Diagnose Like a Pathologist: Transformer-Enabled Hierarchical Attention-Guided Multiple Instance Learning for Whole Slide Image Classification

Conghao Xiong[1] , **Hao Chen**[2] , **Joseph J.Y. Sung**[3] and **Irwin King**[1]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]Department of Computer Science and Engineering and Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology
[3]Lee Kong Chian School of Medicine, Nanyang Technological University
{chxiong21, king}@cse.cuhk.edu.hk, jhc@cse.ust.hk, josephsung@ntu.edu.sg

Abstract

Multiple Instance Learning (MIL) and transformers are increasingly popular in histopathology Whole Slide Image (WSI) classification. However, unlike human pathologists who selectively observe specific regions of histopathology tissues under different magnifications, most methods do not incorporate multiple resolutions of the WSIs, hierarchically and attentively, thereby leading to a loss of focus on the WSIs and information from other resolutions. To resolve this issue, we propose a Hierarchical Attention-Guided Multiple Instance Learning framework to fully exploit the WSIs. This framework can dynamically and attentively discover the discriminative regions across multiple resolutions of the WSIs. Within this framework, an Integrated Attention Transformer is proposed to further enhance the performance of the transformer and obtain a more holistic WSI (bag) representation. This transformer consists of multiple Integrated Attention Modules, which is the combination of a transformer layer and an aggregation module that produces a bag representation based on every instance representation in that bag. The experimental results show that our method achieved state-of-the-art performances on multiple datasets, including Camelyon16, TCGA-RCC, TCGA-NSCLC, and an in-house IMGC dataset. The code is available at https://github.com/BearCleverProud/HAG-MIL.

1 Introduction

Histopathology tissue analysis is widely recognised as the gold standard for cancer diagnosis, making histopathology **W**hole **S**lide **I**mage (WSI) analysis an important deep learning application in the real world. Due to the advancements of scanning and storage devices, the number of digitalised histology tissues, *i.e.*, WSIs, is increasing. This makes deep learning methods applicable to WSI classification tasks [Dimitriou *et al.*, 2019; Srinidhi *et al.*, 2021; Li *et al.*, 2022], and there are already some successful applications on breast [Huang *et al.*, 2006; Peng and King, 2008], prostate [Campanella *et al.*, 2019], skin [Ianni *et al.*, 2020],
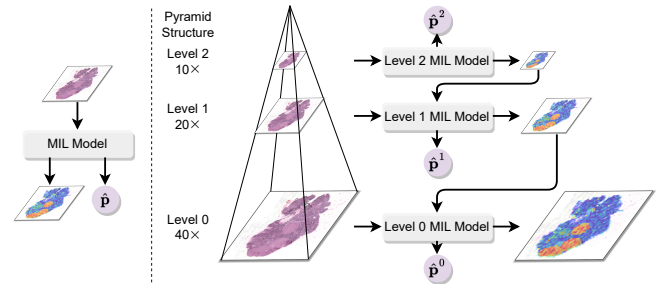

Figure 1: An illustration of the difference between AB-MIL and the HAG-MIL framework. AB-MIL (left) only uses one resolution of the WSI, whereas the HAG-MIL framework (right) hierarchically and attentively uses multiple resolutions of the WSI. The high-attention regions are passed to the models at higher resolutions for further checking.

and pancreas cancer [Keikhosravi *et al.*, 2020].

Though deep learning is a common solution to computer vision [He *et al.*, 2016; Liu *et al.*, 2021], WSIs, as gigapixel images, pose a unique challenge for deep learning since they are too large to be fed into a neural network. The **M**ultiple **I**nstance **L**earning (MIL) [Maron and Lozano-Pérez, 1997] framework provides a solution to this problem. Under this framework, the WSI is regarded as a bag of patches (instances) that are images cropped from the WSI, which can be directly fed into a neural network [Ilse *et al.*, 2018].

Most MIL methods assume that the patches are independent and identically distributed [Ilse *et al.*, 2018; Campanella *et al.*, 2019]. However, patches have strong correlations with each other, *e.g.*, a tumour patch might have similar morphological structures to another tumour patch in the same WSI. Neglecting these correlations may result in insufficient exploitation of the WSI. The transformer [Vaswani *et al.*, 2017] is well-suited for MIL with patch correlations, as it can capture the dependencies between patches [Shao *et al.*, 2021].

The lack of fine-grained annotations is another obstacle to achieving good performance with deep learning models. Due to the large size of the WSI, it can take several hours for an expert pathologist to annotate each pixel, which is prohibitively time-consuming and expensive; therefore, the WSI datasets usually include only slide-level labels. Furthermore, the malignant tissues may only occupy less than 10%

of the entire WSI [Litjens *et al.*, 2018], making most of the regions irrelevant to the bag label "tumour". It is therefore important to find the discriminative patches from WSIs since they are the only patches providing strong indications of the bag label. These factors also contribute to the thriving of the weakly-supervised learning on WSI classification tasks [Lu *et al.*, 2019; Chikontwe *et al.*, 2020; Zhao *et al.*, 2020; Lu *et al.*, 2021; Tourniaire *et al.*, 2021; Li *et al.*, 2021a; Li *et al.*, 2021b], as they do not need pixel-level annotations.

Identifying the discriminative patches in a large dataset can be a challenging task, akin to finding a needle in a haystack. The WSI datasets often have less than a thousand of WSIs whereas hundreds of thousands of patches can be cropped from each WSI. These discriminative patches are overwhelmed by a flood of irrelevant ones, preventing the model from learning the most essential features of the patches. In fact, many works address this issue through techniques like contrastive learning [Li *et al.*, 2021a; Wang *et al.*, 2021; Tu *et al.*, 2022], attention mechanism [Ilse *et al.*, 2018; Tourniaire *et al.*, 2021; Lu *et al.*, 2021; Li *et al.*, 2021b], hard mining [Li *et al.*, 2019], and so on [Wang *et al.*, 2020].

We address the problem of finding the most discriminative patches from the following two aspects.

Multi-resolution Previous methods for WSI classification often attach little importance to utilisation of multiple resolution explicitly [Chen *et al.*, 2022]. For instance, a patch at $40\times$ magnification (level 0) may contain fine-grained cellular features, such as stroma, tumour cells, lymphocytes and other texture and morphology features. However, a patch at $10\times$ magnification (level 2) may have the tissue microenvironment, including the organisations of the cell structures and interactions between clusters of cells and so on. Using only one resolution leads to a lack of information and multiple resolutions should be exploited. Unlike current methods, human pathologists intentionally inspect histopathology tissues under multiple magnifications, switching their focus on different regions and magnifications of the WSIs based on their prior knowledge and observations. They diagnose those in a hierarchical and attentive way, which is not the case with current model behaviours. Therefore, to incorporate multiple resolutions of the WSIs and from being inspired by the human pathologists, we propose a **H**ierarchical **A**ttention-**G**uided **M**ultiple **I**nstance **L**earning (HAG-MIL) framework to minimise the disparity between model and human behaviours.

Figure 1 illustrates how human pathologists make diagnoses and how our framework is designed to resemble their approach. Information from lower resolutions can guide the pathologists to regions of interest at higher resolutions for further refinement of their judgment. Similarly, the HAG-MIL framework first inspects patches from lower resolutions, discovers the suspicious regions and passes them to another model at higher resolutions for further confirmation. The process is repeated until the highest resolution is reached. By reserving only a certain number of patches at each resolution, the number of patches in a bag is maintained below a specified bag size. This enables the models to operate at the highest resolution of the WSIs. Also, by eliminating irrelevant patches, the models can concentrate more on the most discriminative ones, thereby improving their performance.

Structure Flaws in the Transformer Structure Although transformers are popular models in all fields, they only output one bag representation at the last layer using the class token [Devlin *et al.*, 2019; Shao *et al.*, 2021; Chen *et al.*, 2022]. This structure has two shortcomings: (1) the bag representation should be obtained based on all instance representations, not just the class token and (2) the bag representation can evolve through different transformer layers and relying solely on the bag representation from the last layer may result in a loss of information from previous bag representations. Hence, we propose the **I**ntegrated **A**ttention **T**ransformer (IAT), consisting of multiple stacked **I**ntegrated **A**ttention **M**odules (IAMs). This module uses an aggregation module to aggregate all instance representations into a bag representation, instead of relying on the class token. Multiple bag representations from different modules are fused into a final bag representation, which allows the transformer model to take into account the evolution of bag representations and produce a more effective and holistic bag representation.

In summary, the benefits of our method include: (1) the alignment of our method with the visual assessment of human pathologists by its effective utilisation of multiple resolutions of the WSIs, (2) preventing discriminative patches from being overwhelmed by irrelevant ones by discarding certain amount of irrelevant patches from the WSI bags, (3) enabling the models performing at the highest resolutions, and (4) generating a more holistic and effective bag representation for the WSI by taking into account all instance representations in that bag and the evolution of the bag representation through different layers. Our contributions are summarised below:

1. We present the HAG-MIL framework to dynamically, attentively and hierarchically locate the most discriminative patches across multiple resolutions of the WSIs like a human pathologist.

2. We propose the IAT to produce a more effective and holistic bag representation by generating a bag representation in each IAM based on every instance representation, and then fusing them together.

3. We conduct extensive experiments on various datasets including Camelyon16, TCGA-RCC, TCGA-NSCLC, and our in-house IMGC dataset. The results demonstrate that our method consistently outperforms other methods.

2 Related Work

2.1 MIL in WSI Analysis

MIL can be divided into two approaches: (1) the *instance-level* and (2) the *embedding-level* approaches. In the first approach, a classifier is trained at the instance level using bag labels, and the predicted instance labels are then aggregated to form the bag label prediction [Hou *et al.*, 2016; Campanella *et al.*, 2019; Kanavati *et al.*, 2020; Chikontwe *et al.*, 2020]. In the second approach, a bag representation is first obtained from instance representations, and a classifier is trained on the bag representations [Ilse *et al.*, 2018; Hashimoto *et al.*, 2020; Li *et al.*, 2020; Tourniaire *et al.*, 2021; Lu *et al.*, 2021; Li *et al.*, 2021a; Shao *et al.*, 2021; Zhang *et al.*, 2022]. The second approach is shown to be superior to the

first one, as the instance labels tend to be noisy [Wang *et al.*, 2018]; therefore, in our work, we adopt the second approach.

Despite its effectiveness, MIL has an apparent disadvantage that it treats every patch in a bag as equally important. Patches should be assigned different importance scores so that the discriminative patches have high importance scores. **A**ttention-**B**ased **MIL** (AB-MIL) [Ilse *et al.*, 2018] calculates the attention scores from instance representations and the attention scores reflect the importance of the corresponding patches. In this way, the model obtains the bag representation by taking a weighed average of the instance representations. Many recent works adapt the idea from AB-MIL [Li *et al.*, 2021a; Lu *et al.*, 2021; Zhang *et al.*, 2022].

2.2 Transformers in WSI Analysis

Transformers [Vaswani *et al.*, 2017; Liu *et al.*, 2021] have demonstrated their efficacy due to their ability to capture long-range dependencies. There are also many works on WSI classification tasks that utilise transformers to model correlations between patches [Shao *et al.*, 2021; Huang *et al.*, 2021; Myronenko *et al.*, 2021; Chen *et al.*, 2021; Zheng *et al.*, 2022; Chen *et al.*, 2022]. In our work, we further improve the quality of the bag representation using an aggregation module, which produces a bag representation based on all instance representations, and bag representations from different IAMs are then fused into a holistic and effective bag representation.

2.3 Utilising Multiple Resolutions of the WSIs

Pathologists make diagnoses with multiple magnifications of histopathology tissues, and it is useful to utilise multiple resolutions of the WSIs [Bejnordi *et al.*, 2015; Tokunaga *et al.*, 2019; Hashimoto *et al.*, 2020; Li *et al.*, 2021a; Zhou and Lu, 2022]. However, the previous works only used multiple resolutions in the same bag [Hashimoto *et al.*, 2020] or concatenate the features from different resolutions together [Li *et al.*, 2021a]. Moreover, the two-stage training can be reduced to an end-to-end process [Zhou and Lu, 2022]. Different from other works, we resemble the pathologists and make use of multiple resolutions in a hierarchical and attentive way.

3 Methodology

3.1 Problem Formulation

Given the features extracted by a pre-trained neural network $\boldsymbol{F} \in \mathbb{R}^{n \times d}$, where n is the number of patches in the WSI and d is the hidden dimension of the pre-trained neural network, the MIL model predicts the bag label of that WSI slide $\hat{Y}$ with only the ground truth bag label Y supervision available. The bag label Y would be positive if and only if there is at least one positive instance in the bag, and it is defined as Eq. (1):

$$Y = \begin{cases} 1 & \exists i, y_i = 1, \\ 0 & \forall i, y_i = 0, \end{cases} \tag{1}$$

where y_i is the ground truth label of the i-th patch. One issue with this formulation is that it cannot identify the discriminative instances, since each instance has the same weight under this framework. AB-MIL [Ilse *et al.*, 2018] is thus a suitable solution. AB-MIL assigns attention scores to the instances, which reflect the importance of each instance in the

final decision-making stage. Mathematically, assuming the hidden representation of the i-th instance is $\boldsymbol{h}_i \in \mathbb{R}^{1 \times d}$ with the corresponding attention score a_i, then the bag-level representation $\boldsymbol{h}_b \in \mathbb{R}^{1 \times d_0}$ is calculated through Eq. (2):

$$\boldsymbol{h}_b = \sum_{i=1}^{n} a_i \boldsymbol{h}_i. \tag{2}$$

Under the AB-MIL framework, the characteristic of WSIs having multiple resolutions is neglected. In our work, we incorporate multiple resolutions of WSIs with AB-MIL and propose the HAG-MIL framework. The HAG-MIL framework is designed based on AB-MIL and its motivation is that the attention scores from lower resolutions can be used as a compass to guide the model at a higher resolution. The differences between AB-MIL and our HAG-MIL framework are illustrated in Fig. 1. The inputs of the HAG-MIL framework are the features from different resolutions, and after the model at the lowest resolution has assigned attention scores to the features, the features with low attention in the higher resolution are discarded in the model at a higher resolution. The process is repeated until the highest resolution is reached.

3.2 Hierarchical Attention-Guided MIL

The overview of our proposed HAG-MIL framework is illustrated in Fig. 2. Given the WSI resolutions level l to level 0, the instance representations $(\boldsymbol{F}^l, \cdots, \boldsymbol{F}^0)$ extracted from multiple resolutions of any given WSI slide and the ground truth bag label Y, the HAG-MIL framework learns a map from the instance representations $(\boldsymbol{F}^l, \cdots, \boldsymbol{F}^0)$ to the probability distribution of the bag label $(\hat{\boldsymbol{p}}^l, \cdots, \hat{\boldsymbol{p}}^0)$ and the attention scores of each instance $(\boldsymbol{a}^l, \cdots, \boldsymbol{a}^0)$.

Our HAG-MIL framework is defined in a recursive way. Given the instance representations $(\boldsymbol{F}^l, \cdots, \boldsymbol{F}^0)$, the IATs on different levels of resolutions $\mathcal{I} = \{\mathcal{I}_j | j = 0, \cdots, l\}$, the HAG-MIL framework is given in Eq. (3):

$$\boldsymbol{a}^j, \hat{\boldsymbol{p}}^j = \begin{cases} \mathcal{I}_j\left(\boldsymbol{F}^j\right), \text{ if } j = l, \\ \mathcal{I}_j\left(\boldsymbol{F}_d^j\right), \text{ if } j \neq l, \end{cases} \tag{3}$$

where $\boldsymbol{F}_d^j$ is the distilled features at level j, defined in Eq. (4):

$$\boldsymbol{F}_d^j = \boldsymbol{F}^j\left[\text{find\_topk\_ids}\left(\boldsymbol{a}^{j+1}, k_{j+1}\right)\right], \tag{4}$$

where k_{j+1} is the number of highest attention instances to be reserved at level $j + 1$, and find_topk_ids($\boldsymbol{a}^{j+1}, k_{j+1}$) first finds the indices of the instances with top-k_{j+1} attention scores at level $j + 1$ and then generate the distilled indices of instances at level j, which contains **exactly the same content** as the instances at level $j + 1$.

Finding the instances in different resolutions containing exactly the same content is a bit tricky, since patching and segmenting outcomes tend to be different in different resolutions, especially near the borderlines. For example, as shown in Fig. 2, patch i at level 1 has four sub-patches at level 0, *i.e.*, $4i, 4i + 1, 4i + 2, 4i + 3$; however, when we run the patching and segmenting algorithm at level 0 resolution, patches $4i + 2$ and $4i + 3$ will be dropped since they are backgrounds, while
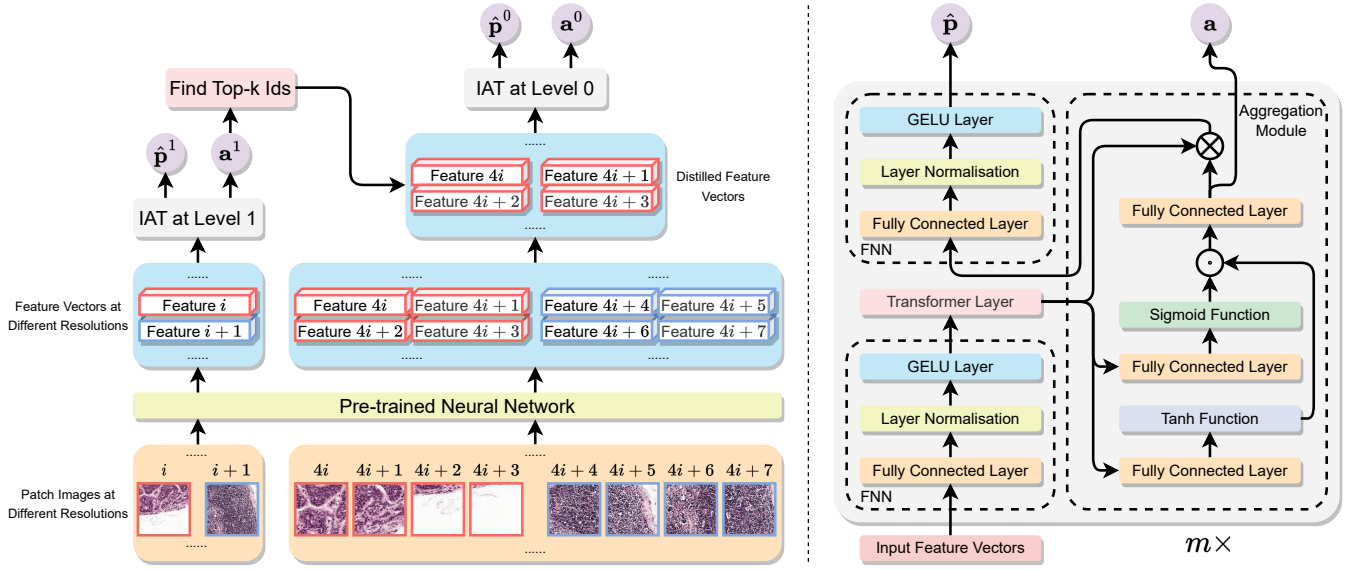
Figure 2: Overview of the HAG-MIL framework (left) and the model architecture of our proposed transformer model IAT (right). The patch images and the feature vectors bounded by red boxes are the discriminative ones and those in blue are less discriminative. The patch images cropped from different resolutions of WSIs are fed into a pre-trained network to obtain the feature vectors. The feature vectors are then passed to the framework in a hierarchical way to obtain the attention score $\boldsymbol{a}^1$ and $\boldsymbol{a}^0$, and the probability distribution of the bag label $\hat{\boldsymbol{p}}^1$ and $\hat{\boldsymbol{p}}^0$.

for the patch at level 1, patch i as a whole is reserved. If some patches are dropped as stated before and some are not, since they are not near the borderlines, it is difficult to retrieve the sub-patches by indexing in higher resolutions.

To retrieve patches more effectively, a quadtree structure for each WSI is constructed. This structure utilises the hierarchical nature of WSIs, where one patch node has four child nodes that contain the exact same content. As shown in Fig. 2, the patches are stored in the same sequence so that there is a clear map from an index in lower resolution to indices in higher resolutions, which is given in Eq. (5):

$$\text{find\_sub\_patch\_ids}(i) = (4i, 4i + 1, 4i + 2, 4i + 3). \quad (5)$$

Then the find_topk_ids can be expressed as Eq. (6):

$$\begin{aligned} &\text{find\_topk\_ids}(\boldsymbol{a}^{j+1}, k_{j+1}) \\ &= \text{find\_sub\_patch\_ids}(\text{argsort}(\boldsymbol{a}^{j+1}, k_{j+1})). \end{aligned} \quad (6)$$

After obtaining the pair of bag label probability distribution $\hat{\boldsymbol{p}}^j$ and attention scores $\boldsymbol{a}^j$ at level j, the model at level j is updated through back-propagation. Following the loss functions of **CL**ustering-constrained-**A**ttention **M**ultiple-instance learning (CLAM) [Lu *et al.*, 2021], the loss function in our HAG-MIL framework is calculated through Eq. (7):

$$\mathcal{L}^j = \sum_{i=1}^{N} \text{CE}\left(\hat{\boldsymbol{p}}^j, Y\right) + \lambda\, \text{SS}\left(\hat{\boldsymbol{p}}_p^j, \boldsymbol{y}_p^j\right), \quad (7)$$

where $\text{CE}(\cdot, \cdot)$ and $\text{SS}(\cdot, \cdot)$ stand for the **C**ross-**E**ntropy loss and the **S**mooth **S**upport Vector Machine loss [Berrada *et al.*, 2018], λ is the balancing coefficient, $\hat{\boldsymbol{p}}_p^j \in \mathbb{R}^{k \times C}$ is the patch label probability distribution of the top-k highest attention patches and $\boldsymbol{y}_p^j = [Y, \cdots, Y]^T \in \mathbb{R}^{k \times 1}$ are the generated

labels of the top-k highest attention patches. The patch-level loss is designed upon the idea that the patches with the highest attention scores should have the same label as the bag label.

3.3 Integrated Attention Transformer

The architecture of the proposed model together with the HAG-MIL framework is illustrated in Fig. 2. The vanilla transformer layer outputs only the contextual hidden representation of the input. In contrast, our IAM produces not only the contextualised hidden representation of the input but also the attention scores and the bag representation of the WSI slide. Furthermore, our IAM does not require a class token and comprises three constituents.

Feedforward Neural Network (FNN) The FNN in Fig. 2 is a sequence of modules, a fully connected layer, a layer normalisation module, and a **G**aussian **E**rror **L**inear **U**nit (GELU) layer [Hendrycks and Gimpel, 2016]. Given the input feature matrix $\boldsymbol{H} \in \mathbb{R}^{n \times d_1}$ and a learnable matrix $\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}$, the $\text{FNN}(d_1, d_2)$ is given in Eq. (8):

$$\text{FNN}(d_1, d_2)\,(\boldsymbol{H}) = \text{GELU}\,(\text{LayerNorm}\,(\boldsymbol{H}\boldsymbol{W})). \quad (8)$$

Transformer Layer After the first FNN, we incorporate a transformer layer [Vaswani *et al.*, 2017] to capture latent relationships between patches and provide cross-patch attention and additional contextual information to the original representations extracted by the pre-trained neural network. However, the time and space complexity of the vanilla transformer layer is $\mathcal{O}(n^2)$, where n is the number of patches and it is usually extremely huge for WSIs. To reduce time and memory costs, we employ Nyström attention [Xiong *et al.*, 2021] as our base transformer layer [Shao *et al.*, 2021], which has achieved competitive performance with the vanilla transformer layer and the time and space complexity are both

$\mathcal{O}(n)$. In addition, we also include the adaptive model initialization [Liu *et al.*, 2020] to stabilise the training process.

Aggregation Module We use the gated attention mechanism [Ilse *et al.*, 2018] as the aggregation module in our work, and we formulate it as a two-component module that can be easily extended in future research.

The aggregation module can be divided into two components: (1) the attention component $s(\cdot) : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times 1}$, which assigns attention scores $\boldsymbol{a} = s(\boldsymbol{H}) \in \mathbb{R}^{n \times 1}$ to the instances in the bag, and (2) the aggregation component $g(\cdot, \cdot) : (\mathbb{R}^{n \times 1}, \mathbb{R}^{n \times d}) \to \mathbb{R}^{1 \times d}$, which takes the attention scores $\boldsymbol{a}$ and the instance representations $\boldsymbol{H}$ as inputs and produces the bag representation $\boldsymbol{h}_b$. The aggregation module $f_a(\cdot) : \mathbb{R}^{n \times d} \to \mathbb{R}^{1 \times d}$ is the composition of the two components as described in Eq. (9):

$$\boldsymbol{h}_b = f_a(\boldsymbol{H}) = g(s(\boldsymbol{H}), \boldsymbol{H}). \quad (9)$$

There are multiple options for the attention component $s(\cdot)$, including mean pooling, max pooling, the attention mechanism, the gated attention mechanism and so on. We propose using the gated attention mechanism [Ilse *et al.*, 2018] as the aggregation module $s(\cdot)$ and it is shown in Eq. (10):

$$\boldsymbol{a} = s(\boldsymbol{H}) = (\tanh{(\boldsymbol{HV})} \odot \sigma{(\boldsymbol{HU})}) \boldsymbol{w}_a, \quad (10)$$

where $\boldsymbol{w}_a \in \mathbb{R}^{d_2 \times 1}$, $\boldsymbol{V}, \boldsymbol{U} \in \mathbb{R}^{d_1 \times d_2}$ are learnable matrices, $\boldsymbol{H} \in \mathbb{R}^{n \times d_1}$ is the hidden representation of the instances, $\sigma(\cdot)$ is the sigmoid function, $\tanh(\cdot)$ is the tanh function and $\odot$ stands for the element-wise production.

The aggregation component $g(\cdot, \cdot)$ can be the weighted average on the instance representation based on the attention scores. The bag representation $\boldsymbol{h}_b \in \mathbb{R}^{1 \times d}$ is given according to Eq. (2). Therefore, $f_a(\cdot)$ in our paper is given by Eq. (11):

$$f_a(\boldsymbol{H}) = ((\tanh{(\boldsymbol{HV})} \odot \sigma{(\boldsymbol{HU})}) \boldsymbol{w}_a)^T \boldsymbol{H}. \quad (11)$$

Finally, since an IAT consists of m IAMs, it produces m bag representations. Bag representations from different IAMs may capture different features of the bag; therefore, different bag representations should have different importance scores when aggregated into the final bag representation. Hence, we use an attention mechanism to fuse the bag representations from each IAM into the IAT bag representation. Assuming the bag representation from the i-th IAM is $\boldsymbol{h}_{b_i}$, and the learnable vector $\boldsymbol{w}_b \in \mathbb{R}^{m \times 1}$ with each element corresponding to the contribution of each $\boldsymbol{h}_{b_i}$, the final bag representation $\boldsymbol{h}_{bf} \in \mathbb{R}^{1 \times d}$ is given in Eq. (12):

$$\boldsymbol{h}_{bf} = \boldsymbol{w}_b^T \begin{bmatrix} \boldsymbol{h}_{b_1} \\ \vdots \\ \boldsymbol{h}_{b_m} \end{bmatrix}. \quad (12)$$

The last step is to feed the $\boldsymbol{h}_{bf}$ into a classification layer to obtain the bag-level distribution $\hat{\boldsymbol{p}}$.

4 Experiments and Results

We first compared our method with other methods on various datasets. Then we conducted extensive sensitivity and ablation studies to validate the effectiveness of the HAG-MIL framework and the aggregation module. Finally, we visualised the attention scores from our method and compared them with those from the CLAM model.

4.1 Dataset Descriptions

Camelyon16 Dataset [Litjens *et al.*, 2018] It contains 397 WSIs of lymph nodes with and without metastasis in tissue sections of women with breast cancer. The training dataset includes 157 normal WSIs and 111 tumour WSIs. The test dataset is provided separately and contains 129 test WSIs, including both tumour and normal WSIs. We further split the training dataset into a new training and validation dataset by 7:3 and evaluate all of the methods on the official test dataset.

IMGC Dataset The **I**ntestinal **M**etaplasia **G**astric **C**ancer (IMGC) dataset is an in-house dataset consisting of 1,882 WSIs, obtained from the Periodic Acid-Schiff-Alcian Blue stained biopsies of the patients diagnosed with intestinal metaplasia. The patients were followed to determine whether they ultimately developed gastric cancer. We used only a portion of this dataset due to its extreme imbalance. The training dataset includes 130 normal and 43 cancer cases, the validation dataset includes 34 normal and 13 cancer cases, and the test dataset includes 253 normal and 18 cancer cases.

TCGA-RCC Dataset It includes 940 WSIs in total. Individually, project Kidney Chromophobe Renal Cell Carcinoma (TGCA-KICH) has 121 WSIs from 109 cases. Project Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) has 519 WSIs from 513 cases, and project Kidney Renal Papillary Cell Carcinoma (TCGA-KIRP) has 300 WSIs from 276 cases. We split the dataset into training, validation, and test datasets by the ratio of 6:1.5:2.5, respectively.

TCGA-NSCLC Dataset It includes 1,053 WSIs in total. Individually, project Lung Squamous Cell Carcinoma (TCGA-LUSC) has 512 WSIs from 478 cases, and project Lung Adenocarcinoma (TCGA-LUAD) has 541 WSIs from 478 cases. We split the dataset into training, validation, and test datasets by the ratio of 6:1.5:2.5, respectively.

4.2 Implementation Details

Evaluation Metrics The evaluation criteria are **A**rea **U**nder the **C**urve (AUC), F1, and accuracy scores. The AUC score reflects the overall ability to distinguish different classes. The F1 score measures the true performance when the data is imbalanced. Together, these metrics can holistically assess the model performances. The thresholds for F1 and accuracy scores are determined by Youden's index in the IMGC dataset due to its extreme unbalance and they are set to 0.5 for other datasets, following previous works.

Training Settings Baseline methods include max/mean pooling, AB-MIL [Ilse *et al.*, 2018], CLAM-SB, CLAM-MB [Lu *et al.*, 2021], TransMIL [Shao *et al.*, 2021], and DTFD-MIL [Zhang *et al.*, 2022]. Training is terminated after 20 epochs of no decrease in validation loss. Official implementations of the methods were used when available, every effort was made to reproduce the results. For the IMGC dataset, all methods were run at level 0 resolution. For other datasets, baseline methods were run at level 1 resolution, following their respective works. The models were run five times and the average performances were reported.

Hyper-parameters Details The learning rate and weight decay are set to 10^{-5}. Dropout is set to 0.25 for the gated attention mechanism and 0.3 for the transformer layer. The Swin-Transformer [Liu *et al.*, 2021] large is used to extract the fea-

METHOD	Camelyon16			Gastric		
	AUC↑	F1↑	Accuracy↑	AUC↑	F1↑	Accuracy↑
Max Pooling	$0.715_{0.007}$	$0.549_{0.012}$	$0.682_{0.005}$	$0.871_{0.007}$	$0.601_{0.053}$	$0.776_{0.071}$
Mean Pooling	$0.654_{0.075}$	$0.437_{0.035}$	$0.640_{0.013}$	$0.839_{0.011}$	$0.551_{0.014}$	$0.704_{0.020}$
AB-MIL [Ilse *et al.*, 2018]	$0.903_{0.013}$	$0.850_{0.018}$	$0.868_{0.015}$	$0.886_{0.016}$	$0.582_{0.028}$	$0.743_{0.038}$
CLAM-SB [Lu *et al.*, 2021]	$0.834_{0.068}$	$0.782_{0.059}$	$0.819_{0.043}$	$0.838_{0.045}$	$0.533_{0.063}$	$0.673_{0.091}$
CLAM-MB [Lu *et al.*, 2021]	$0.876_{0.024}$	$0.800_{0.031}$	$0.826_{0.019}$	$0.873_{0.018}$	$0.578_{0.023}$	$0.742_{0.033}$
TransMIL [Shao *et al.*, 2021]	$0.838_{0.047}$	$0.787_{0.033}$	$0.809_{0.027}$	$0.834_{0.016}$	$0.560_{0.046}$	$0.731_{0.082}$
DTFD-MIL [Zhang *et al.*, 2022]	$0.933_{0.009}$	$0.858_{0.017}$	**$0.899_{0.012}$**	**$0.890_{0.011}$**	$0.565_{0.020}$	$0.725_{0.030}$
HAG-MIL (Ours)	**$0.946_{0.011}$**	**$0.874_{0.025}$**	$0.887_{0.022}$	**$0.890_{0.011}$**	**$0.616_{0.027}$**	**$0.781_{0.027}$**
Improvement	0.013	0.016	-0.012	0.000	0.015	0.005

Table 1: Results on Camelyon16 and our gastric cancer test set. The best ones are in red bold, and the second best ones are in blue italic underline. The subscript in each cell is the standard derivation.

METHOD	TCGA-RCC			TCGA-NSCLC		
	AUC↑	F1↑	Accuracy↑	AUC↑	F1↑	Accuracy↑
Max Pooling	$0.970_{0.000}$	$0.847_{0.008}$	$0.880_{0.005}$	$0.863_{0.001}$	$0.774_{0.002}$	$0.774_{0.002}$
Mean Pooling	$0.959_{0.010}$	$0.828_{0.015}$	$0.862_{0.014}$	$0.851_{0.000}$	$0.777_{0.002}$	$0.777_{0.002}$
AB-MIL [Ilse *et al.*, 2018]	$0.978_{0.002}$	$0.864_{0.013}$	$0.892_{0.010}$	$0.903_{0.007}$	$0.817_{0.010}$	$0.817_{0.010}$
CLAM-SB [Lu *et al.*, 2021]	$0.970_{0.007}$	$0.846_{0.011}$	$0.878_{0.010}$	$0.905_{0.004}$	$0.824_{0.007}$	$0.824_{0.007}$
CLAM-MB [Lu *et al.*, 2021]	$0.975_{0.005}$	$0.871_{0.012}$	$0.895_{0.010}$	$0.900_{0.004}$	$0.818_{0.010}$	$0.818_{0.010}$
TransMIL [Shao *et al.*, 2021]	$0.972_{0.003}$	$0.850_{0.007}$	$0.876_{0.011}$	$0.902_{0.008}$	$0.825_{0.015}$	$0.825_{0.016}$
DTFD-MIL [Zhang *et al.*, 2022]	$0.976_{0.009}$	$0.884_{0.028}$	$0.898_{0.027}$	$0.902_{0.023}$	$0.830_{0.027}$	$0.838_{0.030}$
HAG-MIL (Ours)	**$0.982_{0.001}$**	**$0.894_{0.009}$**	**$0.914_{0.006}$**	**$0.921_{0.008}$**	**$0.849_{0.015}$**	**$0.849_{0.016}$**
Improvement	0.004	0.010	0.016	0.016	0.019	0.011

Table 2: Results on TCGA-RCC and TCGA-NSCLC test set.

tures from the IMGC dataset. ResNet-50 [He *et al.*, 2016] is used for other datasets, following previous works. Therefore, the input dimension for the IMGC dataset and other datasets are 1,536 and 1,024, respectively. The dimensions of the stacked IAMs are determined via experimental validation. Interestingly, our experiment found that the optimal structure for the model consisted of four modules with dimensions of 1,024, 1,536, 512, and 1,024, respectively, forming a bottleneck structure. The dimension of the final bag representation is 1,024. The loss functions for models at different resolutions are different because we tolerate slight errors in lower resolutions, but the model at the highest resolution must be as accurate as possible. Therefore, label smoothing is set to 0.1 for low resolutions, and is not used for the highest resolution. The balancing coefficient λ is set to 1.

4.3 Comparison with State-of-the-art Methods

Table 1 and Table 2 show all results of the experiments. Our method had the highest AUC and F1 scores and the highest or competitive accuracy scores across all datasets, which proved the effectiveness of all our proposed methods.

Camelyon16 is quite challenging due to the small area of tumour regions. From Table 1, max/mean pooling can barely learn the distinctions between tumour and normal slides, and the CLAM and TransMIL are inferior to our method by at least 4.3% on the AUC score. Our method achieved a 1.3% improvement in AUC score and a 1.6% improvement in F1

score compared to DTFD-MIL, demonstrating its efficacy.

For the IMGC dataset, the fact that our method has achieved the best performance has proved that our proposed IAT is superior to other methods. Furthermore, the data imbalance problem on the IMGC dataset is much more severe than the other datasets, as the number of patients who developed gastric cancer is much less than those who did not. Hence, the result proved that our model performs well on extremely imbalanced datasets as well.

For the TCGA-RCC and TCGA-NSCLC datasets, they share a common characteristic that the tumour regions are large, which sometimes takes around 80% of the whole WSI, making most of the methods to have similar performance, and even the most naive approaches like mean pooling and max pooling have achieved very high AUC scores. Our method has outperformed all others in every metric on these two datasets by a significant margin, particularly on the TCGA-NSCLC dataset. In addition, the highest AUC score of our method on the TCGA-RCC dataset demonstrated that our model also excels in the multi-class classification task.

4.4 Sensitivity and Ablation Studies

We performed sensitivity and ablation analyses to assess the impact of our HAG-MIL framework and IAT on performance.

The HAG-MIL Framework The number of patches (k_j) reserved for higher resolution is a crucial parameter in our framework. k_j should not be too small to avoid missing tu-
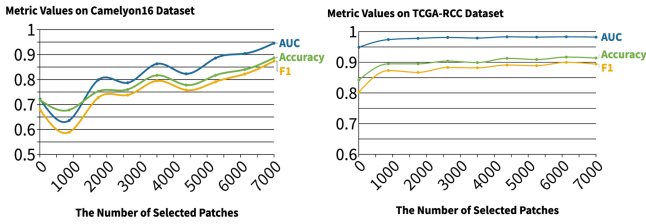
Figure 3: AUC, F1 and accuracy scores with different k values on the Camelyon16 (left) and TCGA-RCC (right) datasets.

METHOD	Camelyon16		
	AUC↑	F1↑	Accuracy↑
CLS w/o gated	$0.917_{0.040}$	$0.834_{0.026}$	$0.855_{0.020}$
CLS w/ gated	$0.929_{0.020}$	$0.855_{0.044}$	$0.869_{0.038}$
Aggr w/o gated	$0.928_{0.012}$	$0.848_{0.040}$	$0.863_{0.038}$
Aggr w/ gated	$\mathbf{0.946}_{0.011}$	$\mathbf{0.874}_{0.025}$	$\mathbf{0.887}_{0.022}$

Table 3: The results on the use of several aggregation modules and the class token on the Camelyon16 dataset.

mour patches, but it should also not be too large to avoid time and GPU memory issues. Since tumour regions are small in Camelyon16 and large in the TCGA series dataset, they tend to have different trends with changes of k_j; therefore, we conducted experiments on both Camelyon16 and TCGA-RCC datasets. For convenience, we set $k_1 = k_2$. Eight different k_j values are sampled within the range of [0, 7,000] at equal intervals, with the starting point of $k_j = 7$. In general, larger values of k_j tend to result in improved model performance. However, the extent of this influence varies across different datasets. The k_j value heavily influences how the model performs on Camelyon16 due to the small tumour regions. Our findings also indicate that increasing the value of k_j during the inference stage can further enhance model performance. However, the k_j value does not influence the performance on the TCGA-RCC dataset much because the tumour regions are large, and there is a high probability that the true discriminative patches are in the top-k_j patches even when $k_j = 7$.

Aggregation Module We compared transformers with and without an aggregation module and included two types of aggregation modules in our experiments: the attention mechanism [Ilse *et al.*, 2018] and the gated attention mechanism [Ilse *et al.*, 2018]. In order for a fair comparison, the patch-level loss is also included in transformers using the class token, and the attention scores used by our HAG-MIL framework are generated by an aggregation module in the final layer. The bag representations are still produced by the class token. The results are shown in Table 3. From these results, we can draw two conclusions: (1) the gated attention mechanism is a better option for the aggregation module as it outperforms the attention mechanism transformer by 1.8% and 1.2% in terms of the AUC score, and (2) our transformer with both two aggregation modules outperform the class token transformer by 1.7% and 1.1% in terms of the AUC score.

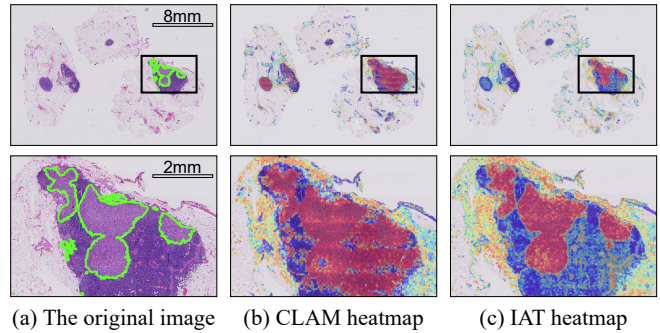(a) The original image (b) CLAM heatmap (c) IAT heatmap

Figure 4: Figures depicting a WSI from Camelyon16 with ground truth annotations (a), alongside heatmaps generated by CLAM-SB (b) and IAT (c). The second row of figures displays magnified views of the regions enclosed by black boxes in the first row.

4.5 Visualisation of the Attention Scores

We demonstrate the interpretability of our method by visualising the attention scores as heatmaps, which indicate the locations of the discriminative regions. For this experiment, we used the Camelyon16 dataset with ground truth pixel-level annotations and compared the heatmaps of our method with those of the CLAM-SB model. The original image and visualisation results are shown in Fig. 4. The ground truth tumour regions are bounded by green lines and the red regions in the heatmaps indicate the model-predicted high-attention regions. While the CLAM-SB model can capture most of the macro-metastasis in the WSI, it has a much higher false positive rate than our method. Additionally, the CLAM-SB model performs poorly at identifying micro-metastasis and isolated tumour cells due to their small size. However, our method accurately predicts macro-metastasis and can also localise micro-metastasis and isolated tumour cells with relatively smaller attention scores compared to macro-metastasis. It is clear that heatmaps from our method have better contrast and higher consistency with ground truth annotations.

5 Conclusion

In this paper, we present the HAG-MIL framework and the IAT for WSI classification tasks. The HAG-MIL framework fully utilises the WSIs by exploiting multiple resolutions of the WSIs, and it enables the large models, like our IAT, to run at the highest resolution of the WSIs. Diagnosing like a human pathologist, the HAG-MIL framework is capable of dynamically localising the most discriminative instances from the enormous number of instances in the WSI bag across multiple resolutions in a hierarchical and attentive way. Within the HAG-MIL framework, the IAT takes into account the evolvement of the bag representations through different transformer layers. It produces a more holistic and effective bag representation than the vanilla transformer by aggregating multiple bag representations generated by the IAMs. The experimental results show that our method outperforms other methods and obtains a more accurate heatmap. Our method can be further extended to more applications such as super-resolution image analysis, natural language processing, *etc*.

Acknowledgments

References

[Bejnordi *et al.*, 2015] Babak Ehteshami Bejnordi, Geert Litjens, Meyke Hermsen, Nico Karssemeijer, and Jeroen A. W. M. van der Laak. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In *Medical Imaging 2015: Digital Pathology*, 2015.

[Berrada *et al.*, 2018] Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Smooth loss functions for deep top-k classification. In *ICLR*, 2018.

[Campanella *et al.*, 2019] 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, August 2019.

[Chen *et al.*, 2021] Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew FK. Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *ICCV*, 2021.

[Chen *et al.*, 2022] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*, 2022.

[Chikontwe *et al.*, 2020] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple Instance Learning with Center Embeddings for Histopathology Classification. In *MICCAI*, 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, June 2019.

[Dimitriou *et al.*, 2019] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D. Caie. Deep learning for whole slide image analysis: An overview. *Frontiers in Medicine*, 2019.

[Hashimoto *et al.*, 2020] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale Domain-adversarial Multiple-instance CNN for Cancer Subtype Classification with Unannotated Histopathological Images. In *CVPR*, June 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv:1606.08415 [cs.LG]*, 2016.

[Hou *et al.*, 2016] 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 *CVPR*, 2016.

[Huang *et al.*, 2006] Kaizhu Huang, Haiqin Yang, Irwin King, and M.R. Lyu. Maximizing sensitivity in medical diagnosis using biased minimax probability machine. *IEEE Transactions on Biomedical Engineering*, 53(5):821–831, 2006.

[Huang *et al.*, 2021] Ziwang Huang, Hua Chai, Ruoqi Wang, Haitao Wang, Yuedong Yang, and Hejun Wu. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In *MICCAI*, 2021.

[Ianni *et al.*, 2020] Julianna D. Ianni, Rajath E. Soans, Sivaramakrishnan Sankarapandian, Ramachandra Vikas Chamarthi, Devi Ayyagari, Thomas G. Olsen, Michael J. Bonham, Coleman C. Stavish, Kiran Motaparthi, Clay J. Cockerell, Theresa A. Feeser, and Jason B. Lee. Tailored for Real-World: A Whole Slide Image Classification System Validated on Uncurated Multi-Site Data Emulating the Prospective Pathology Workload. *Scientific Reports*, 10(1):3217, February 2020.

[Ilse *et al.*, 2018] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *ICML 2018*. PMLR, July 2018.

[Kanavati *et al.*, 2020] 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):9297, June 2020.

[Keikhosravi *et al.*, 2020] Adib Keikhosravi, Bin Li, Yuming Liu, Matthew W. Conklin, Agnes G. Loeffler, and Kevin W. Eliceiri. Non-disruptive collagen characterization in clinical histopathology using cross-modality image synthesis. *Communications Biology*, 3(1):1–12, July 2020.

[Li *et al.*, 2019] Meng Li, Lin Wu, Arnold Wiliem, Kun Zhao, Teng Zhang, and Brian C. Lovell. Deep Instance-Level Hard Negative Mining Model for Histopathology Images. *arXiv:1906.09681 [cs]*, June 2019.

[Li *et al.*, 2020] Zhenliang Li, Liming Yuan, Haixia Xu, Rui Cheng, and Xianbin Wen. Deep Multi-Instance Learning with Induced Self-Attention for Medical Image Classification. In *BIBM 2020*, December 2020.

[Li *et al.*, 2021a] 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 *CVPR*, 2021.

[Li *et al.*, 2021b] Jiayun Li, Wenyuan Li, Anthony Sisk, Huihui Ye, W. Dean Wallace, William Speier, and Corey W. Arnold. A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in Biology and Medicine*, 131:104253, 2021.

[Li *et al.*, 2022] Xintong Li, Chen Li, Md Mamunur Rahaman, Hongzan Sun, Xiaoqi Li, Jian Wu, Yudong Yao, and Marcin Grzegorzek. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. *Artificial Intelligence Review*, 55(6):4809–4878, August 2022.

[Litjens *et al.*, 2018] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, and Jeroen van der Laak. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6), June 2018.

[Liu *et al.*, 2020] Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *EMNLP*, 2020.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[Lu *et al.*, 2019] 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:1910.10825 [cs, q-bio]*, November 2019. arXiv: 1910.10825.

[Lu *et al.*, 2021] Ming Y. Lu, Drew F. K. 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, June 2021.

[Maron and Lozano-Pérez, 1997] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in NIPS*, 1997.

[Myronenko *et al.*, 2021] Andriy Myronenko, Ziyue Xu, Dong Yang, Holger R. Roth, and Daguang Xu. Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging. In *MICCAI*, 2021.

[Peng and King, 2008] Xiang Peng and Irwin King. Robust BMPM training based on second-order cone programming and its application in medical diagnosis. *Neural Networks*, 21(2-3):450–457, 2008.

[Shao *et al.*, 2021] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *NIPS*, 2021.

[Srinidhi *et al.*, 2021] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep neural network models for computational histopathology: A survey. *MIA*, 2021.

[Tokunaga *et al.*, 2019] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive Weighting Multi-Field-Of-View CNN for Semantic Segmentation in Pathology. In *CVPR*, June 2019.

[Tourniaire *et al.*, 2021] Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, and Herve Delingette. Attention-based Multiple Instance Learning with Mixed Supervision on the Camelyon16 Dataset. In *MICCAI Workshop on Computational Pathology*, September 2021.

[Tu *et al.*, 2022] Chao Tu, Yu Zhang, and Zhenyuan Ning. Dual-curriculum contrastive multi-instance learning for cancer prognosis analysis with whole slide images. In *NeurIPS*, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[Wang *et al.*, 2018] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74, 2018.

[Wang *et al.*, 2020] 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*, 2020.

[Wang *et al.*, 2021] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *MICCAI*, 2021.

[Xiong *et al.*, 2021] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. *AAAI*, 2021.

[Zhang *et al.*, 2022] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *CVPR*, June 2022.

[Zhao *et al.*, 2020] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, and Jianhua Yao. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *CVPR*, June 2020.

[Zheng *et al.*, 2022] Yi Zheng, Rushin H. Gindra, Emily J. Green, Eric J. Burks, Margrit Betke, Jennifer E. Beane, and Vijaya B. Kolachalama. A graph-transformer for whole slide image classification. *TMI*, 2022.

[Zhou and Lu, 2022] Yuanpin Zhou and Yao Lu. Deep hierarchical multiple instance learning for whole slide image classification. In *ISBI*, 2022.