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

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah Coupland, Yalin Zheng

1. Dataset Description

CAMELYON-16 [1] is a WSI dataset that has been released for the detection of breast cancer metastasis. This dataset contains 400 WSIs in total, including 270 for training and 130 for testing (officially splitting)[1]. There are 159 normal and 111 tumor slides in the training set. Although CAMELYON-16 has both pixel-level annotation and slide labels, for this specific MIL application, we only use the slide labels for training and testing, except for the average FROC calculations. The challenge of this dataset is that most positive slides contain only small portions of tumors over the whole tissue regions.

TCGA Lung Cancer Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC) are two sub-type of cancers in the TCGA lung cancer dataset, with 534 LUAD and 512 LUSC slides, respectively. There are only slide-level labels available for this dataset. Compared to CAMELYON-16, tumor regions in tumor slides are significantly larger in this dataset.

2. Implementation Details

Following [7, 8], we use the ResNet-50 model [3] pretrained with ImageNet [2] as the backbone network, to extract an initial feature vector from each patch, which has a dimension of 1024. The last convolutional module of the ResNet-50 is removed and a global average pooling is applied to the final feature maps to generate the initial feature vector. The initial feature vector is then reduced to a 512-dimensional feature vector by one fully-connected layer, which is served as the ultimate feature representation of a patch. An Adam optimizer [5] with weight decay of 0.0001 is used for the model training. All the models are trained for 200 epochs with an initial learning rate of 0.0001, which is reduced to 20% of itself after 100 epochs. The batch size is set to be 1, meaning that in each iteration, one slide is processed. All the experiments were conducted with a NVIDIA V100 GPU.

It should be noted that the fixed initial features obtained by pretraining on ImageNet are not the only option. Alternatively, we can extract the features of patches through self-supervised learning (SSL) [6].

[1]Two slides in the test set are officially recognized as being incorrectly annotated thus are excluded in the experiments.

3. Proof of Proposition 1

Observing Eq.(6) (main paper), the bag feature representation (embedding) in AB-MIL can be seen as the averaging pooling result of the re-scaled instance features, i.e.,

$$\boldsymbol{F} = \sum_{k=1}^{K} a_k \boldsymbol{h}_k = \frac{1}{K} \sum_{k=1}^{K} a_k K \boldsymbol{h}_k = \frac{1}{K} \sum_{i=1}^{K} \hat{\boldsymbol{h}}_k, \qquad (1)$$

with the re-scaled instance feature $\hat{\boldsymbol{h}}_k = a_k K \boldsymbol{h}_k$. Meanwhile, an image's feature representation obtained by the global average pooling operation in a deep learning classification model as in Eq.(1) (main paper) can also be rephrased as,

$$\boldsymbol{f} = \frac{1}{WH} \sum_{w,h}^{W,H} \boldsymbol{u}_{w,h} \qquad (2)$$

where $\boldsymbol{u}_{w,h} \in \mathbb{R}^D$ is the feature vector from $\boldsymbol{U}$ located at w, h. Comparing Eq.(1) and Eq.(2), it is obvious that the bag feature representation $\boldsymbol{F}$ of AB-MIL essentially has the same formation with the image's feature representation $\boldsymbol{f}$ of a deep learning model for image classification. Note that in both AB-MIL and deep learning based classification model, a classifier is operated upon either $\boldsymbol{F}$ or $\boldsymbol{f}$ for the bag or image classification. One can subsequently conclude that the AB-MIL has the same framework structure with that of the deep-learning framework for image classification as described in Section.3.1.1 (main manuscript), except for the only difference that the feature vectors $\boldsymbol{u}_{w,h}$ from the feature maps $\boldsymbol{U}$ have spatial relations with each other, while the spatial relations between feature vector $\hat{\boldsymbol{h}}_k$ are not explicitly considered in AB-MIL (or the spatial relations may only be encoded in the attention scores.). However, the spatial relations play no role in the inference of attention map in Grad-CAM. Therefore it is safe to apply the mechanism of Grad-CAM to AB-MIL to directly infer the signal strength for an instance to be positive or negative. Resembling to

Eq.(2) in the main paper, the signal strength for instance k to be class c ($c = 0$ for negative and $c = 1$ for positive) can then be derived as,

$$L_k^c = \sum_d^D \beta_d^c \hat{h}_{k,d}, \quad \beta_d^c = \frac{1}{K} \sum_{k=1}^K \frac{\partial s_c}{\partial \hat{h}_{k,d}} \tag{3}$$

where $\hat{h}_{k,d}$ is the d_{th} element of $\hat{\boldsymbol{h}}_k$, and s_c is the output logit for class c from the MIL classifier. By applying soft-max, the corresponding probability is then,

$$p_k^c = \frac{\exp\left(L_k^c\right)}{\sum_{t=1}^C \exp\left(L_k^t\right)} \tag{4}$$

The value of the attention score in $\hat{\boldsymbol{h}}_k$, however, may affect the availability of the probability derivation by Eq.(4). When a certain patch (instance) is deactivated by the attention module, i.e., $a_k \to 0$, the corresponding $\hat{\boldsymbol{h}}_k$ (Eq.(3)) tends to be a zero vector. In this case, L_k^c will be closed to zero for all the classes, resulting in p_k^c (Eq.(4)) being close to 0.5 (for both $c = 0$ and $c = 1$), which means the derived probabilities reveal little information. In short, the proposed derivation of instance probability will be only applicable to the instances assigned with large enough attention scores. However, those patches assigned with low values of attention scores are deemed by the trained model as unimportant ones for the bag-level prediction; therefore, their probabilities are not crucial.

4. More on instance probability derivation

Free-Response ROC (FROC) is usually used to evaluate the detection ability of a model. A higher value of FROC indicates a better detection capability. In Tab.1, we report the average FROC values of different methods on the CAMELYON-16 test set. A FROC value is defined as the average detection sensitivity at 6 predefined numbers of false-positive per slide: 1/4, 1/2, 1, 2, 4 and 8. The average FROC values are measured based on the corresponding probability maps that are generated from attention scores, derived instance probabilities or direct instance probability outputs, for various methods, including AB-MIL [4], DS-MIL [6] and Max Pooling, as shown in the table. Of these, the probability map of Max Pooling is formed by the direct instance probability outputs of a trained MIL model of Max Pooling, since it is an instance-level model. Therefore, the calculated average FROC of MaxPooling is an metric explicitly related to the model's detection capability, and its value can serve as a benchmark for comparison. A threshold of 0.5 is applied to all the probability maps for calculating the average FROC values.

Overall, the results from Tab.1 suggest that the probability maps from the derived instance probabilities can better reveal the detected locations of positive activation, compared to the probability maps from attention scores. Specifically, we can see from Tab.1 that when using the probability map from the derivation, the average FROC score of AB-MIL is similar to that of MaxPooling, of these two the corresponding slide-level AUC scores also have similar values. In contrast, the average FROC scores of AB-MIL and DS-MIL by the probability maps from attention scores achieve much lower values, which are not in accordance with their slide-level AUC scores. For instance, DS-MIL can achieve a slide-level AUC of 0.899 (best among the three methods), but the average FROC by the probability maps from attention scores is merely 0.262.

Fig.2 presents the color heatmaps of 10 sub-fields from 10 slides by a trained classic AB-MIL model. The heat maps are from normalized attention scores (attention-based) and the proposed instance probability derivation (derivation-based), respectively. The attention scores directly from the attention module are normalized as [4, 6–8],

$$a_k^{'} = \left(a_k - a_{\text{min}}\right) / \left(a_{\text{max}} - a_{\text{min}}\right), \tag{5}$$

where a_{min} and a_{max} are the minimum and maximum attention scores for patches in a slide, respectively.

We can see that the derivation-based heatmaps provide better contrast for the predicted tumor regions over the non-tumor regions, and present a higher level of consistency and accuracy. On the one hand, there are always strong activations in the attention-based heatmaps, no matter whether there exist ground-truth tumor regions or not, which can provide misleading information when these heatmaps are used for offline analysis. This deficiency mainly results from the normalization of the attention scores. An attention-based heatmap utilizes the comparisons of attention scores of patches in a slide for rescaling (Eq.(5)). As a result, the strongest activations in an attention-based heatmap are not always those with the highest positive probabilities, but those with larger attention scores over other patches. On the other hand, a probability value in a derivation-based heatmap comes from the comparison of derived logits of different classes, i.e., the soft-max operation in Eq.(9) in the main paper. Therefore, a heat-map of derived probabilities is explicitly relevant to the class estimations by the trained MIL model.

We also notice that the sparse tissue regions surrounding the major tissues tend to receive comparably high attention scores, although usually they can be easily recognized as non-tumor regions from the derivation-based heatmaps. Probably this phenomenon results from the fact that the model benefits from choosing these neutral regions to represent the negative slides instead of the non-tumor tissue regions; therefore the attention module automatically learns to activate these regions. This phenomenon is also in accordance with the performance differences of DTFD-MIL(MaxS) and DTFD-MIL(MaxMinS) presented in Tab.1

Table 1. Average FROC and slide-level AUC on the CAMELYON-16 test set.

Method	AUC	Probability Map From	FROC
Classic AB-MIL [4]	0.854	attention score	0.251
		probability by derivation	0.375
DS-MIL [6]	0.899	attention score	0.262
MaxPooling	0.854	direct instance output	0.387

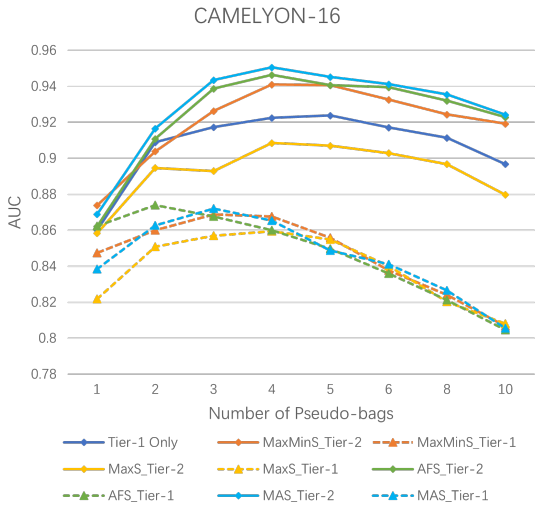
in the main paper, i.e., the distillation of the instance with maximum derived probability (MaxS) in a pseudo-bag to Tier-2 model shows inferior performance to the distillation of the two instances with minimum and maximum derived probability (MaxMinS) in a pseudo-bag, where the minimum ones correspond to the sparse tissue regions mentioned above.

Please note that the average FROC values and the heatmaps presented here are to demonstrate the effectiveness of the instance probability derivation for positive detection in AB-MIL models. The extension of the instance probability derivation to the proposed double-tier framework may not be straightforward, and it is remained to be solved in future work.

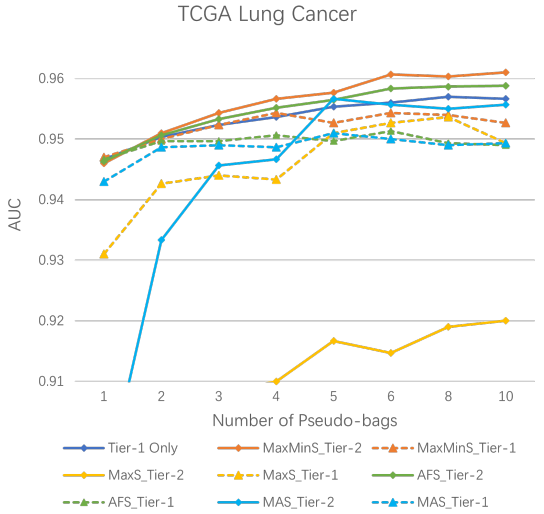
5. With Tier-1 Model only

In the proposed DTFD-MIL framework, a pseudo-bag acts as an independent unit, and plays the role of a regular bag for MIL in Tier-1. The Tier-1 MIL model is trained and tested directly on pseudo-bags, while the Tier-2 model takes responsibility for the parent bags' inference. One may, however, wonder about the performances of only the Tier-1 MIL model if it is trained on pseudo-bags but is tested on the original bags. We conduct ablation experiments to further explore this case, in which only the Tier-1 MIL model is kept after the Tier-2 MIL model is removed from the framework, and the remained Tier-1 model is trained on the pseudo-bags while tested on the original bags. This case is denoted as 'Tier-1 only' in Fig.1. For CAMELYON-16 test set (Fig.1a), the 'Tier-1 only' case performs better than the Tier-1 MIL models in the original DTFD-MIL framework, which are trained and tested on pseudo-bags. Obviously, the 'Tier-1 only' model as well benefits from the increasing number of pseudo-bags when the number is smaller than a certain value. However, the Tier-2 MIL models in the original DTFD-MIL framework still outperform the 'Tier-1 only', which further demonstrates the advantage of using the Tier-2 MIL models. The performances on the TCGA lung cancer dataset presented in Fig.1b also show a similar trend to that on CAMELYON-16, although the performance gaps between different methods are smaller.

These experimental results further demonstrate the effectiveness of the proposed ideas of pseudo-bag and double-tier MIL.



(a)



(b)

Figure 1. AUC scores of the proposed DTFD-MIL with different feature distillation approaches and of the 'Tier-1 only' case on CAMELYON-16 and TCGA Lung Cancer, respectively. Note that some curves in sub-figure (b) only present the upper parts for better comparison.

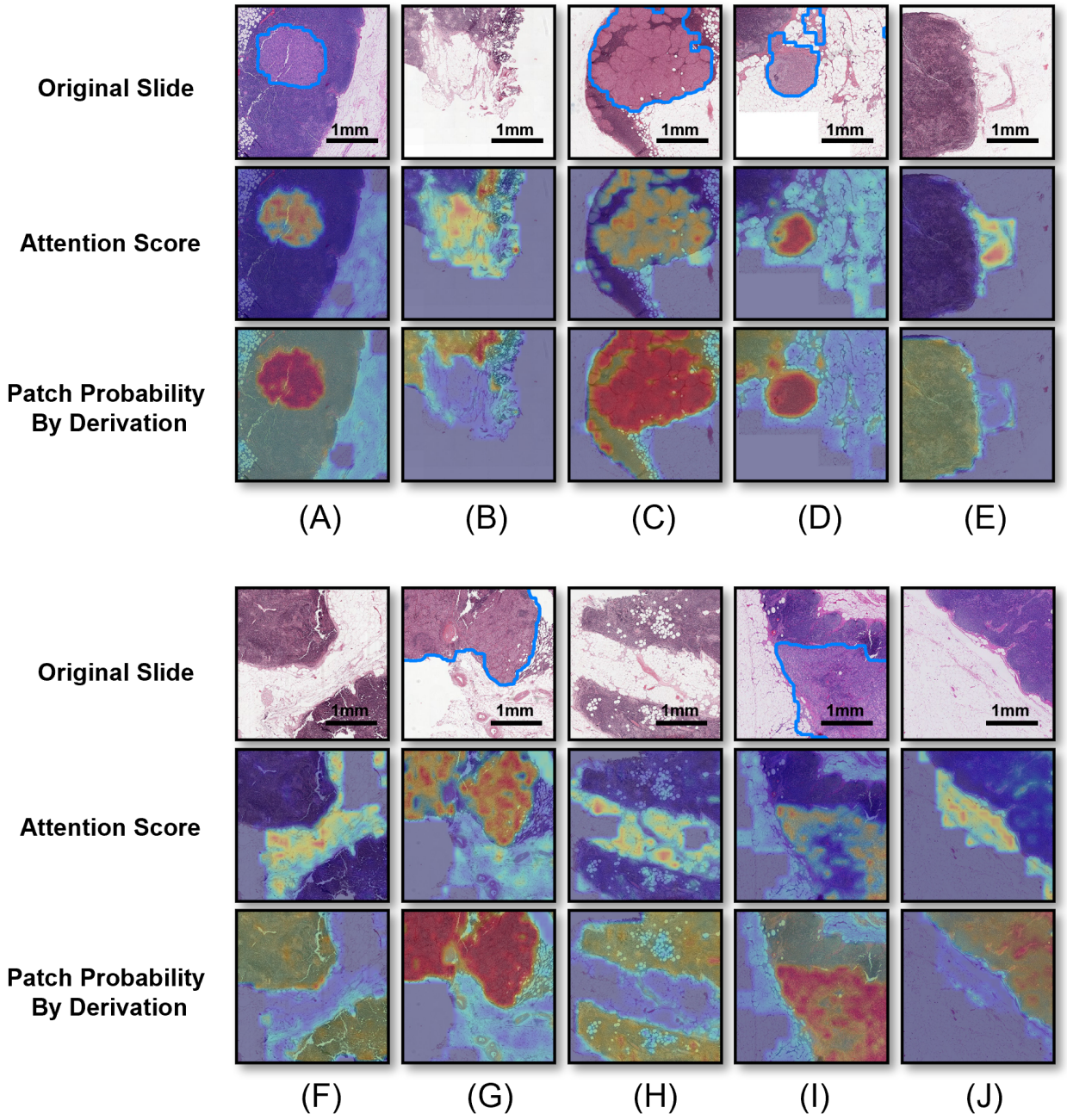


Figure 2. Color heatmaps of 10 sub-fields of 10 slides by attention score and the proposed instance probability derivation, respectively. In the row of 'Original Slide', the tumor regions are delineated by blue lines.

References

[1] 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. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition, pages 770–778, 2016. 1

[4] 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

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] 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. 1, 2, 3

[7] 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. 1, 2

[8] 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. 1, 2