CAMIL: CONTEXT-AWARE MULTIPLE INSTANCE LEARNING FOR CANCER DETECTION AND SUBTYPING IN WHOLE SLIDE IMAGES

Olga Fourkioti, Matt De Vries & Chris Bakal
The Institute of Cancer Research, London, United Kingdom
{ofourkioti;mvries;cbakal}@icr.ac.uk

ABSTRACT

The visual examination of tissue biopsy sections is fundamental for cancer diagnosis, with pathologists analyzing sections at multiple magnifications to discern tumor cells and their subtypes. However, existing attention-based multiple instance learning (MIL) models used for analyzing Whole Slide Images (WSIs) in cancer diagnostics often overlook the contextual information of tumor and neighboring tiles, leading to misclassifications. To address this, we propose the Context-Aware Multiple Instance Learning (CAMIL) architecture. CAMIL incorporates neighbor-constrained attention to consider dependencies among tiles within a WSI and integrates contextual constraints as prior knowledge into the MIL model. We evaluated CAMIL on subtyping non-small cell lung cancer (TCGA-NSCLC) and detecting lymph node (CAMELYON16 and CAMELYON17) metastasis, achieving test AUCs of 97.5%, 95.9%, and 88.1%, respectively, outperforming other state-of-the-art methods. Additionally, CAMIL enhances model interpretability by identifying regions of high diagnostic value[1].

1 INTRODUCTION

Deep learning (DL) methods have revolutionized the development of highly accurate diagnostic machines (Morales et al., 2021) that rival or even surpass the performance of expert pathologists (Tong et al., 2014; Melendez et al., 2015; Quellec et al., 2016; Das et al., 2018; Srinidhi et al., 2019; Wang et al., 2021). These advancements have been facilitated by the emergence of weakly supervised learning, which eliminates the need for laborious pixel-level annotations. Models trained using weakly supervised learning, relying solely on slide-level labels, have demonstrated exceptional classification accuracy on whole slide imaging (WSI) data, paving the way for scalable computational decision support systems in clinical practice (Xu et al., 2014; Courtiol et al., 2018; Xu et al., 2019; Zhou et al., 2021).

In the context of cancer histopathology, WSIs are not processed as a single image by DL models. Instead, WSIs are frequently subdivided into smaller tiles, which serve as an input. The task is, then, to classify the WSI based on the features extracted from the individual tiles. Most current methods for weakly supervised WSI classification use the Multiple Instance Learning (MIL) framework, which considers each WSI as a 'bag' of tiles and attempts to learn the slide-level label without prior knowledge about the labels of the individual tiles.

A major bottleneck in the deployment of MIL models, and the weakly-supervised learning paradigm in general, is that the MIL model is either permutation invariant, meaning that the tiles within a WSI exhibit no ordering among each other (Sharma et al., 2021; Xie et al., 2020), or permutation-aware without explicit information guidance. In other words, the spatial relationship of one tile to another is either ignored, or the dependencies between the tiles are implicitly modeled during training without requiring direct instructions (Shao et al., 2021; Landini et al., 2020; Campanella et al., 2019).

However, explicit knowledge about a tile's spatial arrangement is particularly relevant in cancer histopathology, where cancer and normal cells are not necessarily distributed randomly inside an

[1]Our code is available at https://github.com/olgarithmics/ICLR_CAMIL.

image. Contextual insights into the cellular landscape, such as the spatial dispersion of cells, the arrangement of cell clusters, and the broader characteristics of the tissue microenvironment, provide a more comprehensive view of the tile's local environment, enabling a better assessment of subtle variations and abnormalities that may indicate the presence of cancer.

In this paper, we propose a novel framework dubbed Context-Aware Multiple Instance Learning (CAMIL) to harness the dependencies among the individual tiles within a WSI and impose contextual constraints as prior knowledge on the multiple instance learning model. By explicitly accounting for contextual information, CAMIL aims to enhance the detection and classification of localized tumors and mitigate the potential misclassification of isolated or noisy instances, thereby contributing to an overall improvement in performance for both individual tiles and WSIs. Moreover, the attention weights enhance the interpretability of the model by highlighting sub-regions of high diagnostic value within the WSI.

2 Related work

Under the MIL formulation, the prediction of a WSI label (i.e., cancerous or not) can come either directly from the tile predictions (instance-based) (Campanella et al., 2019; Landini et al., 2020; Hou et al., 2016; Xu et al., 2019), or from a higher-level bag representation resulting from the aggregation of the tile features (bag embedding-based) (Ilse et al., 2018; Lu et al., 2021; Sharma et al., 2021; Wang et al., 2018). The bag embedding-based approach has empirically demonstrated superior performance (Sharma et al., 2021; Wang et al., 2018). Most recent bag embedding-based approaches employ attention mechanisms (Vaswani et al., 2017), which assign an attention score to every tile reflecting its relative contribution to the collective WSI-level representation. Attention scores enable the automatic localization of sub-regions of high diagnostic value in addition to informing the WSI-level label (Zhang et al., 2021; BenTaieb & Hamarneh, 2018; Lu et al., 2021).

Attention-based MIL models vary in how they explore tissue structure in WSIs. Many are permutation invariant, assuming the tiles are independent and identically distributed. Building upon this assumption, Ilse et al. (2018) proposed a learnable attention-based MIL pooling operator that computes the bag embedding as the average of all tile features in the WSI weighted by their respective attention score. This operator has been widely adopted and modified with the addition of a clustering layer (Lu et al., 2021; Li et al., 2021b; Yao et al., 2020) to further encourage the learning of semantically-rich, separable and class-specific features. Another variation of the same model uses 'pseudo bags' (Zhang et al., 2022), splitting the WSI into several smaller bags to alleviate the issue of the limited number of training data. Recently, data augmentation has been adopted to inflate the number of bags (Gadermayr et al., 2023; Liu et al., 2023; Shao et al., 2023).

However, permutation invariant operators cannot inherently capture the structural dependencies among different tiles at the input. The lack of bio-topological information has partially been remedied by the introduction of feature similarity scores instead of positional encodings to model the mutual tile dependencies within a WSI (Xie et al., 2020; Tellez et al., 2021; Adnan et al., 2020). For instance, DSMIL (Li et al., 2021a) utilizes a non-local operator to compute an attention score for each tile by measuring the similarity of its feature representation against that of a critical tile. To consider the correlations between the different tiles of a WSI, transformer-based architectures have been introduced, which usually make use of a learnable position-dependent signal to incorporate the spatial information of the image (Zhao et al.; Tu et al., 2019). For instance, TransMIL (Shao et al., 2021) is a transformer-like architecture trained end-to-end to optimize for the classification task and produce attention scores while simultaneously learning the positional embeddings.

In CAMIL, we provide explicit guidance regarding the context of every tile as we argue that it can provide a valuable, rich source of information. Unlike most existing MIL approaches where the relationships developed between neighboring tiles are omitted, in our approach, we propose a neural network architecture that explicitly leverages the dependencies between neighboring tiles of a WSI by enforcing bio-topological constraints to enhance performance effectively.

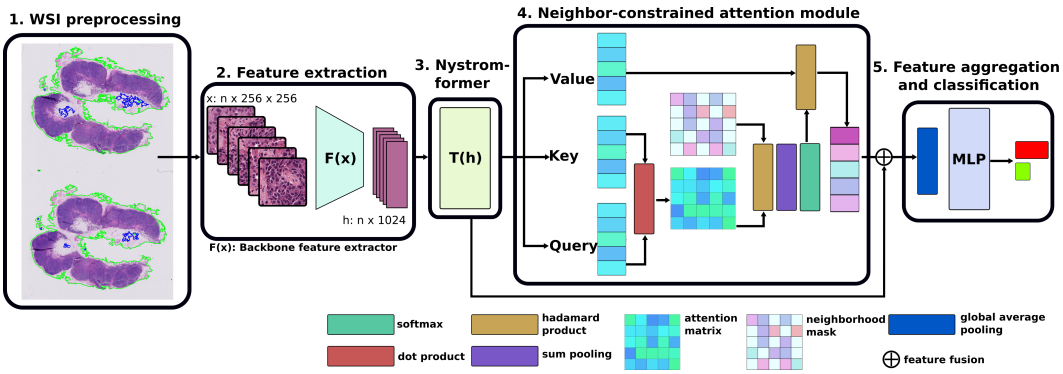

Figure 1: An overview of the CAMIL model architecture. First, WSIs are preprocessed to separate tissue from the background. Then, the WSIs are split into fixed-size tiles of size 256×256 and fed through a pre-trained feature extractor to obtain feature representations of size 1024 for each tile. A Nystromformer module then transforms these feature embeddings. These transformed feature embeddings are then used as input to our neighbor-constrained attention module. This module allows attending over each patch and its neighboring patches, generating a neighborhood descriptor of each tile's closest neighbors, and calculating their attention coefficients. The output layer then aggregates the tile-level attention scores produced in the previous layer to emit a final slide classification score.

3 Material and Methods

CAMIL operates on the principle that the context and characteristics of a tile's surroundings hold substantial potential for enhancing the accuracy of whole slide classification. To illustrate this concept, we can draw a parallel between our framework and the examination process of a pathologist analyzing a biopsy slide. Similar to how a pathologist inspects sub-regions to comprehensively understand its broader surroundings, CAMIL expands the tile's view to examine the broader neighborhood of each tile thoroughly. This extension allows CAMIL to gather additional information and facilitates a better understanding and assessment of the surrounding microenvironment and tissue context.

In CAMIL, we recalibrate each tile's individual attention score by aggregating the attention scores of its surroundings. For example, tiles with high attention scores surrounded by other high-scoring tiles should be considered important. Conversely, the presence of a tile classified by the model as important in a low-scoring neighborhood could be, in some cases, attributed to noise, and this should be reflected in its final attention score.

The overview of CAMIL can be seen in Figure 1. It can be decomposed into five elements:

1. A WSI-preprocessing phase automatically segments the tissue region of each WSI and divides it into many smaller tiles (e.g., 256 x 256 pixels).
2. A tile and feature extraction module, consisting of a stack of convolutional, max pooling, and linear layers transform the original tile input to low dimensional feature representations: $\boldsymbol{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_i, \ldots, \boldsymbol{h}_N\}, \boldsymbol{h}_i \in \mathbb{R}^{n \times d}$, where d is the embedding dimensions of a tile, n the number of tiles within a WSI (n differs among different WSIs), and N the number of WSIs.
3. A Nystromformer module (Xiong et al., 2021) transforms the tile embeddings to a concise, descriptive hidden feature representation. It is crucial in aggregating global contexts, capturing the overall information and patterns across multiple tiles.
4. A neighbor-constrained attention mechanism in CAMIL, coupled with a contrastive learning block, encapsulates the neighborhood prior and focuses on aggregating local concepts.
5. The feature aggregator and classification layer combine the local concepts derived from the previous layer with the features that describe the global contexts obtained from the transformer module. These features are merged to generate a prediction at the slide level.

We elaborate on each step in the following subsections.

3.1 WSI PREPROCESSING

We followed methods in Lu et al. (2021) to segment tissue regions and split the whole slide image into individual non-overlapping tiles (1. in Figure 1) (details of the hyperparameters used here are shown in the appendix).

3.2 FEATURE EXTRACTOR

Effective feature representations significantly impact predictive accuracy, as demonstrated by the success of self-supervised contrastive learning (Chen et al., 2020). Therefore, to extract rich, meaningful feature representations from individual tiles, we first train a feature extractor following the SimCLR (Chen et al., 2020) approach. SimCLR is one of the most popular self-supervised learning frameworks that enable semantically rich feature representations to be learned by minimizing the distance between different augmented versions of the same image data.

Similar to the training approach followed by Li et al. (2021a), the data sets utilized in SimCLR are composed of patches derived from WSIs. These patches are densely cropped with no overlap and treated as separate images for the purpose of SimCLR training. During training, two different augmentations are done on the same tile. These two augmentations are chosen from four possible augmentations (color distortion, zoom, rotation, and reflection) using a stochastic data augmentation module. These two augmentations of the same tile are fed through a ResNet-18 (He et al., 2015) pre-trained on ImageNet (Deng et al., 2009) with an additional projection head, which is a multi-layer perceptron (MLP) with two hidden layers that map the feature representations to a space where a contrastive loss is applied. The final convolutional block of ResNet-18 and the projection head are then fine-tuned by minimizing the contrastive loss (temperature-scaled cross entropy) between $\boldsymbol{z}_i, \boldsymbol{z}_j$, corresponding to two 'correlated' (differentially augmented) views of the same tile. Here, we minimized the normalized temperature-scaled cross entropy (NT-Xent) defined as

$$l_{ij} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)} \tag{1}$$

The trained network is then used as the base feature extractor ($F(\boldsymbol{x})$ in Figure 1) to produce the feature representations $\boldsymbol{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_i, \ldots, \boldsymbol{h}_N\}, \boldsymbol{h}_i \in \mathbb{R}^{n \times d}$ of each WSI, where n is the number of tiles and d is the embedding dimension to represent each tile, and N the number of WSIs. This trained feature extractor is frozen when training CAMIL and is only used to extract features and calculate distances between neighboring patches. These distances between neighboring patches are calculated using the sum of squared differences between the features and are used in the following neighbor-constrained attention module.

3.3 TRANSFORMER MODULE TO CAPTURE GLOBAL CONTEXTS

To encode the feature embeddings $\boldsymbol{H}$, our approach focuses on capturing the inter-tile relationships and dependencies, enhancing the global context understanding, and facilitating comprehensive feature aggregation. This is achieved using a transformer layer, represented as $T(\boldsymbol{h})$ in Figure 1, which is particularly effective for managing the complex structure of WSIs. To address the challenge of memory overload due to the long-range dependencies in large WSIs, we adopt the Nystromformer architecture (Xiong et al., 2021), enabling CAMIL to model intricate feature interactions through an efficient, approximate self-attention mechanism. This produces a "transformed" feature set $\boldsymbol{T} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_i, \ldots, \boldsymbol{t}_N\}$, with each $\boldsymbol{t}_i \in \mathbb{R}^{n \times d}$, where,

$$\boldsymbol{t}_i = \text{softmax}\left(\frac{\boldsymbol{Q_1}(\boldsymbol{h_i})\tilde{\boldsymbol{K}}_1^T(\boldsymbol{h_i})}{\sqrt{\text{d}_\text{k}}}\right)\left(\boldsymbol{A}\right)^+ \text{softmax}\left(\frac{\tilde{\boldsymbol{Q}}_1(\boldsymbol{h_i})\boldsymbol{K}_1^T(\boldsymbol{h_i})}{\sqrt{\text{d}_\text{k}}}\right)\boldsymbol{V_1}(\boldsymbol{h_i}), \tag{2}$$

where $\tilde{\boldsymbol{Q}}_1(\boldsymbol{h_i})$ and $\tilde{\boldsymbol{K}}_1(\boldsymbol{h_i})$ are the m selected landmarks (see Xiong et al. (2021) and Appendix) from the original n-dimensional sequence of $\boldsymbol{Q}_1$ and $\boldsymbol{K}_1$, $\boldsymbol{A}^+ = \text{softmax}\left(\frac{\tilde{\boldsymbol{Q}}_1(\boldsymbol{h_i})\tilde{\boldsymbol{K}}_1^T(\boldsymbol{h_i})}{\sqrt{\text{d}_\text{k}}}\right)^+$ is the approximate inverse of $\boldsymbol{A}$. Softmax is applied along the rows of the matrix. $\boldsymbol{K_1}(\boldsymbol{h_i})$, $\boldsymbol{Q_1}(\boldsymbol{h_i})$,

and $\boldsymbol{V_1}(\boldsymbol{h_i})$ are the first key, query, and value representations of $\boldsymbol{h_i}$ shown as $T(\boldsymbol{h})$ in Figure 1 and defined in Vaswani et al. (2017).

3.4 NEIGHBOR-CONSTRAINED ATTENTION MODULE TO CAPTURE LOCAL CONTEXTS

The neighbor-constrained attention module in CAMIL is designed to capture specific features and patterns within localized areas of the slide. It focuses on the immediate neighborhood of each tile and aims to capture the local relationships and dependencies within that specific region. By doing so, the module can emphasize the relevance and importance of nearby tiles, effectively incorporating fine-grained details and local nuances into the model.

To model the tile and its surroundings, we construct a weighted adjacency matrix. Consider an undirected graph $G = (V, E)$, where V represents the set of nodes representing image tiles, and E represents the set of edges between nodes indicating adjacency. The graph can be represented by an adjacency matrix $\boldsymbol{A}$ with elements $A_{i,j}$, where $A_{i,j} = s_{ij}$ if there exists an edge $(v_i, v_j) \in E$ and $A_{i,j} = 0$ otherwise. Each image tile must be connected to other tiles and can be surrounded by eight adjacent patches. Each element of the matrix s_{ij} represents the degree of similarity or resemblance between two connected tiles and is calculated as follows:

$$s_{ij} = \begin{cases} \exp -(\sqrt{(\boldsymbol{h}_i - \boldsymbol{h}_j)^2}), & (v_i, v_j) \in E \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

This design ensures injecting a bio-topological prior such that the weight of a tile is dependent on adjacent tiles with a similar pattern.

The transformed tile representations $\boldsymbol{T} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_i, \ldots, \boldsymbol{t}_N\}$ are again transformed by the weight matrices $\boldsymbol{W}_q \in \mathbb{R}^{n \times d_q}$, $\boldsymbol{W}_k \in \mathbb{R}^{n \times d_k}$ and $\boldsymbol{W}_v \in \mathbb{R}^{n \times d_v}$ into three distinct representations: the query representation $\boldsymbol{Q}(\boldsymbol{t}_i) = \boldsymbol{W}_q^\intercal \boldsymbol{t}_i$, the key representation $\boldsymbol{K}(\boldsymbol{t}_i) = \boldsymbol{W}_k^\intercal \boldsymbol{t}_i$ and the value representation $\boldsymbol{V}(\boldsymbol{t}_i) = \boldsymbol{W}_v^\intercal \boldsymbol{t}_i$, where $d_q = d_k = d_v = d$. The dot product of every query with all the key vectors produces an attention matrix whose elements determine the correlation between the different tiles of a WSI (4. in Figure 1).

The similarity mask is element-wise multiplied with the dot product of the query and key embeddings, generating a masked attention matrix whose non-zero elements reflect the contribution of a tile's neighbors to the tile score.

After obtaining the attention coefficients that correspond to the neighbors of every tile, the last step is to aggregate this contextual information to generate a single attention weight. For each tile, we sum the coefficients of their neighbors. The resultant tile score vector is passed through a softmax function to ensure all weights sum to one.

Therefore, the attention coefficient of the ith tile of a WSI is given by the following equation, where $\langle . \rangle$ denotes the inner product between two vectors:

$$w_i = \frac{\exp\left(\sum_{j=1}^{N} \langle \boldsymbol{Q}(\boldsymbol{t}_i), \boldsymbol{K}(\boldsymbol{t}_j) \rangle s_{ij}\right)}{\sum_{k=1}^{N} \exp\left(\sum_{j=1}^{n} \langle \boldsymbol{Q}(\boldsymbol{t}_k), \boldsymbol{K}(\boldsymbol{t}_j) \rangle s_{kj}\right)}. \tag{4}$$

The feature embeddings $\boldsymbol{t} \in \mathbb{R}^{n \times d_v}$ are then computed and weighted by their respective attention score to give a neighbor-constrained feature vector, $\boldsymbol{l}_i$, for each tile:

$$\boldsymbol{l}_i = w_i \boldsymbol{V}(\boldsymbol{t}_i) \tag{5}$$

3.5 FEATURE AGGREGATION AND SLIDE-LEVEL PREDICTION

The mechanism utilized to fuse the local and global value vectors allows for the adaptive blending of local and global information described in Equation 6. The sigmoid function applied to the local values serves as a weighting factor, enabling the model to emphasize local characteristics when they are deemed more relevant while still retaining the contribution from the global contexts.

$$\boldsymbol{m} = \sigma(\boldsymbol{l}) \odot \boldsymbol{l} + (1 - \sigma(\boldsymbol{l})) \odot \boldsymbol{t}, \tag{6}$$

where $\sigma()$ denotes the sigmoid non-linearity.

The collective, WSI-level representation $\boldsymbol{m} \in \mathbb{R}^{n \times d}$ is adaptively computed as the weighted average of the $\boldsymbol{z}$ fused vector:

$$\boldsymbol{z} = \sum_{i=1}^{N} a_i(\boldsymbol{m}_i)\,, \tag{7}$$

such that:

$$a_i = \frac{\exp \boldsymbol{w}^T(\tanh(\boldsymbol{V}\boldsymbol{t}_i^T) \odot \sigma(\boldsymbol{U}\boldsymbol{t}_i^T))}{\sum_{j=1}^{K} \exp \boldsymbol{w}^T(\tanh(\boldsymbol{V}\boldsymbol{t}_j^T) \odot \sigma(\boldsymbol{U}\boldsymbol{t}_j^T))}\,, \tag{8}$$

where $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{w}$ are learnable parameters, $\odot$ is an element-wise multiplication, and $\tanh()$ is the hyperbolic tangent function.

CAMIL achieves a synergistic effect by combining the value vector of the neighbor-constrained attention module with that of the transformer layer. The transformer layer captures global interactions and dependencies across the entire slide, while the neighbor-constrained attention module complements it by capturing local details and context. Together, they enable CAMIL to integrate both local and global perspectives effectively.

Finally, the slide-level prediction is given via the classification layer $\boldsymbol{W}_c \in \mathbb{R}^{c \times d}$:

$$\boldsymbol{y}_{\text{slide}} = \boldsymbol{W}_c \cdot \left(\sum_i \boldsymbol{z}_i \right)^T \tag{9}$$

where c corresponds to the number of classes and $\sum$ the sum pooling operation applied on z. The representation obtained from the high-attended patches is used to minimize a cross-entropy loss, and a final classification score is produced.

4 Experiments and Results

To demonstrate the performance of CAMIL in capturing informative contextual relationships and improving classification and localization, various experiments were performed on three histopathology datasets: CAMELYON16 (Ehteshami Bejnordi et al., 2017), CAMELYON17 (Bándi et al., 2019), and TCGA-NSCLC. Additional information about the datasets, including details about the training and test sets and our baseline models, can be found in the appendix.

4.1 Classification performance

We evaluated the performance of our context-aware pooling operator by comparing its performance other attention-based MIL models, including CLAM-SB, CLAM-MB (Lu et al., 2021), TransMIL (Shao et al., 2021), DTFD-MIL (Zhang et al., 2022), DSMIL Li et al. (2021a) and GTP (Zheng et al., 2022). CLAM-SB and CLAM-MB utilize an attention-based pooling operator within the Attention-Based MIL (AB-MIL) framework (Ilse et al., 2018). They focus on the features of individual tiles and incorporate a clustering layer to enhance performance further. TransMIL is a transformer-based aggregator operator, DTFD-MIL leverages class activation maps to estimate the probability of an instance being positive under the AB-MIL framework, and DSMIL uses dual instance and bag classifiers to refine predictions. Lastly, GTP combines a graph-based representation of a WSI and a vision transformer.

The results of using CAMIL to classify WSI in the CAMELYON16, TCGA-NSCLC, and CAMELYON17 datasets are presented in Table 1. The evaluation of the model's performance in all experiments includes the area under the receiver operating characteristic curve (AUC) and the slide-level

Table 1: Classification results on CAMELYON16, TCGA-NSCLC, and CAMELYON17

	CAMELYON16		TCGA-NSCLC		CAMELYON17	
Method	ACC(↑)	AUC(↑)	ACC(↑)	AUC(↑)	ACC(↑)	AUC(↑)
CLAM-SB	$0.877_{0.029}$	$0.933_{0.002}$	$0.903_{0.011}$	$0.972_{0.004}$	$0.802_{0.039}$	$0.849_{0.041}$
CLAM-MB	$0.894_{0.010}$	$0.938_{0.005}$	$0.904_{0.010}$	$0.973_{0.004}$	$0.803_{0.036}$	$0.858_{0.046}$
TransMIL	$0.905_{0.005}$	$0.950_{0.005}$	$0.905_{0.011}$	$0.974_{0.003}$	$0.804_{0.021}$	$0.873_{0.031}$
DTFD-MIL	$0.889_{0.007}$	$0.941_{0.005}$	$0.899_{0.010}$	$0.964_{0.003}$	$0.797_{0.029}$	$\mathbf{0.884}_{0.032}$
GTP	$0.883_{0.026}$	$0.921_{0.026}$	$\mathbf{0.916}_{0.015}$	$0.973_{0.006}$	$0.800_{0.037}$	$0.762_{0.108}$
DSMIL	$0.874_{0.066}$	$0.949_{0.006}$	$0.853_{0.031}$	$0.954_{0.015}$	$0.815_{0.031}$	$0.863_{0.043}$
CAMIL-L	$0.910_{0.010}$	$0.953_{0.002}$	$0.914_{0.011}$	$0.975_{0.004}$	$0.828_{0.027}$	$0.881_{0.031}$
CAMIL-G	$0.891_{0.001}$	$0.950_{0.009}$	$0.907_{0.012}$	$0.973_{0.004}$	$0.818_{0.039}$	$0.875_{0.036}$
CAMIL	$\mathbf{0.917}_{0.006}$	$\mathbf{0.959}_{0.001}$	$\mathbf{0.916}_{0.007}$	$\mathbf{0.975}_{0.003}$	$\mathbf{0.843}_{0.024}$	$0.881_{0.039}$

accuracy (ACC), which is determined by the threshold of 0.5. When implementing our baselines, we fine-tuned the hyperparameters used in the previously published original work to achieve the best performance. The GTP model has the same configurations as the one described in the original paper (Zheng et al., 2022). However, for the CAMELYON16 dataset, the batch size (k) was set to 2 due to memory limitations.

CAMIL outperforms other MIL models in ACC and AUC across CAMELYON16, TCGA-NSCLC, and CAMELYON17 datasets, narrowly trailing DTFD-MIL on CAMELYON17 by 0.003 in AUC. Its effectiveness notably identifies sparse cancerous regions in WSI, where tumor cells are often minimal, such as in the CAMELYON datasets, where tumor cells may account for as little as 5% of any WSI, which is particularly common in metastatic sites (Cheng et al., 2021). Specifically, CAMIL outperforms the other models on the CAMELYON16 dataset by significant margins, achieving at least 0.9% better in the AUC and 1.2% in ACC than the existing models on CAMELYON16 and 3.8% in ACC on CAMELYON17.

GTP, with its MinCUT pooling layer, which aims to reduce the complexity of self-attention, performs well on TCGA-NSCLC but less so on CAMELYON datasets. Reducing the complexity may make computation more manageable. However, on large and complex datasets such as CAMELYON16 and CAMELYON17, this pooling operation may result in loss of information, particularly in that of fine-grained details. TCGA-NSCLC is a smaller dataset, allowing the MinCUT pooling in GTP to retain sufficient information and remain competitive.

4.2 Localization

To evaluate the localization capability of CAMIL compared to our baselines, we examine both qualitative and quantitative evidence. Similar to the experimental design of Tourniaire et al. (2023b), we compute the Dice score to quantify the ability of the different models to identify cancerous evidence in cancerous slides. For normal slides, we compute the tile-level specificity. The reference ground-truth masks are computed at the 5th magnification level using expert tumor delineations. Additionally, a tile is considered cancerous if it contains at least 20% annotated tumor. To produce the predicted masks, we use the scaled attention scores for CAMIL, both CLAM models, TransMIL and DSMIL, the tile level logits for DTFD-MIL, and the GraphCAM for GTP. We apply a threshold of 0.5 to the model's output probabilities to generate the masks from the tile-level predictions. The results for the Dice score and Specificity are shown in Table 2

Table 2: Localization on CAMELYON16

Method	Dice(↑)	Specificity(↑)
CLAM-SB	$0.459_{0.037}$	$0.987_{0.008}$
CLAM-MB	$0.406_{0.007}$	$0.573_{0.045}$
TransMIL	$0.103_{0.004}$	$\mathbf{0.999}_{0.001}$
DTFD-MIL	$\mathbf{0.525}_{0.033}$	$\mathbf{0.999}_{0.001}$
GTP	$0.418_{0.068}$	$0.851_{0.116}$
DSMIL	$0.259_{0.083}$	$0.863_{0.043}$
CAMIL	$0.515_{0.058}$	$0.980_{0.040}$

CAMIL performs well, albeit slightly behind the DTFD-MIL model. We believe the decreased localization performance might be attributed to integrating the Nystromformer module in our model

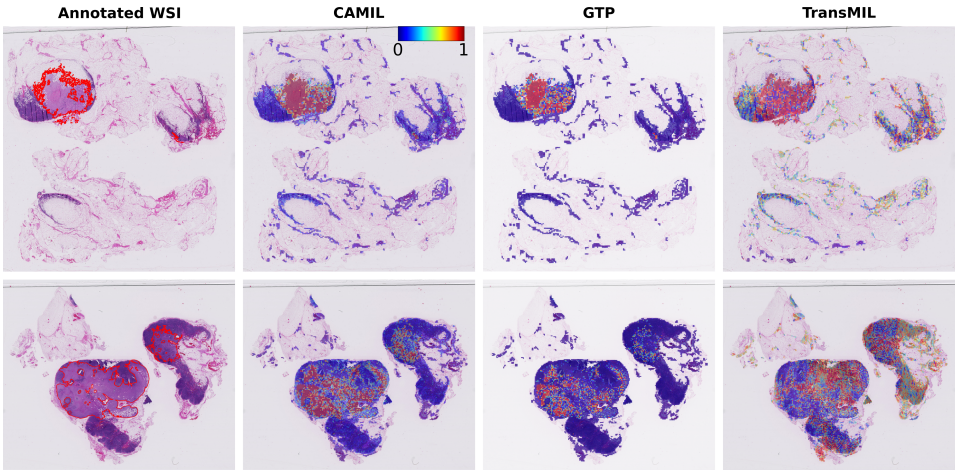

Figure 2: A visual example of the tumor regions and attention maps produced by different models on the CAMELYON16 cancer dataset. The left column shows the original whole slide image, including the pathologist ground truth annotations with tumor regions delineated by red lines. The other columns show attention maps from left to right for CAMIL, GTP, and TransMIL, pinpointing diagnostically significant locations on the CAMELYON16 cancer dataset.

design, akin to its role in TransMIL. TransMIL, as indicated by the attention maps in our qualitative assessment and its slide-level performance, demonstrates the ability to grasp the general patterns within a WSI and distinguish between normal and cancerous slides. However, despite this, confirmed by its low Dice score, it falls short in effectively pinpointing specific cancerous evidence within slides. Integrating the Nystromformer into our model design might introduce a trade-off between slide-level accuracy and localization performance, resulting in improved slide-level accuracy with an expense of slightly decreased localization performance.

Figure 2 provides a qualitative comparison of the attention maps generated by CAMIL, GTP, and TransMIL on the CAMELYON16 cancer dataset. We visually compare these methods as they leverage spatial information to enhance prediction. These attention maps underscore CAMIL's high localization performance, as it can discern the boundaries separating normal tissue from tumor tissue. Although effectively pinpointing the regions of interest, GTP attention maps appear fragmented and less dense in cancer-associated regions. This fragmentation could be attributed to the MinCUT-pooling operation, which may reduce the representation's granularity and affect the heatmap's coherence.

TransMIL appears to sufficiently capture long-term dependencies within the WSIs, as evidenced by the attention maps of Figure 2. Specifically, TransMIL can identify the presence of cancer and precisely pinpoint the cancer-associated regions. However, these maps also reveal TransMIL's inability to capture intricate details and local nuances. The attention scores are not only confined to the cancer regions but expand beyond those to the surrounding normal tissue, impeding the precise localization of tumor boundaries, indicating the model's limitations in representing close proximity relationships within the WSIs.

5 Ablation Studies

Additionally, we performed ablation studies to evaluate the effectiveness of the Nystromformer module and that of the neighboring-constrained attention module in our model. Specifically, we examined the effect of the Nystromformer block by retaining it while excluding the neighbor-constrained attention module denoted as CAMIL-G. Table 1 demonstrates that using only the Nystromformer block leads to satisfactory performance comparable to that of the TransMIL model, which also incorporates the Nystromformer. In a distinct ablation study, we omitted the Nystromformer block and retained the neighbor-constrained attention module, referred to as CAMIL-L. CAMIL-L exhibits a marginal improvement over the CAMIL-G model, thereby underlining its crucial role in augmenting the model's performance. Optimal results were achieved through the amalgamation of both models,

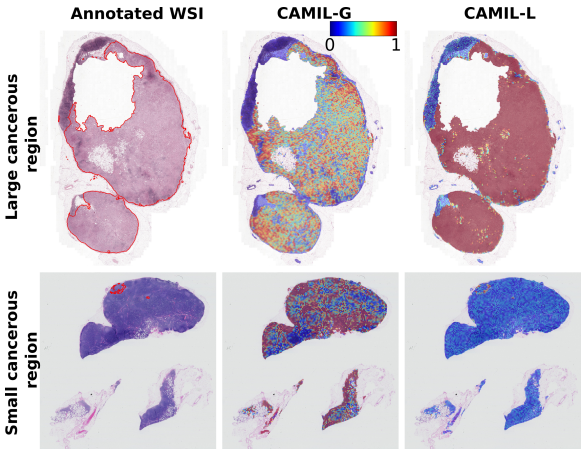

Figure 3: A visual example of the tumor regions and attention maps produced by different models on the CAMELYON16 cancer dataset. From left to right are the ground truth annotations with tumor regions delineated by red lines, the heatmaps for CAMIL-G, and the heatmaps for CAMIL-L.

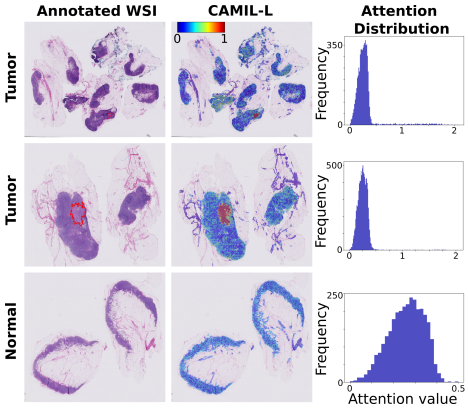

Figure 4: A visual example CAMIL-L fine-grained localization abilities. From left to right are the ground truth annotations with tumor regions delineated by red lines, generated heatmaps for CAMIL-LCAMIL-L, and the attention distribution of the neighbor-constrained attention module's weights (before applying the softmax function).

harnessing both the potential of the Nystromformer block to model long-term dependencies and the prowess of the neighbor-attention module to comprehend local visual concepts.

5.1 Visualising global and local concepts

We also visualized the attention maps generated by the two versions of our model. Notably, in Figure 3, CAMIL-G demonstrates commendable performance in understanding global concepts and overall patterns, as it effectively detects tumor regions. However, it struggles with highly localized tumors, encountering difficulties capturing intricate, short-term dependencies within the image.

On the other hand, the attention maps produced by CAMIL-L excel in capturing the fine details within a WSI. This proficiency can be attributed to the context-aware module, which utilizes a similarity mask. This mask aggregates attention weights of similar neighboring feature representations, resulting in robust activations. In contrast, less favorable weights, particularly those associated with negative regions, do not contribute as strongly. This observation is substantiated by the histograms of the unnormalized attention coefficients, which are not constrained within the range of 0 to 1 (before applying the softmax function) provided in Figure 4. These histograms underscore a notable pattern: while a significant portion of the attention coefficients falls below the 0.5 threshold, cancerous cases display outliers within the range of 0.5 to 2. These outliers represent stronger activations, which distinctly characterize regions affected by cancer. Conversely, all other activations consistently remain below the 0.5 threshold, signifying a reduced emphasis on non-cancerous areas.

6 Conclusion

We have introduced CAMIL, a novel MIL vision transformer-based method that considers the tumor microenvironment context while determining tile-level labels in WSIs, mirroring the approach of a skilled pathologist. This is achieved by employing a unique neighbor-constrained attention mechanism, which assesses the dependencies between tiles within a WSI and incorporates contextual constraints as prior knowledge into the MIL model. We have demonstrated that using the transformer and the neighborhood-attention mechanism together is imperative in successful performance across datasets through our ablation studies. Importantly, CAMIL achieves state-of-the-art across multiple datasets regarding tile-level ACC, AUC, and F1 scores and patch-level localization and interpretability.

REFERENCES

Mohammed Adnan, Shivam Kalra, and Hamid R. Tizhoosh. Representation learning of histopathology images using graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pp. 4254–4261. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPRW50498.2020.00502.

Aïcha BenTaieb and Ghassan Hamarneh. Predicting cancer with a recurrent visual attention model for histopathology images. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, volume 11071 of *Lecture Notes in Computer Science*, pp. 129–137. Springer, 2018. doi: 10.1007/978-3-030-00934-2_15.

Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. doi: 10.1109/TMI.2018.2867350.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen P. 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*, pp. 1–9, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Jun Cheng, Yuting Liu, Wei Huang, Wenhui Hong, Lingling Wang, Xiaohui Zhan, Zhi Han, Dong Ni, Kun Huang, and Jie Zhang. Computational Image Analysis Identifies Histopathological Image Features Associated With Somatic Mutations and Patient Survival in Gastric Adenocarcinoma. *Frontiers in Oncology*, 11, 2021. ISSN 2234-943X.

Pierre Courtiol, Eric W. Tramel, Marc Sanselme, and Gilles Wainrib. Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. *CoRR*, abs/1802.02212, 2018.

Kausik Das, Sailesh Conjeti, Abhijit Guha Roy, Jyotirmoy Chatterjee, and Debdoot Sheet. Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification. In *15th IEEE International Symposium on Biomedical Imaging, ISBI 2018, Washington, DC, USA, April 4-7, 2018*, pp. 578–581. IEEE, 2018. doi: 10.1109/ISBI.2018.8363642.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, December 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.14585. URL https://doi.org/10.1001/jama.2017.14585.

Michael Gadermayr, Lukas Koller, Maximilian Tschuchnig, Lea Maria Stangassinger, Christina Kreutzer, Sebastien Couillard-Despres, Gertie Janneke Oostingh, and Anton Hittmair. Mixup-mil: Novel data augmentation for multiple instance learning and a study on thyroid cancer diagnosis. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan,

Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 477–486, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43987-2.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Le Hou, Dimitris Samaras, Tahsin M. Kurç, Yi Gao, James E. Davis, and Joel H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2424–2433. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.266.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2132–2141, 2018.

Gabriel Landini, Giovanni Martinelli, and Filippo Piccinini. Colour deconvolution: stain unmixing in histological imaging. *Bioinformatics*, 09 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa847. btaa847.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2021a.

Jiayun Li, Wenyuan Li, Anthony E. 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. *Comput. Biol. Medicine*, 131:104253, 2021b. doi: 10.1016/j.compbiomed.2021.104253.

Pei Liu, Luping Ji, Xinyu Zhang, and Feng Ye. Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification, 2023.

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

Jaime Melendez, Bram van Ginneken, Pragnya Maduskar, Rick H. H. M. Philipsen, Klaus Reither, Marianne Breuninger, Ifedayo M. O. Adetifa, Rahmatulai Maane, Helen Ayles, and Clara I. Sánchez. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays. *IEEE Trans. Medical Imaging*, 34(1):179–192, 2015. doi: 10.1109/TMI.2014.2350539.

Sandra Morales, Kjersti Engan, and Valery Naranjo. Artificial intelligence in computational pathology - challenges and future directions. *Digit. Signal Process.*, 119:103196, 2021. doi: 10.1016/j.dsp.2021.103196.

Gwénolé Quellec, Mathieu Lamard, Michel Cozic, Gouenou Coatrieux, and Guy Cazuguel. Multiple-instance learning for anomaly detection in digital mammography. *IEEE Trans. Medical Imaging*, 35(7):1604–1614, 2016. doi: 10.1109/TMI.2016.2521442.

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 Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 2136–2147, 2021.

Zhuchen Shao, Liuxi Dai, Yifeng Wang, Haoqian Wang, and Yongbing Zhang. Augdiff: Diffusion based feature augmentation for multiple instance learning in whole slide image, 2023.

Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A. Moskaluk, Sana Syed, and Donald E. Brown. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In Mattias P. Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schlaefer, and Floris Ernst (eds.), *Medical Imaging with Deep Learning, 7-9 July 2021, Lübeck, Germany*, volume 143 of *Proceedings of Machine Learning Research*, pp. 682–698. PMLR, 2021.

Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep neural network models for computational histopathology: A survey. *CoRR*, abs/1912.12378, 2019.

David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(2):567–578, 2021. doi: 10.1109/TPAMI.2019.2936841.

Tong Tong, Robin Wolz, Qinquan Gao, Ricardo Guerrero, Joseph V. Hajnal, and Daniel Rueckert. Multiple instance learning for classification of dementia in brain MRI. *Medical Image Anal.*, 18 (5):808–818, 2014. doi: 10.1016/j.media.2014.04.006.

Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, and Hervé Delingette. MS-CLAM: mixed supervision for the classification and localization of tumors in whole slide images. *Medical Image Anal.*, 85:102763, 2023a. doi: 10.1016/j.media.2023.102763. URL https://doi.org/10.1016/j.media.2023.102763.

Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, and Hervé Delingette. Ms-clam: Mixed supervision for the classification and localization of tumors in whole slide images. *Medical Image Analysis*, 85:102763, 2023b. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102763. URL https://www.sciencedirect.com/science/article/pii/S1361841523000245.

Ming Tu, Jing Huang, Xiaodong He, and Bowen Zhou. Multiple instance learning with graph neural networks. *CoRR*, abs/1906.04881, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

Xiaodong Wang, Ying Chen, Yunshu Gao, Huiqing Zhang, Zehui Guan, Zhou Dong, Yuxuan Zheng, Jiarui Jiang, Haoqing Yang, Liming Wang, Xianming Huang, Lirong Ai, Wenlong Yu, Hongwei Li, Changsheng Dong, Zhou Zhou, Xiyang Liu, and Guanzhen Yu. Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. *Nature Communications*, 12(1):1637, March 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21674-7. Number: 1 Publisher: Nature Publishing Group.

Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognit.*, 74:15–24, 2018. doi: 10.1016/j.patcog.2017.08.026.

Chensu Xie, Hassan Muhammad, Chad M. Vanderbilt, Raul Caso, Dig Vijay Kumar Yarlagadda, Gabriele Campanella, and Thomas J. Fuchs. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning. In Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Hervé Lombaert, and Christopher Pal (eds.), *International Conference on Medical Imaging with Deep Learning, MIDL 2020, 6-8 July 2020, Montréal, QC, Canada*, volume 121 of *Proceedings of Machine Learning Research*, pp. 843–856, 2020.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 14138–14148. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17664.

Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. CAMEL: A weakly supervised learning framework for histopathology image segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 10681–10690, 2019. doi: 10.1109/ICCV.2019.01078.

Yan Xu, Jun-Yan Zhu, Eric I-Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Anal.*, 18(3):591–604, 2014. doi: 10.1016/j.media.2014.01.010.

Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas J. Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Anal.*, 65:101789, 2020. doi: 10.1016/j.media.2020.101789.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 18780–18790, 2022. doi: 10.1109/CVPR52688.2022.01824.

Jingwei Zhang, Ke Ma, John S. Van Arnam, Rajarsi Gupta, Joel H. Saltz, Maria Vakalopoulou, and Dimitris Samaras. A joint spatial and magnification based attention framework for large scale histopathology classification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pp. 3776–3784. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPRW53098.2021.00418.

Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern H. Menze, Xinjuan Fan, and Jianhua Yao. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 4836–4845. Computer Vision Foundation / IEEE. doi: 10.1109/CVPR42600.2020.00489.

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. *IEEE Trans. Medical Imaging*, 41(11):3003–3015, 2022. doi: 10.1109/TMI.2022.3176598. URL https://doi.org/10.1109/TMI.2022.3176598.

Changjiang Zhou, Yi Jin, Yuzong Chen, Shan Huang, Rengpeng Huang, Yuhong Wang, Youcai Zhao, Yao Chen, Lingchuan Guo, and Jun Liao. Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning. *Comput. Medical Imaging Graph.*, 88:101861, 2021. doi: 10.1016/j.compmedimag.2021.101861.

A Appendix

A.1 Reproducibility statement

All weakly-supervised deep learning models are trained using NVIDIA Tesla P100 GPUs. One GPU is used for training in each experiment. We intend to make the source code of our algorithm publicly available soon.

A.2 Performance metrics

Here, we provide ACC, F1, and AUC for both datasets. F1 scores were omitted from the main text to save space.

Table 3: Classification results on CAMELYON16

METHOD	ACC(↑)	F_1(↑)	AUC(↑)
CLAM-SB	$0.877_{0.029}$	$0.840_{0.023}$	$0.933_{0.002}$
CLAM-MB	$0.894_{0.010}$	$0.845_{0.019}$	$0.938_{0.005}$
TransMIL	$0.905_{0.005}$	$\mathbf{0.888}_{0.018}$	$0.950_{0.005}$
DTFD-MIL	$0.889_{0.007}$	$0.850_{0.004}$	$0.941_{0.005}$
GTP	$0.883_{0.026}$	$0.863_{0.026}$	$0.921_{0.026}$
DSMIL	$0.874_{0.066}$	$0.848_{0.056}$	$0.949_{0.006}$
CAMIL-L	$0.910_{0.010}$	$0.872_{0.016}$	$0.953_{0.002}$
CAMIL-G	$0.891_{0.001}$	$0.866_{0.012}$	$0.950_{0.009}$
CAMIL	$\mathbf{0.917}_{0.006}$	$0.881_{0.010}$	$\mathbf{0.959}_{0.001}$

Table 4: Classification results on TCGA-NSCLC

METHOD	ACC(↑)	F_1(↑)	AUC(↑)
CLAM-SB	$0.903_{0.011}$	$0.867_{0.012}$	$0.972_{0.004}$
CLAM-MB	$0.904_{0.010}$	$0.872_{0.057}$	$0.973_{0.004}$
TransMIL	$0.905_{0.011}$	$0.911_{0.005}$	$0.974_{0.003}$
DTFD-MIL	$0.899_{0.010}$	$0.899_{0.014}$	$0.964_{0.003}$
GTP	$\mathbf{0.916}_{0.015}$	$0.917_{0.016}$	$0.973_{0.006}$
DSMIL	$0.853_{0.031}$	$0.864_{0.024}$	$0.954_{0.015}$
CAMIL-L	$0.914_{0.011}$	$\mathbf{0.921}_{0.009}$	$0.975_{0.004}$
CAMIL-G	$0.907_{0.012}$	$0.918_{0.007}$	$0.973_{0.004}$
CAMIL	$\mathbf{0.916}_{0.007}$	$0.918_{0.005}$	$\mathbf{0.975}_{0.003}$

A.3 WSI PRE-PROCESSING

Using the publicly available WSI-prepossessing toolbox developed by (Lu et al., 2021), for both datasets, we first automatically segmented the tissue region from each slide and exhaustively divided it into 256×256 non-overlapping patches using ×20 magnification (Figure 1). Otsu's method was used to perform automatic WSI thresholding.

To avoid the computational overhead and capitalize on the rich feature representations already learned during its previous training on CAMELYON16, CAMELYON17, and TCGA-NSCLC datasets, we opted to use the pre-trained ResNet-18 feature extractor provided by (Li et al., 2021a). This model was extensively trained on a large set of tiles from the CAMELYON16 dataset, densely cropped without overlap, making it a powerful feature extractor. To make a fair comparison, we used the same contrastive learning-based model as the feature extractor for all our baselines.

A.4 DATASETS

CAMELYON16 is a significant publicly available Whole Slide Image (WSI) dataset for lymph node classification and metastasis detection. It includes 270 training and 129 test slides from two medical centers, all meticulously annotated by pathologists. Some slides have partial annotations, making it a challenging benchmark due to varying metastasis sizes.

The CAMELYON17 dataset consists of 1000 WSIs of a similar type to the CAMELYON16 dataset. However, only half (500 WSI) of these images are labeled and accessible publicly. These images are expertly annotated by pathologic lymph node classification into pN-stage:

- pN0: No micro-metastases or macro-metastases, or isolated tumor cells (ITCs) found.
- pN0(i+): Only ITCs found.
- pN1mi: Micro-metastases found, but no macro-metastases found.
- pN1: Metastases found in 1–3 lymph nodes, of which at least one is a macro-metastasis.
- pN2: Metastases found in 4–9 lymph nodes, of which at least one is a macro-metastasis.

Adopting the approach discussed in Tourniaire et al. (2023b), we created a binary classification problem by treating pN0 as normal and unifying all classes that were not pN0 into a single class, cancerous.

The TCGA-NSCLC dataset comprises two non-small cell lung cancer subtypes, LUAD and LUSC, with 541 slides. Unlike CAMELYON16, it lacks annotations.

A.5 DATA SPLITS

In the case of CAMELYON16, the WSIs are partitioned into a training and test set. The 270 WSIs of the training set are split five times into a training (80%) and a validation (20%) set in a 5-fold cross-validation fashion, and the average performance of the model on the competition test set is reported. The official test set comprising 129 WSIs is used for evaluation. For CAMELYON17, we used a 4-fold validation 65%, 15%, and 25% train, validation, and test splits. Regarding the TCGA-NSCLC dataset, a 5-fold cross-validation across the available images is performed. For each fold, the training set is split into 80% for training purposes and 20% for validation.

A.6 TRANSFORMER MODULE TO CAPTURE GLOBAL CONTEXTS

Working with large WSIs can lead to memory overload, as the self-attention mechanism used in the transformer layer requires computing pairwise interactions between all of the tiles in each WSI. To circumvent the memory overload associated with the long-range dependencies of large WSIs, we adopt the Nystromformer architecture (Xiong et al., 2021) to model feature interactions that otherwise would be intractable.

The Nystromformer approach is based on the Nystrom method, which is a technique for approximating a kernel matrix by selecting a small subset of its rows and columns. In the context of the transformer layer, this means selecting a subset of "landmark" tiles from each WSI to represent the full set of tiles. The landmark tiles are chosen randomly, and their embeddings are used to compute a low-rank approximation of the self-attention matrix. This approximation is then used instead of the full self-attention matrix to compute the final output of the transformer layer.

The Nystromformer architecture is highly scalable regarding sequence length, making it well-suited for processing large WSIs (Xiong et al., 2021). Additionally, by reducing the time complexity of the self-attention mechanism from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$, the Nystromformer approach can significantly reduce the computational cost of processing each WSI.

A.7 LOCALISATION

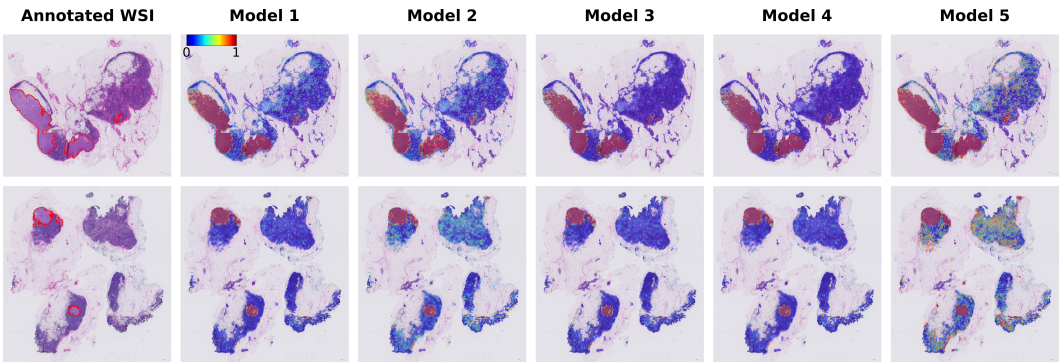

Figure 5: Displayed are attention maps produced on Whole Slide Images (WSIs) throughout the 5-fold cross-validation runs. The leftmost column shows the original WSIs, including ground truth annotations of tumor regions delineated by red lines, while the succeeding columns depict the predicted probabilities from the cross-validated model runs. The colormap signifies the likelihood with which a specific region within a WSI correlates to the target output label.

In addition to the quantitative outcomes, we assessed our approach qualitatively by visually presenting attention maps generated by different validation runs of our model overlaid on expert-annotated

Table 5: CAMIL using ResNet18 feature extractor pre-trained on ImageNet without SimCLR fine-tuning.

Dataset	ACC	AUC
CAMELYON16	$0.723_{0.095}$	$0.743_{0.077}$
TCGA-NSCLC	$0.692_{0.082}$	$0.798_{0.059}$

tumor regions (Figure 5). These attention maps pinpoint diagnostically significant locations in the image that are crucial for accurate tumor identification.

We notice a substantial agreement between the regions of interest identified by experts and those generated by our attention maps. These steps are harmonized with an attention map, forming a transformer relevancy map. Notably, our method consistently highlights the same regions within Whole Slide Images (WSIs) across different cross-validation folds, underscoring the reliability and robustness of our model.

A.8 SimCLR ablation study

SimCLR feature extraction backbone is an integral part of our model, and it plays a pivotal role in generating concise and descriptive feature representations. Leveraging the pre-trained SimCLR weights lays the foundation for generating meaningful similarity scores and, therefore, is crucial for the optimal performance of the neighbor-constrained attention mask similar to many similarity-based histopathology approaches. Incorporating ImageNet weights directly into our model notably decreases its performance (as it does with most models Tourniaire et al. (2023a); Li et al. (2021a)), emphasizing the necessity of SimCLR within our model architecture. Table 5 shows these results when using a ResNet-18 pre-trained on ImageNet on CAMELYON16 and TCGA-NSCLC.