation (represented as @ $1.25\times$ || @ $2.5\times$). Table 4 shows that concatenation improves performance, indicating the value of multi-scale information. However, our sum-pooling, which is inspired by residual learning [18], significantly outperforms concatenation as it leverages the complementarity of the two magnifications more effectively.

5 Conclusion

In this work, we introduced ZOOMMIL, a novel framework for WSI classification. The method is more than an order of magnitude faster than previous state-of-the-art methods during inference while achieving comparable or better accuracy. Essential for our method is the concept of differentiable zooming that allows the model to learn which patches are informative and thus worth zooming in on. We conduct extensive quantitative and qualitative evaluations on three different datasets and demonstrate the importance of each component in our model with a detailed ablation study. Finally, we show that ZOOMMIL is a modular architecture that can easily be deployed in different flavors, depending on the performance-efficiency requirements in a given application. In future work, it would be interesting to further study the attention maps of ZoomMIL and compare them with the visual attention of pathologists.

References

1. Adnan, M., Kalra, S., Tizhoosh, H.: Representation learning of histopathology images using graph neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 988–989 (2020)
2. Anklin, V., Pati, P., Jaume, G., Bozorgtabar, B., Foncubierta-Rodriguez, A., Thiran, J., Sibony, M., Gabrani, M., Goksel, O.: Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 636–646 (2021)
3. Aygüneş, B., Aksoy, S., Cinbiş, R., Kösemehmetoğlu, K., Önder, S., Üner, A.: Graph convolutional networks for region of interest classification in breast histopathology. In: SPIE Medical Imaging 2020: Digital Pathology. vol. 11320 (2020)
4. Bejnordi, B., Litjens, G., Hermesen, M., Karssemeijer, N., van der Laak, J.: A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In: SPIE Medical Imaging 2015: Digital Pathology. vol. 9420 (2015)
5. Bejnordi, B., Veta, M., van Dienst, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017)
6. Bejnordi, B., Zuidhof, G., Balkenhol, M., Hermesen, M., Bult, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.: Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *Journal of Medical Imaging* **4** (2017)
7. BenTaieb, A., Hamarneh, G.: Predicting cancer with a recurrent visual attention model for histopathology images. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 129–137 (2018)
8. Berthet, Q., Blondel, M., Teboul, O., Cuturi, M., Vert, J., Bach, F.: Learning with differentiable perturbed optimizers. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 9508–9519 (2020)
9. Brancati, N., Anniciello, A., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta-Rodríguez, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. arXiv:2111.04740 (2021)
10. Campanella, G., Hanna, M., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K., Brogi, E., Reuter, V., Klimstra, D., Fuchs, T.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* **25**, 1301–1309 (2019)
11. Cordonnier, J., Mahendran, A., Dosovitskiy, A.: Differentiable patch selection for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2351–2360 (2021)
12. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
13. Dong, N., Kampffmeyer, M., Liang, X., Wang, Z., Dai, W., Xing, E.: Reinforced auto-zoom net: Towards accurate and fast breast cancer segmentation in whole-slide images. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshop. pp. 317–325 (2018)

14. Elmore, J., Longton, G., Carney, P., Geller, B., Onega, T., Tosteson, A., Nelson, H., Pepe, M., Allison, K., Schnitt, S., et al.: Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* **313**, 1122–1132 (2015)
15. Gao, Y., Liu, W., Arjun, S., Zhu, L., Ratner, V., Kurc, T., Saltz, J., Tannenbaum, A.: Multi-scale learning based segmentation of glands in digital colorectal pathology images. In: *SPIE Medical Imaging 2016: Digital Pathology*. vol. 9791 (2016)
16. Gomes, D., Porto, S., Balabram, D., Gobbi, H.: Inter-observer variability between general pathologists and a specialist in breast pathology in the diagnosis of lobular neoplasia, columnar cell lesions, atypical ductal hyperplasia and ductal carcinoma in situ of the breast. *Diagnostic Pathology* **9**, 1–9 (2014)
17. Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I.: Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3852–3861 (2020)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770–778 (2016)
19. Ho, D., Yarlagadda, D., D’Alfonso, T., Hanna, M., Grabenstetter, A., Ntiamoah, P., Brogi, E., Tan, L., Fuchs, T.: Deep multi-magnification networks for multi-class breast cancer image segmentation. *Computerized Medical Imaging and Graphics* **88** (2021)
20. Isle, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: *International Conference on Machine Learning (ICML)*. vol. 35 (2018)
21. Jaume, G., Pati, P., Anklin, V., Foncubierta-Rodríguez, A., Gabrani, M.: Histocartography: A toolkit for graph analytics in digital pathology. In: *MICCAI Workshop on Computational Pathology* (2021)
22. Jia, Z., Huang, X., Eric, I., Chang, C., Xu, Y.: Constrained deep weak supervision for histopathology image segmentation. *IEEE Transactions on Medical Imaging* **36**, 2376–2388 (2017)
23. Katharopoulos, A., Fleuret, F.: Processing megapixel images with deep attention-sampling models. In: *International Conference on Machine Learning (ICML)*. vol. 36 (2019)
24. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: *International Conference on Learning Representations (ICLR)* (2015)
25. Kong, S., Hénao, R.: Efficient classification of very large images with tiny objects. *arXiv:2106.02694* (2021)
26. Lerousseau, M., Vakalopoulou, M., Deutsch, E., Paragios, N.: Sparseconvmil: Sparse convolutional context-aware multiple instance learning for whole slide image classification. In: *MICCAI Workshop on Computational Pathology*. pp. 129–139 (2021)
27. Li, B., Li, Y., Eliceiri, K.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 14318–14328 (2021)
28. Li, J., Li, W., Sisk, A., Ye, H., Wallace, W.D., Speier, W., Arnold, C.: A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in Biology and Medicine* **131** (2021)

29. Li, R., Yao, J., Zhu, X., Li, Y., Huang, J.: Graph cnn for survival analysis on whole slide pathological images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 174–182 (2018)
30. Liang, Q., Nan, Y., Coppola, G., Zou, K., Sun, W., Zhang, D., Wang, Y., Yu, G.: Weakly supervised biomedical image segmentation by reiterative learning. *IEEE Journal of Biomedical and Health Informatics* **23**, 1205–1214 (2018)
31. Lu, M., Chen, T., Williamson, D., Zhao, M., Shady, M., Lipkova, J., Mahmood, F.: Ai-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021)
32. Lu, M., Williamson, D., Chen, T., Chen, R., Barbieri, M., Mahmood, F.: Data efficient and weakly supervised computational pathology on whole slide images. *Nature Biomedical Engineering* **5**, 555–570 (2021)
33. Myronenko, A., Xu, Z., Yang, D., Roth, H., Xu, D.: Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 329–338 (2021)
34. Oliveira, S., Neto, P., Fraga, J., Montezuma, D., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Pinto, I., Cardoso, J.: Cad systems for colorectal cancer from wsi are still not ready for clinical acceptance. *Scientific Reports* **11** (2021)
35. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 33, pp. 8024–8035 (2019)
36. Pati, P., Jaume, G., Foncubierta-Rodríguez, A., Feroce, F., Anniciello, A., Scognamiglio, G., Brancati, N., Fiche, M., Dubruc, E., Riccio, D., et al.: Hierarchical graph representations in digital pathology. *Medical Image Analysis* **75** (2021)
37. Qaiser, T., Rajpoot, N.: Learning where to see: A novel attention model for automated immunohistochemical scoring. *IEEE Transactions on Medical Imaging* **38**, 2620–2631 (2019)
38. Raju, A., Yao, J., Haq, M., Jonnagaddala, J., Huang, J.: Graph attention multi-instance learning for accurate colorectal cancer staging. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 529–539 (2020)
39. Shaban, M., Awan, R., Fraz, M., Azam, A., Tsang, Y., Snead, D., Rajpoot, N.: Context-aware convolutional neural network for grading of colorectal cancer histology images. *IEEE Transactions on Medical Imaging* **39**, 2395–2405 (2020)
40. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 35 (2021)
41. Sirinukunwattana, K., Alham, N., Verrill, C., Rittscher, J.: Improving whole slide segmentation through visual context - a systematic study. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 192–200 (2018)
42. Tellez, D., Litjens, G., van der Laak, J., Ciompi, F.: Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 567–578 (2019)
43. Tokunaga, H., Teramoto, Y., Yoshizawa, A., Bise, R.: Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12597–12606 (2019)

- 44. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis* **65** (2020)
- 45. Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., et al.: Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4837–4846 (2020)