DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification

Hongrun Zhang[1], Yanda Meng[1], Yitian Zhao[2], Yihong Qiao[3], Xiaoyun Yang[4], Sarah E. Coupland[1]
Yalin Zheng[1]✉
[1]University of Liverpool, [2]Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences
[3]China Science IntelliCloud Technology Co., Ltd, [4]Remark AI UK Limited, London
{hongrun.zhang, yanda.meng, S.E.Coupland, yalin.zheng}@liverpool.ac.uk
yitian.zhao@nimte.ac.cn, yihong.qiao@intellicloud.ai, xyang@remarkholdings.com

Abstract

Multiple instance learning (MIL) has been increasingly used in the classification of histopathology whole slide images (WSIs). However, MIL approaches for this specific classification problem still face unique challenges, particularly those related to small sample cohorts. In these, there are limited number of WSI slides (bags), while the resolution of a single WSI is huge, which leads to a large number of patches (instances) cropped from this slide. To address this issue, we propose to virtually enlarge the number of bags by introducing the concept of pseudo-bags, on which a double-tier MIL framework is built to effectively use the intrinsic features. Besides, we also contribute to deriving the instance probability under the framework of attention-based MIL, and utilize the derivation to help construct and analyze the proposed framework. The proposed method outperforms other latest methods on the CAMELYON-16 by substantially large margins, and is also better in performance on the TCGA lung cancer dataset. The proposed framework is ready to be extended for wider MIL applications. The code is available at: https://github.com/hrzhang1123/DTFD-MIL

1. Introduction

The automation of whole slide images (WSIs) poses a significant challenge to the field of computer vision. The increasing use of WSIs in histopathology results in digital pathology providing huge improvements in workflow and diagnosis decision-making by pathologists [7,21,24,29,31], but it also stimulates the need for intelligent or automatic analytical tools of WSIs [11, 20, 36, 40, 44, 48, 49]. WSIs have enormous sizes, ranging from 100M pixels to 10G pixels, and this unique characteristic makes it almost infeasible to directly transfer existing machine learning techniques to their applications, since these existing techniques were initially intended for natural images or medical images with much smaller sizes. When it comes to deep learning based models, large scale datasets and high quality annotations are the primary yet crucial conditions to train a high capacity model. However, the enormous sizes of WSIs bring along substantial burden for pixel-level annotation. This problem in turn encourages researchers to develop deep learning based models trained with limited annotations, termed as "Weakly Supervised" or "Semi-Supervised" [22,26,35,41]. A large proportion of existing weakly supervised works for WSI classification are characterized as "multiple instance learning" (MIL) [1,5,8,25]. Under the framework of MIL, a slide (or WSI), acting as a bag, constitutes multiple instances that are hundreds or thousands of patches cropped from the slide. With at least one instance being disease positive, the slide is marked as positive, or otherwise negative.

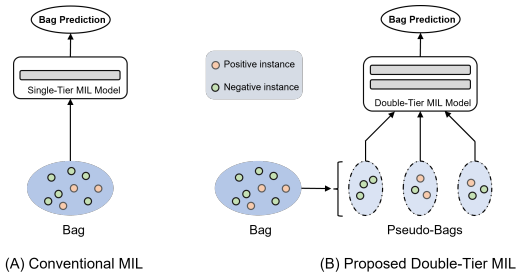

Figure 1. Illustration of the difference between conventional MIL models and the proposed double-tier MIL model.

There exist some successful attempts to solve the MIL problem in various computer vision tasks [19,27,28,30,32]. The innate characteristics of WSIs, however, make it less straightforward to develop MIL solution for WSI classification than the counterparts in other computer vision subfields, as the only direct guidance information for training

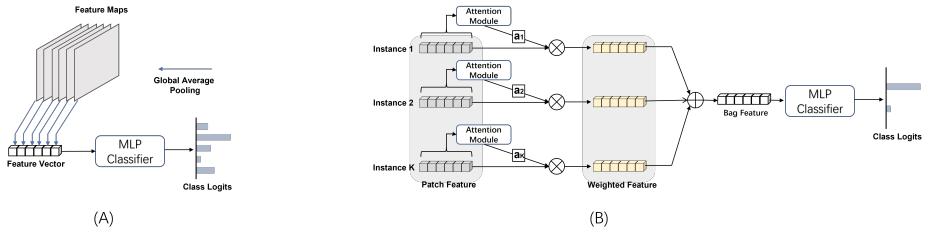

Figure 2. (A) Illustration of an image classification system of deep learning. Global average pooling is applied to the extracted feature maps of an image, leading to a feature vector representing the image. Then the feature vector is forwarded to a classifier which outputs the class logits, then class probabilities by softmax. (B) Illustration of the AB-MIL paradigm. The extracted feature of an instance is weighted by an attention score. The bag feature is obtained through the sum of the weighted instance features, and then fed into a classifier for the bag prediction.

are the labels of a few hundred slides. The most notorious consequence is the over-fitting problem where a machine learning model tends to fall into local minima during optimization while the learnt features are less relevant with respect to target disease, and as a result, the trained model has inferior generalization ability.

Most recent works of MIL for WSIs to tackle the overfitting issue are built on the essential idea of exploiting more information to learn in addition to the labels of the relatively small number of slides in a cohort. The mutual-instance relation is an important direction to explore for this purpose, and it has been empirically proven to be effective. The mutual-instance relation can be specified as spatial [6] or feature distances [18, 35, 37, 46], or can be agnostically learnt by neural modules, such as recurrent neural networks (RNN) [3], transformer [34], and graph convolution network [51].

Many of the aforementioned methods belong to attention-based MIL (AB-MIL) [14], although they differ in the formulations of the attention scores. However, it was believed infeasible to explicitly infer instance probabilities under AB-MIL frameworks [18], and as an alternative, attention scores were usually used as the indications of positive activation [10, 14, 18, 34]. In this paper, we argue that the attention score is not a rigorous metric for this purpose, and instead we contribute to deriving the instance probability under the framework of AB-MIL.

Given the huge size of a WSI, the units to be directly processed are the much smaller patches cropped from WSIs [12]. MIL models for WSI classification essentially aim to recognize the most distinctive patches that correspond mostly to the slide label. However, there are a limited number of slides while there are hundreds or even thousands of patches (instances) in a slide, and the information for learning are only the slide-level labels. Moreover, in many histopathology slides the positive regions corresponding to positive diseases only occupy small portions of tissue, leading to a small ratio of positive instances of a slide. Therefore, it is challenging to guide a model to recognize positive instances under the condition of MIL, since these factors collectively contribute to deteriorating the over-fitting problem.

Although most recent methods utilize mutual-instance relations to improve MIL, they do not explicitly confront the problems originated from the innate characteristics of WSIs as mentioned above. To alleviate the negative impacts of these problems, we introduce the concept of 'pseudo-bag' in the proposed framework. That is, we randomly split the instances (patches) of a bag (slide) into several smaller bags (pseudo-bags), and assign each pseudo bag the label of the original bag, termed as the parent bag. This strategy virtually increases the number of bags while inside each pseudo bag there are fewer instances; it also enables Double-Tier Feature Distillation MIL model (Fig.1). More specifically, a Tier-1 AB-MIL model is applied to the pseudo-bags of all the slides. However, it comes along the risk that a pseudo-bag from a positive parent bag may not be allocated with at least one positive instance, in which case a mislabeled pseudo-bag is introduced. To tackle this issue, we distill a feature vector from each pseudo-bag and establish a Tier-2 AB-MIL model upon the features distilled from all the pseudo-bags of a slide (See Fig.3). Through the distillation process, the Tier-1 model provides initial candidates of distinct features for the Tier-2 model to generate a better representation for the corresponding parent bag. Furthermore, for the purpose of feature distillation, we derive the instance probability under the framework of AB-MIL, by harnessing the fundamental idea of Grad-CAM that was developed for visualizing deep learning features [33].

In essence, we deal with the MIL problem for WSI classification from another perspective with the proposed double-tier MIL framework. The main contributions are: (1). We introduce the concept of pseudo-bags to alleviate the issue of limited number of WSIs. (2) We derive the instance probability under the AB-MIL framework by utilizing the essential idea of Grad-CAM. Given AB-MIL is the base for many MIL works, the instance probability derivation can help with the extension studies of related

MIL methods. (3). By utilizing the instance probability derivation, we formulate a double-tier MIL framework, and the experiments show its superiority to other latest methods on two large public histopathology WSI datasets.

2. Related Works

2.1. Multiple instance learning in WSI analysis

Providing the importance of weakly supervised learning, there is a trend to develop MIL algorithms for WSI analysis where, instead of elaborated pixel-level annotation, only the slide labels are available for training. MIL models generally can be divided into two groups, based on whether the final bag predictions are directly from instance predictions [3, 9, 12, 15, 17, 47], or from the aggregations of features of instance [14, 18, 23, 34, 35, 42, 53]. For the former, the bag predictions are usually obtained through either the average pooling (mean value of probabilities of instances) or the maximum pooling (max value of probabilities of instances). In contrast, the latter one learns a high-level representation of a bag and builds up a classifier upon this bag representation for the bag-level prediction, and it is usually referred to as the bag embedding method. Despite the simplicity and being straightforward, instance-level probability pooling is empirically proven to be inferior to the bag embedding counter-part in performance [34, 42].

Many of the bag embedding-based models adopt the basic idea of AB-MIL, i.e., the bag embedding (or bag representation) is obtained from the weighting of the features of individual instances, while the latest works of this kind differ in the ways to generate the weighting values, which are typically referred to as attention scores. For example, in the original paper [14], the attention scores are learned by a side-branch network, in DS-MIL [18] an attention score is based on the cosine distance between features of an instance and the critical instance, while in Trans-MIL [34], they are the output of a transformer architecture that encode the mutual-correlations between instances. Essentially, these methods are all AB-MIL, and to distinguish, we term the original AB-MIL [14] as the classic AB-MIL. The main component of our proposed method is also attention-based, but it is not restricted to the way the attention scores are generated. Without loss of generality, we adopt the classic AB-MIL as the base MIL model for each tier in the proposed framework. Please note that altering to other variants of AB-MIL will be straightforward, but is not the main focus of this paper.

2.2. Grad-based Class Activation Map

The class activation map (CAM) [52] originally serves as a spatial visualization tool to reveal the locations in an image that correspond to the classification by a deep learning model. As its generalized version, Grad-CAM (Grad-based Class Activation Map) [33] enables the generation of CAM from higher complex architectures of multi-layer perception (MLP). Many works utilized Grad-CAM not only as a powerful tool for offline model analysis but as an embedded component in the designed deep learning model for various applications. For example, one notable capability of CAM is the target localization of a model trained only with image labels; therefore, it prevails in weakly supervised tasks, such as segmentation [4, 13, 16, 43] and detection [38, 45, 50], or even knowledge distillation [39].

In this paper, we demonstrate that the framework of AB-MIL is a special case of the deep learning architecture for image classification. This finding enables the utilization of the mechanism of Grad-CAM to directly derive the positive probability of an instance under the framework of AB-MIL, and the derivation assists in constructing the proposed framework and also helps with the corresponding analysis.

3. Method

3.1. Revisit Grad-CAM and AB-MIL

3.1.1 Grad-CAM

A deep learning model for end-to-end image classification typically comprises of two modules: a deep convolution neural network (DCNN) for high-level feature extraction and a multi-layer perceptron (MLP) for classification. An image is fed into the DCNN to generate the features maps, which become a feature vector by a pooling operation. The feature vector is then forwarded to the MLP for final class probabilities (Fig.2.(a)). Suppose the final output feature maps from the DCNN is $\boldsymbol{U} \in \mathbb{R}^{D\times W\times H}$, with D being the number of channel and W, H being the dimensional sizes, respectively. Imposing global average pooling on the feature maps leads to a feature vector that represents the image,

$$\boldsymbol{f} = \underset{W,H}{\text{GAP}}\left(\boldsymbol{U}\right) \in \mathbb{R}^{D} \tag{1}$$

where $\underset{W,H}{\text{GAP}}\left(\boldsymbol{U}\right)$ denotes the global average pooling with respect to W, H, i.e., the d_{th} element of $\boldsymbol{f}$, $f_d = \frac{1}{WH}\sum_{w=1,h=1}^{W,H} U^d_{w,h}$. Using $\boldsymbol{f}$ as the input, the MLP outputs logits s^c for class $c \in \{1, 2, ..., C\}$, whose value indicates the signal strength for the image belonging to class c and the predicted class probability can then be obtained by soft-max operation accordingly. The class activation map for class c by Grad-CAM is defined as the weighted sum of the feature maps,

$$\boldsymbol{L}^c = \sum_{d}^{D} \beta^c_d \boldsymbol{U}^d, \quad \beta^c_d = \frac{1}{WH}\sum_{w,h}^{W,H}\left(\frac{\partial s^c}{\partial U^d_{w,h}}\right), \tag{2}$$

with $\boldsymbol{L}^c \in \mathbb{R}^{W\times H}$, and $L^c_{w,h}$ being the magnitude value at

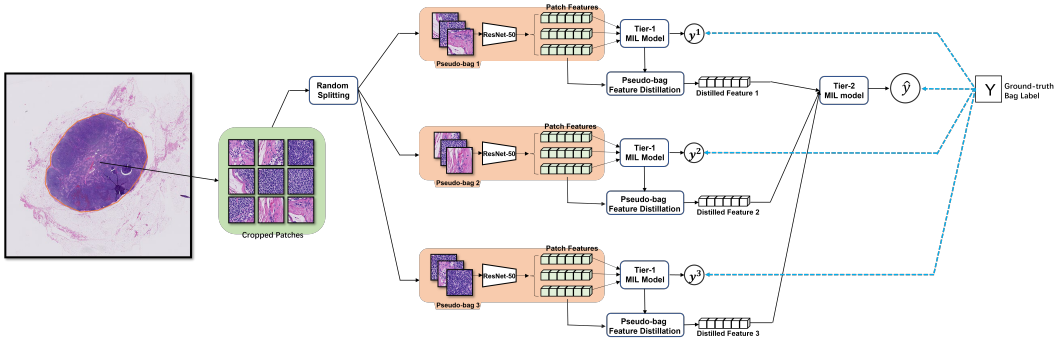

Figure 3. Overview of the proposed DTFD-MIL. A set of patches (we show only 9 for convenience) are first cropped from the tissue regions of a slide. These patches are randomly split into M pseudo-bags ($M = 3$ for example). A tier-1 MIL model is then applied to the 3 pseudo-bags, respectively. Based on the outputs of the Tier-1 model on the 3 pseudo-bags, 3 feature vectors are distilled accordingly and are then forwarded to the Tier-2 MIL model. The ground-truth bag label supervises both the Tier-1 and Tier-2 models during the training, denoted by blue dash lines.

location w, h of $\boldsymbol{L}^c$, indicating how strongly this location tends to be class c,

$$L^c_{w,h} = \sum_{d=1}^{D} \beta^c_d U^d_{w,h} \tag{3}$$

3.1.2 Attention-Based Multiple Instance Learning

Consider a bag of instances $\boldsymbol{X} = \{x_1, x_2, ...x_K\}$ with K being the number of instances in the bag. Each instance $x_k, k \in 1, 2, ..., K$ holds a latent label y_k ($y_k = 1$ for positive, or $y_k = 0$ for negative), which is assumed to be unknown. The goal of MIL is to detect whether there exist at least one positive instance in the bag. The only information revealed for training, however, is the bag label, which is defined as,

$$Y = \begin{cases} 1, & \text{if } \sum_{k=1}^{K} y_k > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

i.e., the bag is positive if at least one instance in it is positive, or negative otherwise. One straightforward solution for this learning problem is to assign each instance the bag label and accordingly train a classifier, and then apply the max or average pooling operation on the individual instance classifications to obtain the bag-level results [42]. Another popular strategy is to learn a bag representation $\boldsymbol{F}$ from the extracted features of instances in the bag, with which the problem becomes a conventional classification task, i.e., a classifier can be trained upon the bag representations. Empirically, the strategy of bag representation learning is proven to be more efficient than the instance pooling strategy, which we refer to as bag embedding-based MIL. The bag embedding is formulated as,

$$\boldsymbol{F} = \mathrm{G}\left(\{\boldsymbol{h}_k \mid k = 1, 2, ..., K\}\right), \tag{5}$$

where G is an aggregation function, and $\boldsymbol{h}_k \in \mathbb{R}^D$ is the extracted feature for instance k. Typically, many works adopt the attention tactic to obtain the bag representation (or embedding) as follows,

$$\boldsymbol{F} = \sum_{k=1}^{K} a_k \boldsymbol{h}_k \ \in \mathbb{R}^D, \tag{6}$$

where a_k is the learnable scalar weight for $\boldsymbol{h}_k$, and D is the dimension of vector $\boldsymbol{F}$ and $\boldsymbol{h}_k$. The paradigm is shown in Fig.2.(b). The attention mechanisms in [14, 18, 23] follow this formulation, therefore they all belong to the category of AB-MIL, but differ in the ways to generate the attention score (weight value) a_k. For example, the weight from the classic AB-MIL [14] is defined as,

$$a_k = \frac{\exp\{\boldsymbol{w}^{\mathrm{T}}(\tanh(\boldsymbol{V}_1\boldsymbol{h}_k) \odot \mathrm{sigm}(\boldsymbol{V}_2\boldsymbol{h}_k))\}}{\sum_{j=1}^{K} \exp\{\boldsymbol{w}^{\mathrm{T}}(\tanh(\boldsymbol{V}_1\boldsymbol{h}_j) \odot \mathrm{sigm}(\boldsymbol{V}_2\boldsymbol{h}_j))\}}, \tag{7}$$

where $\boldsymbol{w}$, $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are the learnable parameters.

3.2. Derivation of Instance Probability in AB-MIL

Despite the better performance of bag embedding-based MIL, it was thought infeasible to unravel instance class probability [18, 42]. In this paper, however, we show that it is possible to derive the predicted probability of each individual instance in a bag under the framework of AB-MIL. The derivation is rooted in the following proposition,

Proposition 1 *The paradigm of AB-MIL is a special case of the framework of the classic deep-learning network for image classification.*

Proof and explanation are in the Supplementary.

Based on Proposition 1, it is safe to apply the mechanism of Grad-CAM to AB-MIL to directly infer the signal

strength for an instance to be a certain class. Resembling to Eq.(2), the signal strength for instance k to be class c ($c = 0$ for negative and $c = 1$ for positive) can then be derived as (see Supplementary),

$$L_k^c = \sum_{d=1}^{D} \beta_d^c \hat{h}_{k,d}, \quad \beta_d^c = \frac{1}{K} \sum_{i=1}^{K} \frac{\partial s_c}{\partial \hat{h}_{k,d}} \tag{8}$$

where s_c is the output logit for class c from the MIL classifier, $\hat{h}_{k,d}$ is the d_{th} element of $\hat{\boldsymbol{h}}_k$, and $\hat{\boldsymbol{h}}_k = a_k K \boldsymbol{h}_k$ with a_k being the attention score for instance k defined in Eq.(6). By applying soft-max, the corresponding probability is then,

$$p_k^c = \frac{\exp\left(L_k^c\right)}{\sum_{t=1}^{C} \exp\left(L_k^t\right)} \tag{9}$$

3.3. Double-Tier Feature Distillation Multiple Instance Learning

In this section, we present the proposed double-tier feature distillation MIL framework.

Given N bags (slides), and in each bag there are K_n instances (patches), i.e., $\boldsymbol{X}_n = \{x_{n,k} \mid k = 1, 2, ..., K_n\}, n \in \{1, 2, ..., N\}$, with the ground-truth of a bag being Y_n. The corresponding feature of a patch, denoted as $\boldsymbol{h}_{n,k}$, is extracted by a backbone network H, i.e., $\boldsymbol{h}_{n,k} = \mathrm{H}(x_{n,k})$. The instances in a bag (slide) are randomly split into M pseudo-bags with approximately even number of instances, $\boldsymbol{X}_n = \{\boldsymbol{X}_n^m \mid m = 1, 2, ..., M\}$. A pseudo-bag is assigned the label of its parent bag's label, i.e., $Y_n^m = Y_n$. In Tier-1, an AB-MIL model, denoted as T_1, is applied to each pseudo-bag. Suppose the estimated bag probability of a pseudo-bag through the Tier-1 model is y_n^m,

$$y_n^m = \mathrm{T}_1\left(\{\boldsymbol{h}_k = \mathrm{H}(x_k) \mid x_k \in \boldsymbol{X}_n^m\}\right), \tag{10}$$

The Tier-1 loss function for training using cross entropy is then defined as,

$$\mathcal{L}_1 = -\frac{1}{MN} \sum_{n=1,m=1}^{N,M} Y_n^m \log y_n^m + (1 - Y_n^m) \log\left(1 - y_n^m\right) \tag{11}$$

The probabilities of instances in each pseudo-bag can then be derived using Eq.(8) and Eq.(9). Based on the derived instance probabilities, a feature from each pseudo-bag is distilled, denoted as $\hat{\boldsymbol{f}}_n^m$ for the m_{th} pseudo-bag of the n_{th} parent bag. All the distilled features are forwarded into a Tier-2 AB-MIL, denoted as T_2, for the inference of the parent bag,

$$\hat{y}_n = \mathrm{T}_2\left(\left\{\hat{\boldsymbol{f}}_n^m \mid m \in (1, 2, ..., M)\right\}\right) \tag{12}$$

The Tier-2 loss function for training T_2 is defined as,

$$\mathcal{L}_2 = -\frac{1}{N} \sum_{n=1}^{N} Y_n \log \hat{y}_n + (1 - Y_n) \log(1 - \hat{y}_n), \tag{13}$$

The overall optimization process is then:

$$\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\} = \arg\min_{\boldsymbol{\theta}_1} \mathcal{L}_1 + \arg\min_{\boldsymbol{\theta}_2} \mathcal{L}_2 \tag{14}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the parameters of T_1 and T_2, respectively. It should be noted that there is a certain proportion of noise labels for the pseudo bags, as a pseudo bag may not be assigned with at least one positive instance by the random allocation. However, deep neural networks are resilient to noise labels to some extent. Besides, the noise level can be roughly controlled by the number of pseudo bags in each parent bag, i.e., M. We show how M's value will affect the performance of the proposed methods in the ablation study section.

Four feature distillation strategies are considered as follows:

- **MaxS** Maximum selection: The feature of the instance in a pseudo-bag that achieves the maximum positive probability from the Tier-1 MIL model is selected to forward to the Tier-2 MIL model.
- **MaxMinS** MaxMin selection: The features of two instances in a pseudo-bag are distilled and concatenated to forward to the Tier-2 model: the one with the maximum probability and the one with the minimum probability in the pseudo-bag. Such a selection is based on the consideration that, if only the instance with maximum positive probability in each pseudo-bag are selected (as in the case of MaxS), the decision boundary of the trained Tier-2 model will tend to be pushed forward too tightly to positive samples, and may miss the genuinely positive cases that are similar to the negative ones [47]. By introducing the instances of maximum and minimum probabilities at the same time, it gives looser space for the Tier-2 model to generate the parent bag's feature embedding.
- **MAS** Maximum attention score selection: The feature of the instance in a pseudo-bag with maximum assigned attention score from the Tier-1 MIL model is distilled to the Tier-2 MIL model.
- **AFS** Aggregated feature selection: The feature aggregated from all the instances in a pseudo-bag as in Eq.(6) is forwarded the Tier-2 model.

We evaluate the performances of all these 4 strategies in the experimental section.

Table 1. Results on CAMELYON-16 test set. The subscripts are the corresponding 95% confidence intervals. The best ones are in bold. For DTFD-MIL, the number of pseudo-bags is 5. The flops are measured with the number of instances of a bag being 120, and the instance feature extraction by ResNet-50 is not considered in the presented model sizes and flops.

Method	CAMELYON-16 Acc	CAMELYON-16 F1	CAMELYON-16 AUC	FLOPs	Model Size
Mean Pooling	$0.626_{(0.616,0.636)}$	$0.355_{(0.346,0.363)}$	$0.528_{(0.518,0.538)}$	62.4M	524.3K
Max Pooling	$0.826_{(0.798,0.854)}$	$0.754_{(0.694,0.813)}$	$0.854_{(0.816,0.891)}$	62.4M	524.3K
RNN-MIL [3]	$0.844_{(0.818,0.870)}$	$0.798_{(0.791,0.806)}$	$0.875_{(0.873,0.877)}$	64.0M	1.57M
Classic AB-MIL [14]	$0.845_{(0.839,0.851)}$	$0.780_{(0.769,0.791)}$	$0.854_{(0.848,0.860)}$	78.1M	655.3K
DS-MIL [18]	$0.856_{(0.843,0.869)}$	$0.815_{(0.797,0.832)}$	$0.899_{(0.890,0.908)}$	117.6M	855.7K
CLAM-SB [23]	$0.837_{(0.809,0.865)}$	$0.775_{(0.755,0.795)}$	$0.871_{(0.856,0.885)}$	94.8M	790.7K
CLAM-MB [23]	$0.823_{(0.795,0.85)}$	$0.774_{(0.752,0.795)}$	$0.878_{(0.861,0.894)}$	94.8M	791.1K
Trans-MIL [34]	$0.858_{(0.848,0.868)}$	$0.797_{(0.776,0.818)}$	$0.906_{(0.875,0.937)}$	613.83M	2.66M
DTFD-MIL (MaxS)	$0.864_{(0.848,0.880)}$	$0.814_{(0.802,0.826)}$	$0.907_{(0.894,0.919)}$	79.4M	986.7K
DTFD-MIL (MaxMinS)	$0.899_{(0.887,0.912)}$	$0.865_{(0.848,0.882)}$	$0.941_{(0.936,0.944)}$	80.1M	986.7K
DTFD-MIL (AFS)	$\mathbf{0.908}_{(0.892,0.925)}$	$\mathbf{0.882}_{(0.861,0.903)}$	$\mathbf{0.946}_{(0.941,0.951)}$	79.4M	986.7K
DTFD-MIL (MAS)	$0.897_{(0.890,0.904)}$	$0.864_{(0.855,0.873)}$	$0.945_{(0.943,0.947)}$	79.4M	986.7K

Table 2. Results on TCGA lung cancer. The subscripts are the corresponding standard variances. The best ones are in bold. For DTFD-MIL, the number of pseudo-bags is 8.

	TCGA Lung Cancer Acc	TCGA Lung Cancer F1	TCGA Lung Cancer AUC
Mean Pooling	$0.833_{0.011}$	$0.809_{0.012}$	$0.901_{0.012}$
Max Pooling	$0.846_{0.029}$	$0.833_{0.027}$	$0.901_{0.033}$
RNN-MIL [3]	$0.845_{0.024}$	$0.831_{0.023}$	$0.894_{0.025}$
Classic AB-MIL [14]	$0.869_{0.032}$	$0.866_{0.021}$	$0.941_{0.028}$
DS-MIL [18]	$0.888_{0.013}$	$0.876_{0.011}$	$0.939_{0.019}$
CLAM-SB [23]	$0.875_{0.041}$	$0.864_{0.043}$	$0.944_{0.023}$
CLAM-MB [23]	$0.878_{0.043}$	$0.874_{0.028}$	$0.949_{0.019}$
Trans-MIL [34]	$0.883_{0.022}$	$0.876_{0.021}$	$0.949_{0.013}$
DTFD-MIL (MaxS)	$0.868_{0.040}$	$0.863_{0.029}$	$0.919_{0.037}$
DTFD-MIL (MaxMinS)	$\mathbf{0.894}_{0.033}$	$\mathbf{0.891}_{0.027}$	$\mathbf{0.961}_{0.021}$
DTFD-MIL (AFS)	$0.891_{0.033}$	$0.883_{0.025}$	$0.951_{0.022}$
DTFD-MIL (MAS)	$0.891_{0.029}$	$0.890_{0.021}$	$0.955_{0.023}$

4. Experiments

In this section, we present the performance of the proposed methods in comparison to other latest MIL works on histopathology WSI, and qualitatively validate the soundness of the derivation of instance probability. We also conduct ablation experiments to further study the proposed methods. More experimental results are presented in the supplementary.

4.1. Datasets

We evaluate the proposed methods on two public histopathology WSI datasets: CAMELYON-16 [2] and The Caner Genome Atlas (TCGA) lung cancer. Please refer to Supplementary for the details of these two datasets.

For pre-processing, we apply the OTSU's threshold method to localize the tissue regions in each WSI. Non-overlapping patches of a size of 256×256 pixels on the 20X magnification are then extracted from the tissue regions. There are in total 3.7 millions patches from the CAMELYON-16 dataset and 8.3 millions patches from the TCGA Lung Cancer dataset.

4.2. Implementation Details

The implementations are described in Supplementary material. For more details, please refer to the released code.

4.3. Evaluation Metrics

For all the experiments, area under curve (AUC) is the primary performance metric to report, since it is more comprehensive and less sensitive to class imbalance. Besides, the slide-level accuracy (Acc) and F1 score are also considered, which are determined by the threshold of 0.5.

For CAMELYON-16, the official training set is further randomly split into training and validation sets with a ratio of 9:1. An experiment is running for 5 times, and the mean values of performance metrics on CAMELYON-16 official test set and the corresponding 95% confidence interval (CI-95) are reported. For TCGA lung cancer, we randomly split the dataset into training, validation, and testing sets with a ratio of 65:10:25 on the patient level. 4-folder cross-validation is adopted, and the mean value of performance metrics of the 4 test folders are reported. Since for each test folder the performances vary significantly, the CI-95 from just 4 values is of less use; therefore, we instead report the corresponding standard variances.

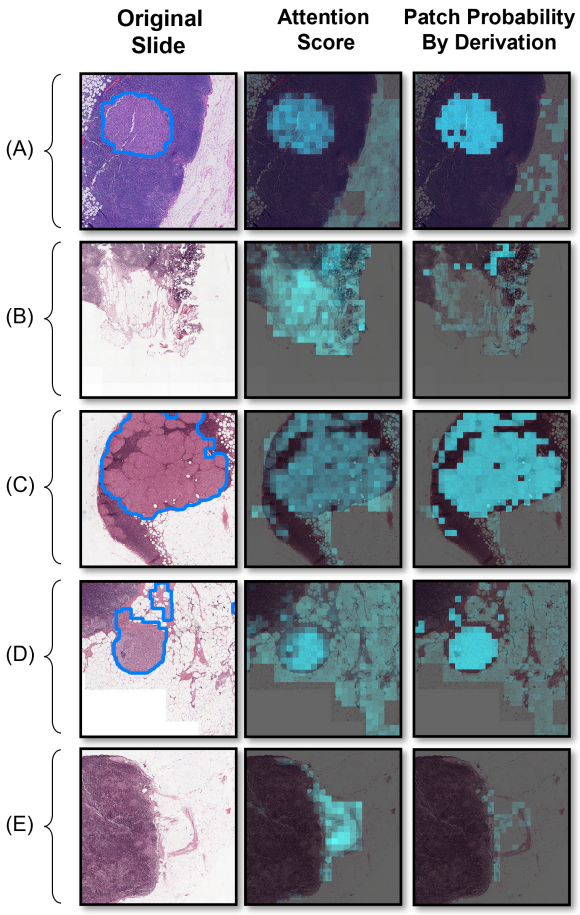

Figure 4. Heat maps of 5 sub-fields of slides by attention score and by the patch probability derivation, respectively. In the column of 'Original Slide', the tumor regions are delineated by the blue lines. In the second and third columns, brighter cyan colors indicate higher probabilities to be tumor for the corresponding locations.

4.4. Performance comparison with existing works

We present the experimental results of the proposed methods on CAMELYON-16 and TCGA lung cancer dataset, in comparison to the following methods: (1) Conventional instance-level MIL, including the Mean-Pooling and Max-Pooling. (2) RNN based RNN-MIL [3]. (3) The classic AB-MIL [14]. (4) Three variants of AB-MIL, including non-local attention pooling DSMIL [18], single-attention-branch CLAM-SB [23] and multi-attention-branch CLAM-MB [23]. (5) transformer-based MIL, Trans-MIL [34]. The results of all the other methods are from the experiments conducted using their official codes under the same settings. As shown in Table.1, the proposed models have similar model sizes and computational complexities with the models of other works except for Trans-MIL, while Trans-MIL is significantly larger in model size and computational complexity.

The results on CAMELYON-16 test set are presented in Tab.1, while the results on TCGA lung cancer are shown in Tab.2. Generally, the instance-level methods (Mean Pooling, Max Pooling) are inferior to the bag embedding-based methods in performances.

For CAMELYON-16, most positive slides contain only small portions of tumor over the whole tissue region. Among the proposed DTFD-MIL methods with different feature distillations, the MaxS has the most inferior performances, yet it still outperforms other existing MIL methods except the most recent Trans-MIL. The other 3 DTFD-MIL achieve similar performances, which are significantly better than others. For example, DTFD-MIL(AFS) is at least 4% better in AUC than other existing methods.

For TCGA lung cancer, except for DTFD-MIL(MaxS), the proposed methods also achieve leading performances, with DTFD-MIL(MaxMinS) obtaining the best AUC value of 96.1%. Due to significantly larger tumor regions in positive slides, however, even the instance-level methods perform well on the TCGA lung cancer dataset, resulting in less obvious superiority of the proposed methods over other existing methods. In comparison, for the much more challenging dataset CAMELYON-16, the proposed methods present stronger robustness to the situation of small portions of tumor regions in positive slides.

4.5. Visualization of Detection Results

To further explore the proposed instance probability derivation, we train a classic AB-MIL model, and generate the heatmaps of 5 sub-fields of 5 slides from CAMELYON-16. These heatmaps come from (1) normalized attentions scores (attention-based); (2) patch probability derivation (derivation-based) by Eq.(8) and Eq.(9), respectively. The attention scores directly from the attention module are normalized as $a_k^{'} = (a_k - a_{\min}) / (a_{\max} - a_{\min})$ [14,18,23,34], where $a_{\min}$ and $a_{\max}$ are the minimum and maximum attention scores of patches in a slide, respectively. For better presentation, we remove the estimated probabilities of patches in the derivation-based heatmaps (the third row) whose values are around 0.5 thus contain little information.

The heatmaps from Fig.4 demonstrate the better ability of the instance probability derivation to localize the positive activations, compared with the attention scores. Specifically, the positive activations in the heatmaps by instance probability derivation are more consistent and accurate, and present better contrast compared to that of attention scores. Moreover, in the ground-truth negative slides, there are always strong false positive regions in the heatmaps of attention scores, while in the heatmaps from instance probability derivation, most of these regions can be correctly recognized as negative. In the supplementary material, we provide more in-depth analysis about why the instance proba-

bility derivation is more efficient for positive activation detection compared to the attention scores.

4.6. Ablation Study

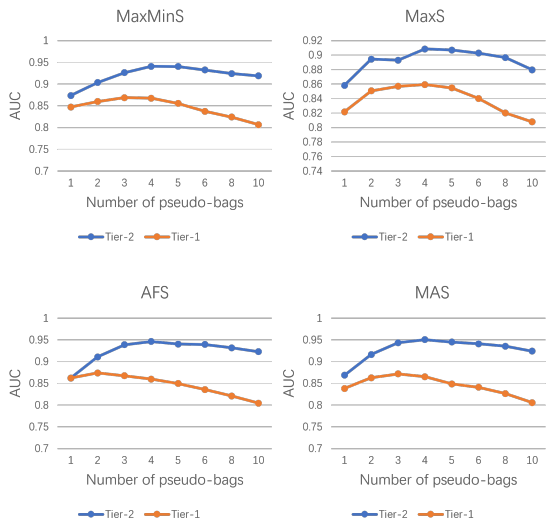

Figure 5. AUC scores of the four feature distillation strategies on CAMELYON-16 test set.

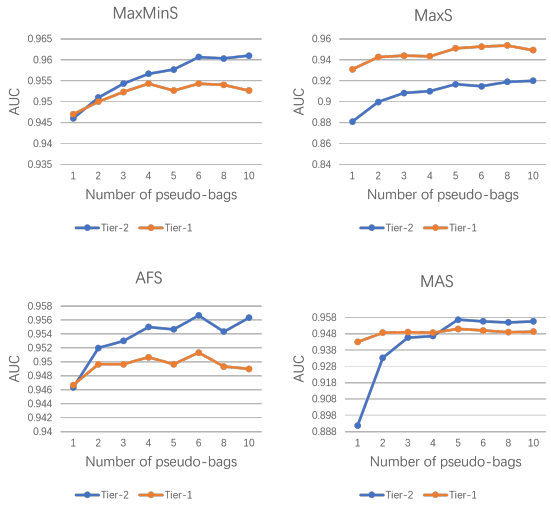

Figure 6. AUC scores of the four feature distillation strategies on TCGA lung cancer dataset.

Fig.5 and Fig.6 present the AUC scores of the proposed methods with respect to different numbers of pseudo-bags on CAMELYON-16 and TCGA lung cancer datasets, respectively. In each sub-figure, the blue curve represents the Tier-2 MIL model, while the red curve represents the Tier-1 MIL model which directly works on pseudo-bags.

From these curves, we can summarize:

(1). The pseudo-bag idea is beneficial to both the Tier-1 and Tier-2 MIL models. However, the Tier-1 models are more sensitive to the number of pseudo-bags in CAMELYON-16: the corresponding AUC scores drop dramatically as the number of pseudo-bags increases from 3. In contrast, Tier-1 models are less sensitive to the number of pseudo-bags in TCGA lung cancer dataset, and they even achieve high-level performances with a proper number of pseudo-bags. This phenomenon mainly results from that the tumors are usually minor regions in CAMELYON-16 positive slides while in TCGA lung cancer the situation reverses; therefore, it is highly possible that a pseudo-bag may not be allocated with at least one positive instance from a positive parent bag in CAMELYON-16. This well justifies our initial motivation to build up a second-tier MIL model upon the distilled features from the corresponding pseudo bags, and in general the performances of Tier-2 models indeed are better than those of Tier-1 models, especially in CAMELYON-16.

(2). Among the four feature distillation strategies, the DTFD-MIL (MaxS) performance is not comparable to the other three, and on TCGA lung cancer dataset the Tier-2 MIL model is even inferior to the Tier-1 MIL model when MaxS feature distillation is used. It implies that adopting the instances with the highest positive responses to form the representation of the bag is not always the optimal option. This phenomenon is also in accordance with the observation from Fig.4, where the strongest activations in a negative slide are from the neutral or even blank regions (corresponding to approximately zero probability of being tumor), instead of the non-tumor tissue regions.

5. Conclusion

The first contribution of this paper is the derivation of instance probability under the framework of AB-MIL, and we qualitatively demonstrate the derived instance probability is a more reliable metric over the widely-used attention scores for the positive region detection. We then propose the DTFD-MIL, which utilizes the idea of pseudo bags and double-tier MIL. The derivation of instance probability serves for the feature distillation in DTFD-MIL. The experimental results demonstrate that the proposed DTFD-MIL indeed provides a new perspective to solve the MIL problem with superior performances, rather than utilizing mutual-instance relations as in other latest works. Finally, we also expect that the derivation of instance probability will be served as a useful tool for developing related MIL models or for the related analysis in future works, just as the role it plays in the proposed DTFD-MIL in this paper.

Acknowledgement H. Zhang and Y. Meng thank the China Science IntelliCloud Technology Co., Ltd for the studentships. The TCGA Lung cancer dataset is from the TCGA Research Network: https://www.cancer.gov/tcga.

References

[1] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013. 1

[2] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. 6

[3] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. 2, 3, 6, 7

[4] Lyndon Chan, Mahdi S Hosseini, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10662–10671, 2019. 3

[5] Zenghai Chen, Zheru Chi, Hong Fu, and Dagan Feng. Multi-instance multi-label image classification: A neural approach. *Neurocomputing*, 99:298–306, 2013. 1

[6] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528. Springer, 2020. 2

[7] Toby C Cornish, Ryan E Swapp, and Keith J Kaplan. Whole-slide imaging: routine pathologic diagnosis. *Advances in Anatomic Pathology*, 19(3):152–159, 2012. 1

[8] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997. 1

[9] Ji Feng and Zhi-Hua Zhou. Deep MIML network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 3

[10] Lucy Godson, Navid Alemi, Jeremie Nsengimana, Graham Cook, Emily L Clarke, Darren Treanor, D Timothy Bishop, Julia A Newton-Bishop, and Ali Gooya. Weakly-supervised learning for image-based classification of primary melanomas into genomic immune subgroups. In *Medical Imaging with Deep Learning*, 2022. 2

[11] Lei He, L Rodney Long, Sameer Antani, and George R Thoma. Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine*, 107(3):538–556, 2012. 1

[12] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2424–2433, 2016. 2, 3

[13] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 3

[14] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pages 2127–2136. PMLR, 2018. 2, 3, 4, 6, 7

[15] Fahdi Kanavati, Gouji Toyokawa, Seiya Momosaki, Michael Rambeau, Yuka Kozuma, Fumihiro Shoji, Koji Yamazaki, Sadanori Takeo, Osamu Iizuka, and Masayuki Tsuneki. Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific Reports*, 10(1):1–11, 2020. 3

[16] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016. 3

[17] Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. Weakly supervised multiple instance learning histopathological tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 470–479. Springer, 2020. 3

[18] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021. 2, 3, 4, 6, 7

[19] Shaohua Li, Yong Liu, Xiuchao Sui, Cheng Chen, Gabriel Tjio, Daniel Shu Wei Ting, and Rick Siow Mong Goh. Multi-instance multi-scale cnn for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–539. Springer, 2019. 1

[20] Yi Li and Wei Ping. Cancer metastasis detection with neural conditional random field. *arXiv preprint arXiv:1806.07064*, 2018. 1

[21] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, 6(1):1–11, 2016. 1

[22] Ming Y Lu, Richard J Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825*, 2019. 1

[23] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. 3, 4, 6, 7

[24] Anant Madabhushi. Digital pathology image analysis: opportunities and challenges. *Imaging in Medicine*, 1(1):7, 2009. 1

[25] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in Neural Information Processing Systems*, pages 570–576, 1998. 1

[26] Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. Spatial uncertainty-aware semi-supervised crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15549–15559, October 2021. 1

[27] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015. 1

[28] Maxime Oquab, Léon Bottou, Ivan Laptev, Josef Sivic, et al. Weakly supervised object recognition with convolutional neural networks. In *Proc. of NIPS*, pages 1545–5963. Citeseer, 2014. 1

[29] Liron Pantanowitz, Paul N Valenstein, Andrew J Evans, Keith J Kaplan, John D Pfeifer, David C Wilbur, Laura C Collins, and Terence J Colgan. Review of the current state of whole slide imaging in pathology. *Journal of Pathology Informatics*, 2, 2011. 1

[30] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014. 1

[31] Hans Pinckaers, Bram van Ginneken, and Geert Litjens. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[32] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015. 1

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 2, 3

[34] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 6, 7

[35] Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A Moskaluk, Sana Syed, and Donald Brown. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, pages 682–698. PMLR, 2021. 1, 2, 3

[36] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[37] Ming Tu, Jing Huang, Xiaodong He, and Bowen Zhou. Multiple instance learning with graph neural networks. *arXiv preprint arXiv:1906.04881*, 2019. 2

[38] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1306, 2018. 3

[39] Chaofei Wang, Jiayu Xiao, Yizeng Han, Qisen Yang, Shiji Song, and Gao Huang. Towards learning spatially discriminative feature representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1326–1335, 2021. 3

[40] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016. 1

[41] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Efstratios Tsougenis, Qitao Huang, Muyan Cai, and Pheng-Ann Heng. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Transactions on Cybernetics*, 50(9):3950–3962, 2019. 1

[42] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018. 3, 4

[43] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1362, 2018. 3

[44] Yinxi Wang, Kimmo Kartasalo, Philippe Weitz, Balazs Acs, Masi Valkonen, Christer Larsson, Pekka Ruusuvuori, Johan Hartman, and Mattias Rantalainen. Predicting molecular phenotypes from histopathology images: A transcriptome-wide expression–morphology analysis in breast cancer. *Cancer Research*, 81(19):5115–5126, 2021. 1

[45] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 434–450, 2018. 3

[46] Chensu Xie, Hassan Muhammad, Chad M Vanderbilt, Raul Caso, Dig Vijay Kumar Yarlagadda, Gabriele Campanella, and Thomas J Fuchs. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning. In *Medical Imaging with Deep Learning*, pages 843–856. PMLR, 2020. 2

[47] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10682–10691, 2019. 3, 5

[48] Hongrun Zhang, Helen Kalirai, Amelia Acha-Sagredo, Xiaoyun Yang, Yalin Zheng, and Sarah E Coupland. Piloting a deep learning model for predicting nuclear bap1 immunohistochemical expression of uveal melanoma from hematoxylin-and-eosin sections. *Translational Vision Science & Technology*, 9(2):50–50, 2020. 1

[49] Hongrun Zhang, Yanda Meng, Xuesheng Qian, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. A regularization term for slide correlation reduction in whole slide image analysis with deep learning. In *Medical Imaging with Deep Learning*, pages 842–854. PMLR, 2021. 1

[50] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4262–4270, 2018. 3

[51] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, et al. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4837–4846, 2020. 2

[52] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 3

[53] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–611. Springer, 2017. 3