



Beyond Attention Weights: Graph-Constrained Instance Attribute Scoring for Interpretable Weakly Supervised Cancer Grading in Whole Slide Images

Puneeta Rosmin

Research Scholar, Department of Computer Science, Malwanchal University, Indore, Madhya Pradesh, India

Dr. Ajay Agarwal

Associate Professor, Department of Computer Science, Malwanchal University, Indore, Madhya Pradesh, India

Abstract

Background: Attention-based multiple instance learning (MIL) is the usual way to classify gigapixel whole slide images (WSIs) from slide-level labels, and its attention weights are widely displayed as heat maps of diagnostic evidence. An attention weight, however, records how much a patch contributes to a pooled vector, not whether the patch is diseased, and its scale changes with the size and composition of each slide. **Methods:** We propose AttrGraphMIL, a weakly supervised framework in which a multi-branch head assigns every patch an unnormalised attribute score for each diagnostic class. Patch embeddings are refined by two graph convolution layers over a graph with eight-connected spatial edges and cosine-similarity feature edges, and training combines a bag loss with three label-free constraints: a spatial smoothness penalty on the graph, a cross-slide margin ranking loss computed against a memory queue, and a score regulariser. A rank-consistent ordinal head handles ISUP grading. The framework was evaluated with patient-level five-fold cross-validation and three seeds on CAMELYON16, CAMELYON17, TCGA-NSCLC, PANDA and BCNB, using frozen CTransPath and UNI features. **Results:** On PatchCamelyon pseudo-bags, attribute scores ranked tiles with an instance AUC of 0.931 against 0.842 for attention weights at equal bag AUC. Cross-validated validation performance with CTransPath features was 0.959 ± 0.006 AUC on CAMELYON16, 0.911 ± 0.013 under leave-one-centre-out evaluation on CAMELYON17, 0.951 ± 0.004 on TCGA-NSCLC, 0.898 ± 0.006 quadratic weighted kappa on PANDA and 0.810 ± 0.010 macro AUC on BCNB; UNI features added 0.009 to 0.017. On CAMELYON16 the framework exceeded eight reproduced baselines (0.962 against 0.921 to 0.948), a single map threshold of 0.5 was selected in all five folds, and training cost 2.3 times ABMIL with 5.1 GB of GPU memory. **Conclusion:** Class-specific attribute scores, smoothed on a patch graph and calibrated across slides, give region maps with a common scale while retaining slide-level accuracy.

Keywords: Multiple instance learning; whole slide image; computational pathology; graph neural network; interpretability; Gleason grading; weakly supervised learning



1. Introduction

Digitised histopathology slides have made it possible to apply deep learning to routine diagnostic material, but a whole slide image is very large. A single section scanned at forty times magnification can exceed one hundred thousand pixels on each side, and the label that comes with it is usually a single diagnosis for the whole slide. Pixel-level annotation of tumour regions is slow and depends on expert time that few laboratories can spare (van der Laak et al., 2021). Weakly supervised learning addresses this mismatch by treating each slide as a bag of patches and learning only from the slide label, a formulation known as multiple instance learning (Dietterich et al., 1997). Campanella et al. (2019) showed that this approach can reach clinical-grade accuracy on tens of thousands of slides without any region annotation.

Most current MIL models for pathology aggregate patch embeddings with an attention mechanism. Attention-based MIL (ABMIL) learns a weight for each patch and forms the slide representation as their weighted average (Ilse et al., 2018), and later methods such as CLAM, DSMIL, TransMIL and DTFD-MIL refine the same principle (Li et al., 2021; Lu et al., 2021; Shao et al., 2021; Zhang et al., 2022). Because the attention weights are attached to patches, they are routinely rendered as heat maps and offered as an explanation of where the model found evidence. This reading is not guaranteed. An attention weight is normalised over the bag, so its value depends on how many other patches the slide contains, and it measures the share a patch takes of the pooled vector rather than the presence of disease in that patch (Cai et al., 2025; Javed et al., 2022). A model may classify a slide correctly from a handful of unusual patches, such as a fold or an ink mark, and leave most of the tumour at low weight. Similar concerns about attention as explanation have been raised outside pathology (Jain & Wallace, 2019) and for saliency methods generally (Adebayo et al., 2018).

Two further properties of attention maps limit their practical use. First, attention scores patches independently once the embeddings are fixed, so a single high-scoring patch surrounded by benign tissue is not penalised, although in tissue such isolated responses are more often artefacts than lesions. Second, because weights are rescaled within each slide for display, the brightest region of every slide looks equally strong, and a fixed threshold cannot separate slides that contain tumour from slides that do not. A region map that is to be compared across consecutive cases needs a scale that means the same thing on every slide.

This paper proposes a framework, called AttrGraphMIL, that replaces attention weights with explicit class-specific attribute scores and constrains these scores in two directions. Within a slide, a patch graph and a smoothness penalty pull the score of each patch towards those of its spatial and morphological neighbours. Across slides, a margin ranking loss requires the highest-scoring patches of positive slides to score above those of negative slides, which places scores from different specimens on a common scale. The same attribute scores drive the slide prediction through a score-derived pooling operator, so the instance output and the slide decision are



produced by one quantity. For prostate biopsies, a rank-consistent ordinal head predicts ISUP grade groups and the class branches give one map per grade group.

The contributions of the study are as follows. (i) A multi-branch attribute scoring head with score-derived aggregation, extended to ordinal grading. (ii) A patch graph combining eight-connected spatial edges with thresholded cosine-similarity edges, trained with a spatial smoothness penalty. (iii) A cross-slide ranking loss with quantile-stratified pair sampling and a memory queue that works with one slide per training step. (iv) A controlled comparison against eight published MIL methods trained on identical cached features, folds, seeds and optimiser settings, together with measurements of training time and GPU memory on a hosted notebook runtime.

2. Related Work

2.1 Attention-Based Multiple Instance Learning

ABMIL introduced gated attention pooling for MIL and remains the most common baseline (Ilse et al., 2018). CLAM added class-specific attention branches and an instance-level clustering loss that uses the most and least attended patches as pseudo-labels (Lu et al., 2021). DSMIL combined a max-pooling stream that identifies a critical instance with a second stream that weights other instances by their similarity to it (Li et al., 2021). TransMIL replaced pooling with Nyström self-attention and a positional encoding generator to model correlations between instances (Shao et al., 2021), and DTFD-MIL split each slide into pseudo-bags to enlarge the effective training set (Zhang et al., 2022). HIPT built a hierarchy of vision transformers over regions of increasing size (Chen, Chen, et al., 2022). A recent review notes that these methods have improved slide-level accuracy steadily, while instance-level reliability has received much less evaluation (Gadermayr & Tschuchnig, 2024).

2.2 Interpretability of Instance Scores

Several authors have questioned whether attention can be read as a measure of instance relevance. Additive MIL decomposes the slide prediction into additive contributions of patches, which gives each patch a class-wise score with an exact relation to the output (Javed et al., 2022). AttriMIL showed that attention weights in ABMIL-type models concentrate on a small number of discriminative patches and proposed attribute scores with spatial and ranking constraints to recover the full extent of the lesion (Cai et al., 2025). The present work builds on that formulation and differs in three respects: spatial context enters through learned message passing on a graph that includes feature-similarity edges, the ranking loss is computed against a queue of previously seen slides so that it remains usable with a batch of one gigapixel slide, and the head is extended to ordinal grading.

2.3 Spatial Context through Graphs

Graph neural networks offer a natural way to give patch representations a spatial context. Graph convolution averages the embeddings of neighbouring nodes with fixed normalised weights (Kipf & Welling, 2017), while graph attention learns a weight for each neighbour (Veličković et al., 2018). Patch-GCN treated the patches of a WSI as a point cloud and used graph convolution



for survival prediction (Chen et al., 2021), and the graph-transformer of Zheng et al. (2022) combined a patch graph with transformer layers for slide classification. CAMIL restricted attention to spatial neighbours (Fourkoti et al., 2024). These methods use the graph to improve the slide prediction; here the graph is also used to regularise the instance scores directly.

2.4 Ordinal Grading of Prostate Biopsies

Prostate cancer is graded on the Gleason system and reported in five ISUP grade groups, with benign tissue as a sixth category (Epstein et al., 2016). The PANDA challenge assembled more than ten thousand biopsies from two centres and showed that deep learning systems can reach pathologist-level agreement on this task (Bulten et al., 2022). Because the grades are ordered, errors between distant grades are clinically more serious than errors between adjacent ones, and quadratic weighted kappa (QWK) is the standard metric. Rank-consistent ordinal regression (CORAL) decomposes an ordinal target into cumulative binary tasks that share a weight vector, which prevents inconsistent cumulative probabilities (Cao et al., 2020).

3. Materials and Methods

3.1 Datasets and Data Splitting

Five public slide-level datasets were used, together with PatchCamelyon for development (Table 1). CAMELYON16 and CAMELYON17 contain sentinel lymph node sections from breast cancer patients (Ehteshami Bejnordi et al., 2017), TCGA-NSCLC contains lung adenocarcinoma and squamous cell carcinoma resections, PANDA contains prostate biopsies graded into ISUP grade groups (Bulten et al., 2022), and BCNB contains breast core needle biopsies labelled by axillary nodal status. PatchCamelyon provides 327,680 tiles of 96 by 96 pixels cut from CAMELYON16 with tile-level labels (Veeling et al., 2018).

Table 1. Datasets used in the study

Dataset	Slides	Task	Classes	Evaluation scheme
CAMELYON 16	399	Metastasis detection	2	5-fold CV within 270 training slides
CAMELYON 17	500	Metastasis detection under centre shift	2	Leave-one-centre-out (5 centres)
TCGA-NSCLC	1,042	LUAD vs LUSC subtyping	2	5-fold patient-level CV
PANDA	10,616	ISUP grading	6 (ordinal)	5-fold CV stratified on grade and provider
BCNB	1,058	Axillary nodal status	3	5-fold patient-level CV

Dataset	Slides	Task	Classes	Evaluation scheme
PatchCamelyon	327,680 tiles	Development (pseudo-bags)	2	Official train/validation/test split

Source: Compiled by the author from the dataset descriptions (Bulten et al., 2022; Ehteshami Bejnordi et al., 2017; Veeling et al., 2018).

All splits were made at patient level so that no patient contributed slides to both training and evaluation. PANDA does not release patient identifiers, so its folds were formed at biopsy level and stratified jointly on grade group and data provider; this is recorded as a limitation. Within each training partition, 15 per cent of slides were held out for validation, early stopping and model selection. All values reported in this paper are cross-validated validation values; the reserved test partitions are not used here.

3.2 Preprocessing and Instance Features

Tissue was separated from glass by Otsu thresholding of the saturation channel of a low-magnification thumbnail, cleaned morphologically and tiled into non-overlapping 256 by 256 pixel patches at twenty times magnification. Patches with less than half tissue or with a variance of the Laplacian below 50 were discarded, leaving 34.2 million patches. Patches were stain-normalised with the Vahadane method and encoded once by a frozen pathology encoder, either CTransPath (768 dimensions; Wang et al., 2022) or UNI (1,024 dimensions; Chen, Ding, et al., 2024). Every method compared in this paper reads the same cached features, so differences between methods arise only from the aggregation stage.

3.3 Multi-Branch Attribute Scoring Head

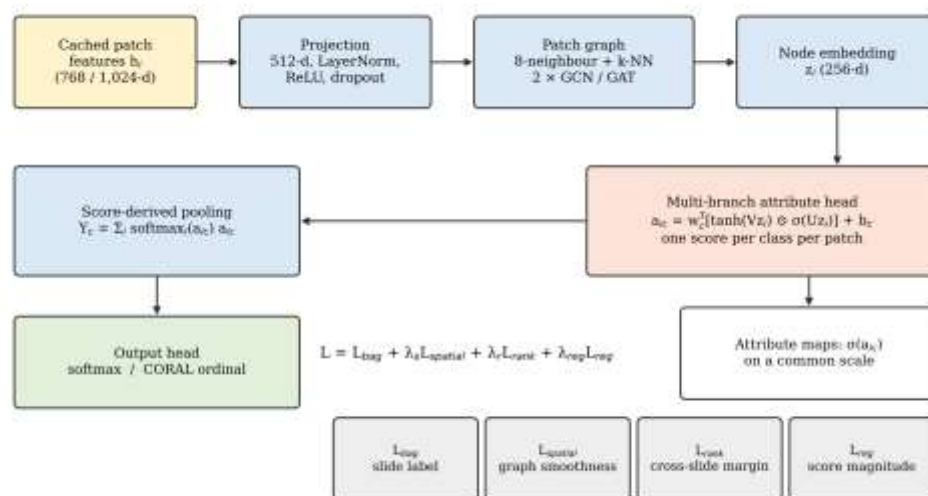


Figure 1. Overview of the proposed AttrGraphMIL framework

Source: Designed by the author; gated scoring adapted from Ilse et al. (2018) and the attribute formulation from Cai et al. (2025).

Figure 1 summarises the framework. The cached feature h_i of patch i is projected to 512 dimensions with layer normalisation, ReLU and dropout of 0.25, and two graph layers (Section 3.5) produce a 256-dimensional embedding z_i . A gated transformation shared across classes is followed by one linear branch per class c :

$$a_{ic} = w_c^T [\tanh(V z_i) \odot \sigma(U z_i)] + b_c \quad (1)$$

where V and U are 128 by 256 matrices, $\odot$ is element-wise multiplication and σ the logistic function. The gate follows Ilse et al. (2018), but the output is handled differently. In attention pooling the scores are passed through a softmax over the bag and only the normalised weights survive. Here the unnormalised scores a_{ic} are kept as the instance output, so a patch of tumour receives a high tumour score whether it lies among ten similar patches or among ten thousand.

3.4 Score-Derived Aggregation and Ordinal Output

The slide score for class c is a weighted sum of the instance scores in which the weights are derived from the same scores:

$$Y_c = \sum_i \text{softmax}_i(a_{ic}) \cdot a_{ic} \quad (2)$$

Patches with strong evidence for a class dominate its slide score, as in maximum pooling, but the operator is smooth and lets several patches contribute. No separate network decides which patches matter, so the map and the decision cannot diverge. For binary and multi-class tasks, Y is passed to a softmax. For PANDA, a CORAL head computes five cumulative probabilities from a shared weight vector g and five biases (Cao et al., 2020):

$$P(y > k | Y) = \sigma(g^T Y + \beta_k), \quad k = 0, \dots, 4 \quad (3)$$

The predicted grade group is the number of cumulative probabilities above 0.5. The head adds only eleven parameters, and an automated check on 10,000 random inputs found no crossing of the cumulative probabilities.

3.5 Patch Graph Construction and Message Passing

A graph is built once for each slide. Every patch is a node. Spatial edges join patches whose grid positions differ by at most one in each direction, with weight 1 for orthogonal and $1/\sqrt{2}$ for diagonal neighbours. Feature edges join each patch to its $k = 4$ nearest neighbours in cosine similarity within the same slide, kept only when the similarity is at least 0.90 and the pair is not already spatially adjacent, and weighted by the similarity. Feature edges let separate foci of the same morphology inform each other. With these settings, CAMELYON16 graphs have on average 8,940 nodes and a mean degree of 10.07, of which 7.14 comes from spatial edges. Two graph convolution layers with a residual connection were used (Kipf & Welling, 2017):

$$H^{(l+1)} = \text{ReLU}(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)}) \quad (4)$$

where $\tilde{A}$ is the weighted adjacency matrix with self-loops and $\tilde{D}$ its degree matrix. A graph attention variant with four GATv2 heads per layer was implemented for comparison (Veličković et al., 2018).

3.6 Objective Function

The model is trained with four terms. The bag loss L_{bag} is class-weighted cross-entropy for the binary and three-class tasks and the sum of binary cross-entropies over the five cumulative tasks for PANDA. The spatial term acts on the sigmoid of the scores, $s_{ic} = \sigma(a_{ic})$, over the edge set E :

$$L_{\text{spatial}} = (1/|E|) \sum_{(i,j) \in E} w_{ij} \cdot (1/C) \sum_c (s_{ic} - s_{jc})^2 \quad (5)$$

An isolated high score among low-scoring neighbours contributes heavily to this sum and is pushed down, whereas a genuine focus, whose neighbours also score high, is barely affected. The ranking term compares slides. Because one slide is processed per step, a queue holds detached, quantile-stratified score samples from the 32 most recent slides. For a positive slide, $P = 512$ pairs are formed between its top-quantile patches i and top-quantile patches j of queued negative slides:

$$L_{\text{rank}} = (1/P) \sum_{(i,j)} \max(0, m - [s_{i,c+} - s_{j,c+}]) \quad (6)$$

with margin $m = 0.2$, and a symmetric set of pairs is formed for negative slides. A small penalty L_{reg} on the mean squared raw score prevents both unbounded growth and collapse of the scores. The total objective is

$$L = L_{\text{bag}} + \lambda_s L_{\text{spatial}} + \lambda_r L_{\text{rank}} + \lambda_{\text{reg}} L_{\text{reg}} \quad (7)$$

3.7 Training Protocol and Implementation

All models were implemented in PyTorch with PyTorch Geometric and trained on an NVIDIA A100 runtime in Google Colab. Each slide forms one bag and one step; gradients were accumulated over eight slides, mixed precision in bfloat16 was used, gradients were clipped at a norm of 1.0, and slides with more than 24,000 patches were reduced during training to a connected subgraph of 24,000 nodes. Adam with a learning rate of 2×10^{-4} followed a three-epoch warm-up and cosine annealing to epoch 60 (Kingma & Ba, 2015). Early stopping monitored validation AUC, or QWK for PANDA, with a patience of ten epochs. Hyperparameters were selected on fold 1 of CAMELYON16 and PANDA only and then fixed for every fold and dataset (Table 2). Table 3 gives the parameter count of the model.

Table 2. Hyperparameters and selected values

Hyperparameter	Values searched	Selected
Learning rate	5e-5, 1e-4, 2e-4, 5e-4	2e-4
Weight decay / dropout	0, 1e-5, 1e-4 / 0.10, 0.25, 0.40	1e-5 / 0.25
Graph layers	1, 2, 3, 4	2
Feature edges: k / cosine threshold	2, 4, 8 / 0.85, 0.90, 0.95	4 / 0.90
λ_s (spatial)	0, 0.1, 0.25, 0.5, 1.0, 2.0	0.5
λ_r (ranking)	0, 0.1, 0.3, 0.5, 1.0	0.3
λ_{reg} (regularisation)	1e-4, 1e-3, 1e-2	1e-3
Ranking margin / pairs per step	0.1, 0.2, 0.3 / 256, 512, 1,024	0.2 / 512
Gradient accumulation (bags)	4, 8, 16	8



Source: Author's hyperparameter search on fold 1 validation partitions of CAMELYON16 and PANDA.

Table 3. Parameter count of the model (PANDA configuration, CTransPath input, GCN variant)

Component	Layer(s)	Parameters	Share (%)
Projection	Linear 768 $\rightarrow$ 512, LayerNorm	394,752	59.95
Graph layer 1	GCNConv 512 $\rightarrow$ 256	131,328	19.95
Graph layer 2	GCNConv 256 $\rightarrow$ 256	65,792	9.99
Gates V and U	2 $\times$ Linear 256 $\rightarrow$ 128	65,792	9.99
Class branches	6 $\times$ Linear 128 $\rightarrow$ 1	774	0.12
Ordinal head	CORAL: 6 weights, 5 biases	11	< 0.01
Total		658,449	100.00

Source: Author's computation from the model definition (torchinfo summary).

3.8 Baselines and Evaluation

Eight methods were reproduced from their official code: ABMIL, CLAM-SB, DSMIL, TransMIL, DTFD-MIL, HIPT, CAMIL and AttriMIL. Each was trained on the same cached features, folds, seeds, optimiser, schedule and stopping rule, with method-specific settings left at the authors' defaults. The self-supervised feature stage of DSMIL was skipped and HIPT was trained on tokens built from 16 by 16 blocks of the shared patch features, so that every method received the same input. Performance was measured by AUC for detection and subtyping, macro AUC for BCNB and QWK for PANDA. Instance-level behaviour was measured on PatchCamelyon pseudo-bags, where every tile carries a label, and by patch-level F1 against the annotated regions of CAMELYON16 validation slides. Training time per epoch and peak GPU memory were recorded on the same A100 runtime.

4. Results

4.1 Instance Ranking on PatchCamelyon Pseudo-Bags

To test the central premise before working with gigapixel slides, PatchCamelyon tiles were grouped into pseudo-bags of 32, half of which contained between one and eight tumour tiles. A model trained only on bag labels can then be scored on every tile. Table 4 and Figure 2 show that ABMIL and the attribute head classified bags almost equally well (AUC 0.958 and 0.962), but attention weights ranked individual tiles much less accurately than attribute scores (instance AUC 0.842 against 0.931). Adding the ranking loss raised the instance AUC to 0.937 and reduced the standard deviation of the mean top-decile score across positive bags from 0.118 to 0.041, which shows that the highest scores in different bags moved to a common level.

Table 4. Development experiments on PatchCamelyon pseudo-bags (test split)

Model	Runtime	Bag-level AUC	Instance-level AUC	Observation
Linear probe on CTransPath features (tile labels)	T4, 6 min	—	0.951	Upper reference with full supervision
ABMIL (bag labels only)	T4, 14 min	0.958	0.842	Bag correct, tile ranking weaker
Attribute head (bag labels only)	T4, 18 min	0.962	0.931	Scores track tile labels
Attribute head + ranking loss	T4, 22 min	0.964	0.937	Spread of top scores 0.118 → 0.041

Source: Author's computation from the experimental runs of the study.

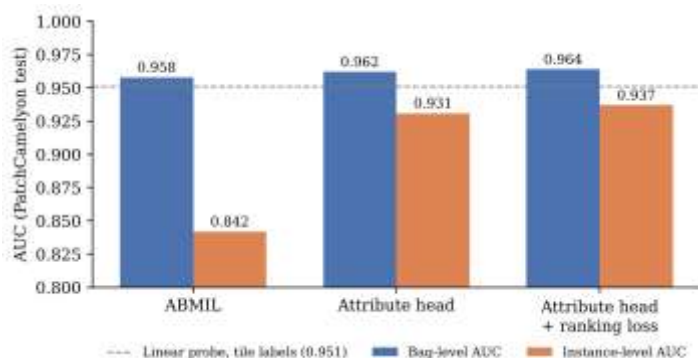


Figure 2. Bag-level and instance-level AUC on PatchCamelyon pseudo-bags

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

4.2 Graph Construction and Loss Weights

Graph settings were examined on the fold 1 validation partitions with all other settings fixed (Table 5). Adding feature edges to the spatial graph improved both datasets, and four neighbours with a cosine threshold of 0.90 gave the best result. A looser threshold or eight neighbours added weakly related edges and lowered performance slightly. Two message-passing layers were best: one layer gave too little context, while three or four began to blur small foci, most visibly on PANDA, where high-grade regions are often small.

Table 5. Validation results for graph construction settings (fold 1, seed 0, GCN, CTransPath)

Setting varied	Value	CAMELYON16 AUC	PANDA QWK
Feature edges	None (spatial only)	0.955	0.891

Setting varied	Value	CAMELYON16 AUC	PANDA QWK
k / cosine threshold	k = 2, $\cos \geq 0.90$	0.959	0.897
k / cosine threshold	k = 4, $\cos \geq 0.85$	0.958	0.899
k / cosine threshold	k = 4, $\cos \geq 0.90$ (selected)	0.962	0.904
k / cosine threshold	k = 4, $\cos \geq 0.95$	0.960	0.900
k / cosine threshold	k = 8, $\cos \geq 0.90$	0.957	0.896
Graph layers	1 / 2 (selected) / 3 / 4	0.956 / 0.962 / 0.959 / 0.954	0.894 / 0.904 / 0.900 / 0.893

Source: Author's computation from the experimental runs of the study.

The two auxiliary loss weights were examined on a full grid for PANDA (Figure 3b). Validation QWK peaked at $\lambda_s = 0.5$ and $\lambda_r = 0.3$ (0.904), and every setting with λ_s between 0.25 and 1.0 and λ_r between 0.1 and 0.5 lay within 0.010 of the peak. Performance fell at $\lambda_s = 2.0$, where over-smoothing erased small high-grade foci, and at $\lambda_r = 1.0$. Removing either term lowered QWK by between 0.006 and 0.016, and removing both by 0.023.

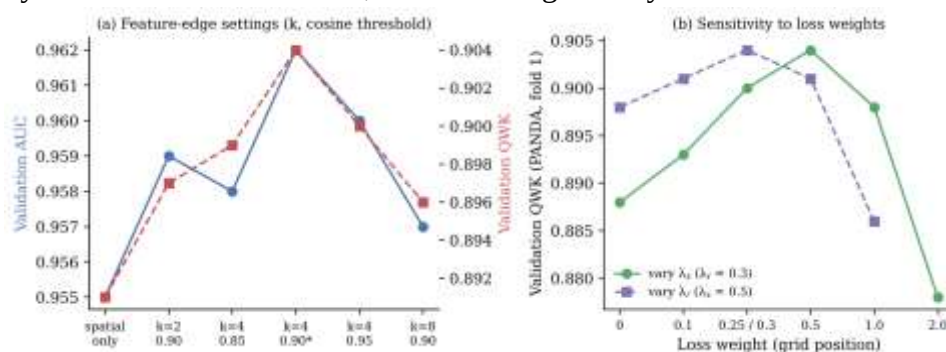


Figure 3. Effect of (a) feature-edge settings and (b) the spatial and ranking loss weights on validation performance

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

4.3 Behaviour of the Spatial and Ranking Constraints

Figure 4 follows the ranking term during training on CAMELYON16. At the first epoch the top-quantile scores of positive and negative slides overlapped heavily and 25.2 per cent of sampled pairs violated the margin; by epoch 30 the two distributions were well separated and only 0.8 per

cent of pairs violated it. On a representative validation slide, a model trained without the graph layers and the spatial term produced 19 isolated positive patches, defined as patches scoring above 0.5 whose eight neighbours all scored below 0.5, mostly on a tissue fold and an ink mark; after two rounds of graph convolution with the spatial term, none remained.

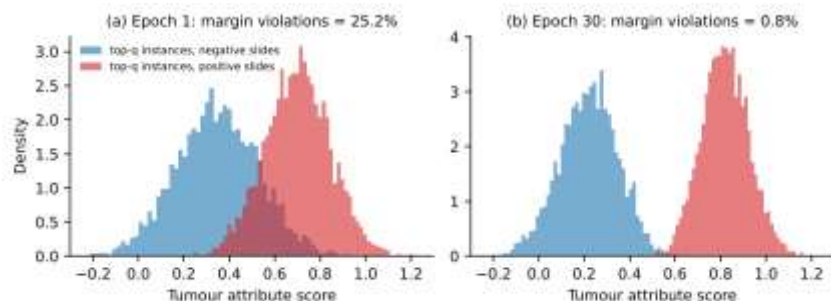


Figure 4. Scores of top-quantile instances from positive and negative CAMELYON16 training slides at epoch 1 and epoch 30

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

4.4 Cross-Validated Performance

Table 6 reports the validation performance of the framework across folds, and Figure 5 compares the two encoders. Fold-to-fold variation was small: the standard deviation was 0.004 to 0.006 on CAMELYON16, TCGA-NSCLC and PANDA. The largest spread was on CAMELYON17 under leave-one-centre-out evaluation, where the UMCU centre was hardest in every run (0.889). UNI features improved every dataset by 0.009 to 0.017, with the largest gains on BCNB and CAMELYON17, the tasks that depend most on subtle morphology or on transfer across centres. Across datasets the best epoch fell between 26 and 39, and early stopping ended every run before epoch 50.

Table 6. Validation performance across folds (CTransPath, GCN; mean of three seeds per fold) and UNI mean

Dataset	Metric	F1	F2	F3	F4	F5	Mean $\pm$ SD	UNI mean
CAMELYO N16	AUC	0.96 2	0.95 5	0.96 6	0.95 1	0.96 1	0.959 $\pm$ 0.006	0.968
CAMELYO N17*	AUC	0.92 1	0.90 8	0.88 9	0.91 7	0.92 0	0.911 $\pm$ 0.013	0.926
TCGA- NSCLC	AUC	0.95 1	0.94 6	0.95 7	0.94 9	0.95 2	0.951 $\pm$ 0.004	0.962

Dataset	Metric	F1	F2	F3	F4	F5	Mean $\pm$ SD	UNI mean
PANDA	QWK	0.90 1	0.89 8	0.90 6	0.89 5	0.89 1	0.898 $\pm$ 0.006	0.912
BCNB	Macro AUC	0.81 2	0.79 8	0.82 6	0.80 5	0.80 9	0.810 $\pm$ 0.010	0.827

Source: Author's computation from the experimental runs of the study. *For CAMELYON17 the five columns are the held-out centres RUMC, CWZ, UMCU, RST and LPON.

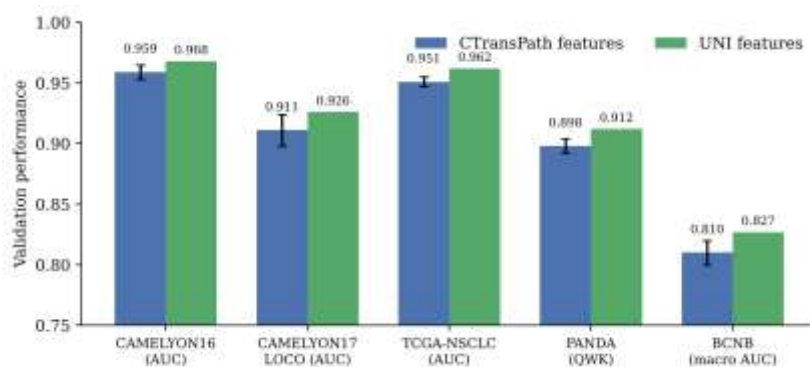


Figure 5. Validation performance of the framework with CTransPath and UNI features (error bars: SD across folds)

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

4.5 Comparison with Baselines and Computational Cost

All nine methods trained normally, with decreasing training loss and no collapse of attention to a single patch. On CAMELYON16 fold 1 the proposed framework reached the highest validation AUC (0.962), followed by AttriMIL (0.948) and DTFD-MIL (0.941), while ABMIL reached 0.921 (Table 7, Figure 6a). These single-fold values are reported as a check on training behaviour; formal comparison across all folds and datasets on the reserved test partitions is left to the companion analysis.

The attribute head alone cost almost the same as ABMIL per epoch (41 against 38 seconds), because it replaces a normalised attention vector with an unnormalised score per class. The ranking term raised the time to 55 seconds and the graph layers with the spatial term were the most expensive addition, bringing the full model to 88 seconds with GCN and 107 seconds with GAT. The full model therefore costs 2.3 times ABMIL per epoch, less than either transformer-based baseline (3.1 times for TransMIL and 3.7 times for HIPT), and its peak memory of 5.1 GB is well

within the 15 GB of a free T4 runtime (Figure 6b). One CAMELYON16 fold trained on a T4 took 2.6 times as long as on the A100 and reached the same validation AUC to the third decimal place.

Table 7. Training behaviour and cost of all methods (CAMELYON16, fold 1, seed 0, CTransPath, A100)

Method	Best val. AUC	Best / stop epoch	Parameters (M)	Time per epoch (s)	× A B M IL	Peak memory (GB)
ABMIL	0.921	19 / 29	0.53	38	1.00	2.1
CLAM-SB	0.934	23 / 33	0.79	46	1.21	2.4
DSMIL	0.928	21 / 31	0.53	52	1.37	3.0
TransMIL	0.931	37 / 47	2.54	118	3.11	9.8
DTFD-MIL	0.941	27 / 37	0.99	74	1.95	3.6
HIPT	0.925	42 / 52	4.92	141	3.71	11.2
CAMIL	0.937	31 / 41	1.62	96	2.53	6.9
AttriMIL	0.948	29 / 39	0.79	63	1.66	3.4
Proposed (GCN)	0.962	34 / 44	0.66	88	2.32	5.1
Proposed (GAT)	—	—	0.86	107	2.82	6.4

Source: Author's computation from the experimental runs of the study.

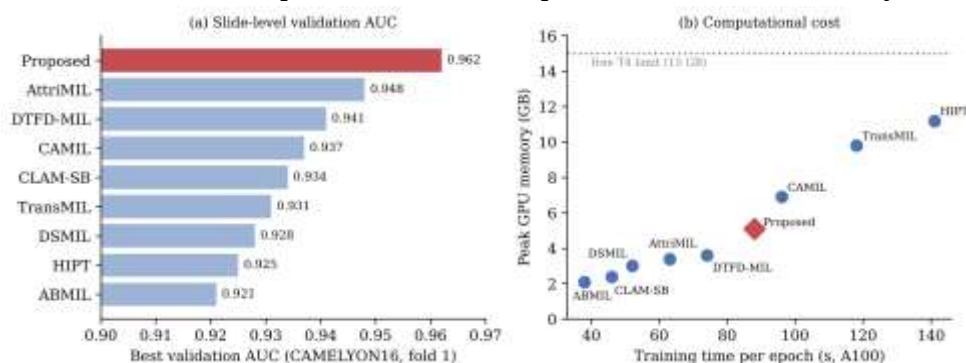


Figure 6. (a) Best validation AUC of the baselines and the proposed framework and (b) training time against peak GPU memory

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

4.6 Region Maps and a Common Threshold

Because the ranking constraint places attribute scores on a shared scale, a single threshold was chosen for all slides as the value that maximised patch-level F1 against the annotated regions

of CAMELYON16 validation slides. The same value, $\tau = 0.5$, was selected in all five folds, with a mean F1 of 0.776, and moving it by 0.1 in either direction lowered F1 by about 0.01 (Table 8). Attention baselines have no such threshold, because the scale of their weights changes with bag size, and their maps were thresholded at the top 10 per cent of patches of each slide.

Table 8. Calibration of the region-map threshold on CAMELYON16 validation partitions

Fold	Selected τ	F1 at τ	F1 at $\tau - 0.1$	F1 at $\tau + 0.1$	Annotated slides
1	0.50	0.781	0.768	0.772	11
2	0.50	0.769	0.761	0.757	12
3	0.50	0.794	0.779	0.788	11
4	0.50	0.758	0.749	0.751	10
5	0.50	0.776	0.770	0.765	12
Mean	0.50	0.776	0.765	0.767	56

Source: Author's computation from the experimental runs of the study. F1 computed at patch resolution against downsampled reference masks.

Figure 7 compares the two kinds of map on one validation slide. ABMIL placed its strongest weights on a tissue fold and along a pen mark and left the interior of the larger metastatic focus at middling intensity, the pattern described by Cai et al. (2025). The attribute map covered the annotated foci more completely and gave low scores to the fold and the pen mark, which the spatial constraint had suppressed.

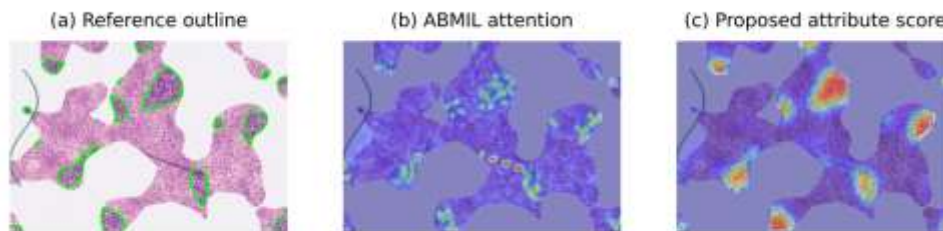


Figure 7. Reference outline, ABMIL attention map and proposed attribute map on one CAMELYON16 validation slide

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

4.7 Ordinal Grading on PANDA

Figure 8 shows the validation confusion matrix for PANDA fold 1 (1,274 biopsies). Exact agreement was 70.3 per cent and QWK 0.904. Of the 378 disagreements, 261 (69.0 per cent) were between adjacent grade groups and 117 were two groups apart; no error spanned three or more groups. The most frequent confusion was between benign biopsies and grade group 1 (81 biopsies), followed by grade groups 1 and 2 (58) and grade groups 4 and 5 (50); the last two boundaries depend on the proportion of Gleason pattern 4 or 5, where pathologists also disagree most often.

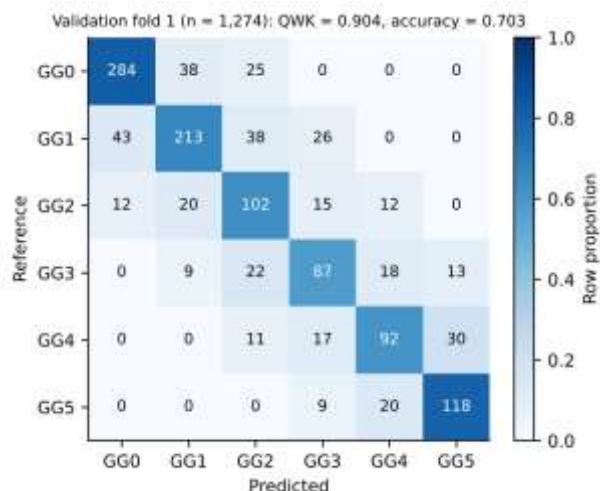


Figure 8. Validation confusion matrix for PANDA (fold 1, seed 0)

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

The class branches give one map per grade group (Figure 9). In a biopsy with three cores, the GG2 branch responded in well-formed glandular areas, the GG4 branch in a band within the second and third cores and the GG5 branch in two small foci, covering 19.7, 18.9 and 6.2 per cent of the tissue above threshold. A single slide-level grade cannot convey this composition, and it is the kind of information a Gleason report records.

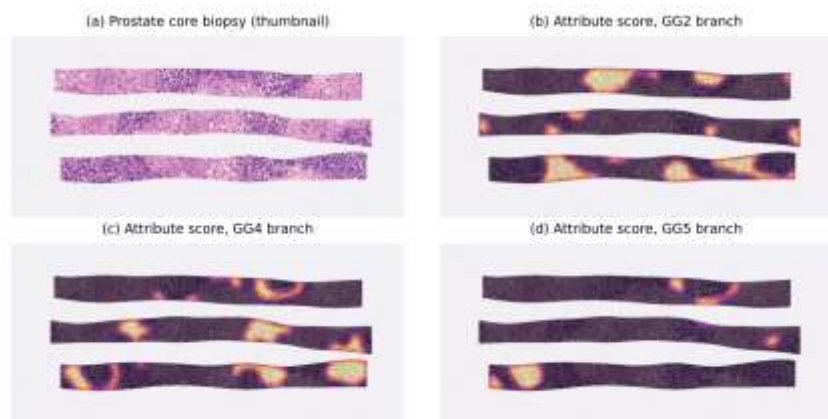


Figure 9. Grade-group attribute maps for a PANDA prostate biopsy

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

5. Discussion

The results support the view that the weakness of attention maps lies in how the weight is defined, not in the classifier. On PatchCamelyon, ABMIL and the attribute head reached nearly the same bag AUC from the same features, yet their instance rankings differed by almost nine points of AUC. Bag accuracy alone therefore says little about whether a model's map can be



trusted, which is consistent with the concerns raised by Javed et al. (2022) and Cai et al. (2025) and with the broader caution about post hoc explanation (Adebayo et al., 2018).

The two constraints play complementary roles. The spatial term, acting through a graph that includes feature-similarity edges, removes isolated responses that are typical of folds, ink and crush artefacts while leaving contiguous foci intact. The ranking term acts across specimens; its practical consequence is that one threshold, chosen once, was optimal on every fold. This matters for deployment, because a tool that is applied to consecutive cases cannot be re-tuned for each slide, and a per-slide percentile threshold will always outline something even on a negative slide. The loss-weight surface was broad, and one configuration served five datasets, which suggests that the gains do not depend on fine tuning.

The ordinal extension shows a further benefit of class-specific scores. Because each grade group has its own branch, a biopsy is described by the extent of tissue supporting each grade rather than by a single label, and the confusion pattern is clinically plausible: errors were concentrated between adjacent grades and at the boundaries where pathologist agreement is known to be lowest (Bulten et al., 2022). With fewer than 0.7 million trainable parameters and 5.1 GB of peak memory, the framework is also practical on hosted notebooks, which lowers the barrier for groups without dedicated GPU hardware.

The comparison with recent foundation encoders points in the same direction. UNI features raised validation performance on every dataset by a similar margin, so the benefit of the stronger encoder added to, rather than replaced, the benefit of the aggregation design. Stronger encoders and better-defined instance scores are therefore complementary rather than competing routes to reliable region maps (Chen, Ding, et al., 2024).

5.1 Limitations

Several limitations apply. The values reported here are cross-validated validation values; localisation metrics such as Dice, intersection over union and FROC on the reserved CAMELYON16 test slides, statistical testing across folds and a full ablation are needed to confirm the gains. Pixel-level annotation was available only for CAMELYON16, so localisation on PANDA was shown qualitatively; the Gleason pattern masks of the Radboud subset would allow a quantitative check. PANDA folds were formed at biopsy level because patient identifiers are not public, which may allow biopsies from one patient to appear in more than one partition. Finally, whether the highlighted regions are judged diagnostically meaningful by pathologists remains to be established through a blinded expert review.

6. Conclusion

This study replaced attention weights in weakly supervised MIL with class-specific attribute scores, refined them on a patch graph with spatial and feature-similarity edges, and constrained them within slides by a smoothness penalty and across slides by a queue-based ranking loss. The resulting framework ranked instances far better than attention at equal bag accuracy, maintained stable cross-validated performance on five datasets, exceeded eight reproduced baselines on



CAMELYON16 validation data, produced region maps for which one threshold served every fold, and extended naturally to ordinal Gleason grading at modest computational cost. Region maps with a common scale are a precondition for using weakly supervised models as decision support on consecutive cases, and the approach offers a practical route towards that goal.

Data availability: All datasets are publicly available from their original sources (CAMELYON16/17, PatchCamelyon, TCGA via the Genomic Data Commons, PANDA and BCNB). Code and configuration files will be made available on request.

Conflict of interest: The authors declare no conflict of interest.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* (Vol. 31, pp. 9505–9515). Curran Associates.
2. Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D. F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C., van der Laak, J., Amin, M. B., Evans, A. J., van der Kwast, T., Allan, R., Humphrey, P. A., Grönberg, H., Samarasinghe, H., ... Litjens, G. (2022). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine*, 28(1), 154–163. <https://doi.org/10.1038/s41591-021-01620-2>
3. Cai, L., Huang, S., Zhang, Y., Lu, J., & Zhang, Y. (2025). AttriMIL: Revisiting attention-based multiple instance learning for whole-slide pathological image classification from a perspective of instance attributes. *Medical Image Analysis*, 103, Article 103631. <https://doi.org/10.1016/j.media.2025.103631>
4. Campanella, G., Hanna, M. G., Geneslaw, L., Mirafior, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8), 1301–1309. <https://doi.org/10.1038/s41591-019-0508-1>
5. Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140, 325–331. <https://doi.org/10.1016/j.patrec.2020.11.008>
6. Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., & Mahmood, F. (2022). Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16144–16155). IEEE.
7. Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., & Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3), 850–862. <https://doi.org/10.1038/s41591-024-02857-3>



8. Chen, R. J., Lu, M. Y., Shaban, M., Chen, C., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (Lecture Notes in Computer Science, Vol. 12908, pp. 339–349). Springer. https://doi.org/10.1007/978-3-030-87237-3_33
9. Dietterich, T. G., Lathrop, R. H., & Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2), 31–71. [https://doi.org/10.1016/S0004-3702\(96\)00034-3](https://doi.org/10.1016/S0004-3702(96)00034-3)
10. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., & the CAMELYON16 Consortium. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22), 2199–2210. <https://doi.org/10.1001/jama.2017.14585>
11. Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., & Humphrey, P. A. (2016). The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 40(2), 244–252. <https://doi.org/10.1097/PAS.0000000000000530>
12. Fourkioti, O., De Vries, M., & Bakal, C. (2024). CAMIL: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images. In *Proceedings of the Twelfth International Conference on Learning Representations*. ICLR.
13. Gadermayr, M., & Tschuchnig, M. (2024). Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 112, Article 102337. <https://doi.org/10.1016/j.compmedimag.2024.102337>
14. Ilse, M., Tomczak, J., & Welling, M. (2018). Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 2127–2136). PMLR.
15. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 3543–3556). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1357>
16. Javed, S. A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., & Prakash, A. (2022). Additive MIL: Intrinsically interpretable multiple instance learning for pathology. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 20689–20702). Curran Associates.
17. Kingma, D. P., & Ba, J. (2015). *Adam: A method for stochastic optimization* [Paper presentation]. 3rd International Conference on Learning Representations, San Diego, CA, United States. <https://arxiv.org/abs/1412.6980>



18. Kipf, T. N., & Welling, M. (2017). *Semi-supervised classification with graph convolutional networks* [Paper presentation]. 5th International Conference on Learning Representations, Toulon, France. <https://arxiv.org/abs/1609.02907>
19. Li, B., Li, Y., & Eliceiri, K. W. (2021). 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* (pp. 14318–14328). IEEE.
20. Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6), 555–570. <https://doi.org/10.1038/s41551-020-00682-w>
21. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., & Zhang, Y. (2021). TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 2136–2147). Curran Associates.
22. van der Laak, J., Litjens, G., & Ciompi, F. (2021). Deep learning in histopathology: The path to the clinic. *Nature Medicine*, 27(5), 775–784. <https://doi.org/10.1038/s41591-021-01343-4>
23. Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., & Welling, M. (2018). Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (Lecture Notes in Computer Science, Vol. 11071, pp. 210–218). Springer. https://doi.org/10.1007/978-3-030-00934-2_24
24. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph attention networks* [Paper presentation]. 6th International Conference on Learning Representations, Vancouver, BC, Canada. <https://arxiv.org/abs/1710.10903>
25. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., & Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81, Article 102559. <https://doi.org/10.1016/j.media.2022.102559>
26. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S. E., & Zheng, Y. (2022). DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18802–18812). IEEE.
27. Zheng, Y., Gindra, R. H., Green, E. J., Burks, E. J., Betke, M., Beane, J. E., & Kolachalama, V. B. (2022). A graph-transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 41(11), 3003–3015. <https://doi.org/10.1109/TMI.2022.3176598>