# MG-Trans: Multi-Scale Graph Transformer With Information Bottleneck for Whole Slide Image Classification

Jiangbo Shi, Lufei Tang, Zeyu Gao, Yang Li, Chunbao Wang, Tieliang Gong, Chen Li, *Member, IEEE*, and Huazhu Fu, *Senior Member, IEEE*

***Abstract*— Multiple instance learning (MIL)-based methods have become the mainstream for processing the megapixel-sized whole slide image (WSI) with pyramid structure in the field of digital pathology. The current MIL-based methods usually crop a large number of patches from WSI at the highest magnification, resulting in a lot of redundancy in the input and feature space. Moreover, the spatial relations between patches can not be sufficiently modeled, which may weaken the model's discriminative ability on fine-grained features. To solve the above limitations, we propose a Multi-scale Graph Transformer (MG-Trans) with information bottleneck for whole slide image classification. MG-Trans is composed of three modules: patch anchoring module (PAM), dynamic structure information learning module (SILM), and multi-scale information bottleneck module (MIBM). Specifically, PAM utilizes the class attention map generated from the multi-head self-attention of vision Transformer to identify and sample the informative patches. SILM explicitly introduces the local tissue structure information into the Transformer block to sufficiently model the spatial relations between patches. MIBM effectively fuses the multi-scale patch features by utilizing the principle of information bottleneck to generate a robust and compact bag-level representation. Besides, we also propose a semantic consistency loss to stabilize the training of the whole model. Extensive studies on three subtyping datasets and seven gene mutation detection datasets demonstrate the superiority of MG-Trans.**

***Index Terms*— Whole slide image analysis, multiple instance learning, vision transformer, information bottleneck.**

## I. Introduction

PATHOLOGICAL slide examination is commonly considered as the gold standard for tumor diagnosis. With the fast development of digital slide scanning technology, traditional tissue specimens can be easily digitized into whole slide images (WSI). A WSI is usually stored in multiple magnifications and huge sizes. Fig. 1 shows an example of WSI, where the magnifications vary from 5× to 40×, and the size at 40× magnification over 10,000 × 10,000 pixels (0.25$\mu$m/pixel). The mutual examination of such a megapixel-sized WSI for pathologists is time-consuming, error-prone, and has high inter- and intra-observer variability. Recently, deep-learning methods [1] have been proposed to automate the WSIs diagnosis process, benefiting a wide range of pathological classification tasks, such as subtyping, staging, and grading [2]. The traditional deep-learning-based models are typically trained on regions of interest (ROIs), which refer to patches cropped from a WSI annotated by pathologists. However, these methods are only semiautomatic due to the reliance on the annotation of ROI from pathologists. Moreover, fine-grained region-level annotation instead of WSI-level is required for model training, which is challenging or even unreachable. To alleviate these problems, multiple instance learning (MIL)-based methods [3], [4], [5], [6], [7] have been proposed. The MIL-based methods only need WSI-level labels, which can automate the diagnosis process entirely and reduce the burden of data annotation.

MIL follows a division-aggregation paradigm and can process all the instances at the same time to generate the bag-level representation for classification. Specifically, the general MIL framework follows a three-step process: 1) cropping a series of patches, *i.e.*, instances, from a WSI at a fixed magnification; 2) extracting deep features of each patch (instance-level representation) individually through a pre-trained encoder such as ResNet50 [8]; 3) aggregating all the patch features to

Manuscript received 7 August 2023; accepted 5 September 2023. Date of publication 8 September 2023; date of current version 30 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62106191; in part by the Key Research and Development Program of Ningxia Hui Nationality Autonomous Region under Grant 2022BEG02025; in part by the Key Research and Development Program of Shaanxi Province under Grant 2021GXLH-Z-095; in part by the Project of China Knowledge Centre for Engineering Science and Technology; in part by the Innovation Team from the Ministry of Education under Grant IRT_17R86; in part by the National Research Foundation, Singapore, through the AI Singapore Program (AISG) under Award AISG2-TC-2021-003; and in part by the consulting research project of the Chinese Academy of Engineering (The Online and Offline Mixed Educational Service System for The Belt and Road Training in MOOC China). *(Corresponding author: Chen Li.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by The Cancer Genome Atlas (TCGA).

Jiangbo Shi, Lufei Tang, Zeyu Gao, Yang Li, Tieliang Gong, and Chen Li are with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: shijiangbo@stu.xjtu.edu.cn; tanglufeiluffy@163.com; betpotti@gmail.com; vigilee@stu.xjtu.edu.cn; gongtl@xjtu.edu.cn; cli@xjtu.edu.cn).

Chunbao Wang is with the Department of Pathology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi 710061, China (e-mail: binglizilliao2012@163.com).

Huazhu Fu is with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: hzfu@ieee.org).

Digital Object Identifier 10.1109/TMI.2023.3313252

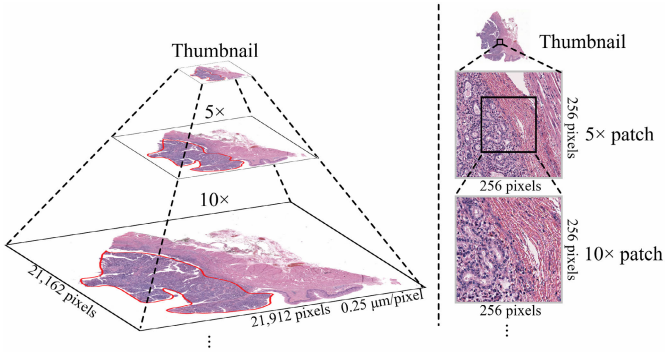



Fig. 1. Left: A whole slide image with a huge size and pyramid structure. The boundary of the cancer region is marked in red; Right: Fixed-size patches cropped from the WSI in different magnifications carry diagnostic information in different granularities.

obtain the bag-level representation for the classification task. Although these methods have achieved great success in many WSI analysis tasks like tumor subtyping [4], staging [6], survival analysis [9], etc., there are still three limitations that correspond to each step in the aforementioned three-step process.

First, a large amount of redundant and irrelevant input instances may hinder the model from focusing on the critical regions with high diagnostically discriminative.[1] Table I reports the average number of cropped patches from 2,839 WSIs in our dataset. We randomly select 100 WSIs and each WSI has around 30,000 patches at 40×, yet only 37.6% of them are within cancer regions as shown in Fig. 1. The instances outside of the cancer regions distract the attention of a MIL model from the diagnosis-discriminative instances, thus leading to the sub-optimal model.

Second, the spatial relations among instances are not sufficiently captured by the pre-trained encoder in existing methods. These spatial relations represent the tissue contextual information which is significant in many WSI classification tasks [10]. To fill this gap, a series of graph-based methods [11], [12] have been proposed to model local spatial relations between instances. They build the WSI graph by treating representations of instances as nodes and assigning edges according to the feature similarity or spatial proximity among nodes. However, the graph structure is fixed during the training process instead of being dynamically re-construction along with the optimization of the instance features. Alternatively, Transformer-based methods [7], [13], [14] directly model the complicated spatial relations among instances by the self-attention mechanism. However, they did not explicitly model the local spatial relations, which are more critical for distinguishing subtypes with subtle differences [15].

Third, the bag-level representation without any constraint may include redundant information, which can negatively affect the accuracy and generalization of the model. The need for a WSI classification model with improved generalization capabilities arises, due to the possibility of data distribution drift between different hospitals or slide scanning devices. Recently, there has been a consensus that reducing the redundancy information in feature space can mitigate the model's dependency on spurious correlation, thus enhancing the generalization of the classification model. Nevertheless, current research on aggregating instance-level representations to the bag-level representation in WSI classification models [3], [4], [5], [6], [11], [16] has not sufficiently taken into account this important aspect.

To solve the limitations of current MIL-based methods, we propose a Multi-scale Graph Transformer (MG-Trans) with Information Bottleneck (IB) for whole slide image classification. The core idea of MG-Trans is to select partial yet sufficient discriminative input patches to generate a robust and compact bag-level representation for the classification task. MG-Trans is mainly composed of three modules including a patch anchoring module (PAM), a dynamic structure information learning module (SILM), and a multi-scale information bottleneck module (MIBM).

The PAM takes the low-scale WSI as input and consists of a Transformer block and a cross-attention layer. The former is the Transformer block aims to generate a class attention map (CA-Map) for selecting a fixed and small number of informative patches. The latter is the cross-attention layer that enhances the interactions between the critical patches selected by CA-Map and all the output patches from the above Transformer block. Specifically, the selected patches serve as queries, while all output patches act as keys and values, which are input into the cross-attention layer. This cross-attention mechanism allows the model to focus on the most critical regions while retaining global contextual information from all the output patches. Consequently, PAM aids in the removal of redundant and irrelevant patches and enables the model to focus solely on highly diagnosis-discriminative regions.

The main model is a Transformer encoder, comprised of multiple standard Transformer blocks followed by several Transformer blocks integrating our proposed SILM. The main objective of SILM is to explicitly model the local tissue structure by constructing a tissue graph as its input. In this graph, the instance features obtained from the previous block and their nearest-neighbor relations are modeled as nodes and edges, respectively. Subsequently, SILM employs several graph convolutional network (GCN) layers with dense connections to extract multi-hop structure features that contain essential local tissue structure information. It is worth noting that the tissue graph is dynamically re-constructed at each training step, allowing it to sufficiently model the local spatial relations among instances based on the latest optimized features.

Recently, the IB principle has been proposed to improve the model's robustness by preserving the most relevant features for a given task [17]. The principle imposes constraints on two Mutual Information (MI) measurements within a model: one between the input and the feature representation, and the other between the feature representation and the output. Our MIBM compresses the bag-level representation by expanding the first MI measurement to the multi-scale scene. Specifically, the MI between each scale's input and representation is minimized. Besides, due to the sparsity of WSI-level supervision, we also propose a semantic consistency loss between the predictions

[1]The high diagnostically discriminative regions represent the area where pathologists make diagnostic conclusions during routine examinations.

of two-scale features to help stabilize the training of the whole model. Extensive experiments on three cancer subtyping datasets and seven gene mutation detection datasets indicate our model achieves new state-of-the-art results in accuracy and generalization. The main contributions of our work are as follows:

- We propose a multi-scale graph Transformer model with information bottleneck for whole slide image classification by reducing the redundancy in the input and feature space and sufficiently modeling the spatial relations between instances.
- We propose a patch anchoring module to project a large number of input patches to a small and fixed number of informational patches, which reduces the redundant information in the input patches.
- We propose a dynamic structure information learning module to explicitly integrate the local tissue structure information into the Transformer block.
- We propose a multi-scale information bottleneck module to generate a compact yet sufficient bag-level representation by effectively fusing the abundant multi-scale instance features.
- Extensive comparisons and ablation studies on three subtyping datasets and seven gene mutation detection datasets show that MG-Trans outperforms other state-of-the-art whole slide image classification methods.

## II. Related Work

This section introduces the MIL-based WSI analysis methods from three aspects: instance sampling strategy, instance relation modeling strategy, and instance feature aggregation strategy.

### A. Instance Sampling Strategy

There are mainly four kinds of instance sampling strategies: 1) Whole sampling. ABMIL [3] designs a gated attention network to learn an adaptive weight for all the patches. Then, based on ABMIL [3], CLAM [16] uses a pre-trained model for patch feature extraction and utilizes a global pooling operator to aggregate all the patch features linearly. Based on CLAM [16], many variations [4], [18], [19] have been proposed and have demonstrated the effectiveness of MIL in WSI analysis. However, these methods utilize all the input patches in WSI, which does not consider filtering out the uninformative ones. 2) Random sampling. To reduce the computation and memory cost, Yao et al. [20] proposed a random sampling method to select a subset of input patches but may ignore some discriminative patches. 3) Clustering-based sampling. C2C [21] divides patches into several clusters to help expose the model to diverse discriminative patches. Wang et al. [19] also utilized the K-means method to cluster the patches into different clusters. Although this kind of method can help the model select diverse patches, the sampled patches can not be promised to be informative. 4) Attention-based sampling. Recently, some work [22], [23] utilizes the attention mechanism to sample the patches. For example, BenTaieb and Hamarneh [23] utilized a recurrent visual attention model to select patches, but the training of the location model relies on pixel-level annotation. Utilizing the multi-head self-attention in vision Transformer (ViT) [24] to locate the informative regions [25], [26] has been widely explored in natural images. Inspired by these works, we design our patch anchoring module based on the class attention map generated from the Transformer block in ViT to effectively sample the most informative input patches. Note that, HIPT [27] also utilizes the multi-head self-attention maps to locate the multi-scale phenotypes. However, it directly presents the attention maps of different heads without additional processing. In this work, we propose a class attention map based on the multi-head self-attention map, which is more effective in helping the model remove redundant instances.

### B. Instance Relation Modeling Strategy

Modeling the spatial relation among patches is beneficial for the model to capture the tissue structure information and enlarge the model's receptive field. There are mainly two kinds of methods to model the spatial relation among instances: 1) Graph-based. The graph-based methods [11], [12] are commonly used to model the instance spatial relation. Specifically, different types of entities (*e.g.*, nuclei, patches, tissues) are selected as the graph nodes, and the spatial relations are modeled typically according to their Euclidean distance or feature similarity. However, the graph nodes and spatial relations between them will not change during the training process. Once the graph structure has been built, it remains static and can not be dynamically reconstructed. This can limit the model's ability to effectively depict the spatial relations between different nodes and accurately capture the complex tissue structure. 2) Transformer-based. Recently, Transformer-based methods [7], [19], [28] have been proposed to model the complicated spatial relations between patches. For example, TransMIL [7] proposes a correlated MIL to explore the spatial information of instances. The vision Transformer utilizes the multi-head self-attention mechanism to capture the global long-distance relations among instances. Each instance has the ability to perceive all other instances simultaneously. However, while the position embedding can help the model capture absolute position information, the local spatial relations of adjacent instances are lost, which is critical for the model to distinguish subtypes with subtle differences. In this work, we design the dynamic structure information learning module and integrate it into the Transformer block to model the local tissue structure in an explicit fashion.

### C. Instance Feature Aggregation Strategy

Naive MIL methods aggregate the instances features to generate the bag-level representation by utilizing the max or mean pooling. Then, the gated attention-based methods [3], [16] are proposed to learn a weight value for each instance and combine them linearly. However, these linear aggregation methods ignore reducing the redundant information in the bag-level representation. Recently, multi-scale patch features are utilized to improve the model's performance. For example, ZoomMIL [5] proposes a multi-level zooming method that directly sums up the multi-scale instance features. Although

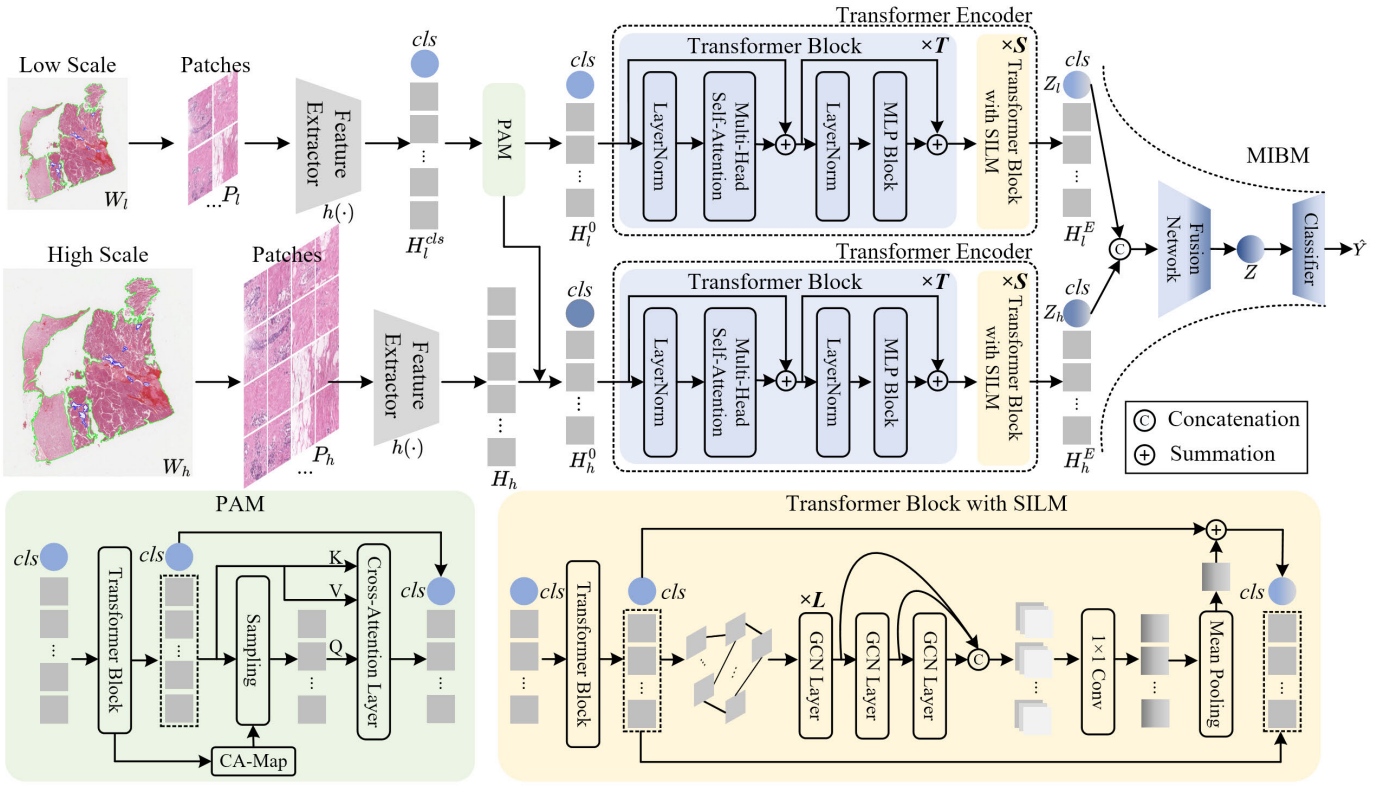



Fig. 2. MG-Trans. The input of our model is the low- and high-scale patches cropped from the original slide. The feature of each patch is extracted by a pre-trained feature extractor. PAM includes a Transformer block and a cross-attention layer to select the informative patches from a large number of redundant and irrelevant input patches. SILM is equipped with the last several Transformer blocks of a Transformer encoder and explicitly models the local tissue structure information. MIBM fuses the multi-scale patch features to generate a compact and sufficient bag-level representation for classification.

this approach can capture a rich set of features, it introduces more redundancy into the bag-level representation.

Lately, there has been a trend to improve the understanding of deep models using the ideas from information theory. The most notable one is the information bottleneck method [29], [30]. Let us denote the input random variable as $X$ and desired output variable as $Y$. The IB method extracts a compressed representation $Z$ from $X$ relevant for predicting $Y$. Formally, $Z$ is usually found by maximizing the following IB Lagrangian:

$$\mathcal{L}_{IB} = I(Y; Z) - \beta I(X; Z), \tag{1}$$

where $\beta > 0$ is the Lagrange multiplier that controls the trade-off between sufficiency (the performance on the task) and minimality (the complexity of the representation). $I(*; *)$ denotes the mutual information which is a metric that quantifies the correlation between two variables. Therefore, the IB method provides a natural approximation of the minimal sufficient statistic. The IB approach has been recently applied to many tasks like computer vision [31] and natural language processing [32]. In this work, we propose a multi-scale information bottleneck module to aggregate the redundant multi-scale patch features to generate a compressed yet robust bag-level representation.

## III. Methodology

### A. Overview

In this section, we describe the details of our MG-Trans model. As shown in Fig. 2, the proposed model mainly includes three modules: patch anchoring module (PAM), dynamic structure information learning module (SILM), and multi-scale information bottleneck module (MIBM). The following subsections will describe these three modules thoroughly.

### B. Problem Formulation

In MIL, a WSI $W = \{W_l, W_h\}$ with pyramid structure[2] is taken as a bag containing multiple instances as $P = \{P_l \in \mathbb{R}^{N_l \times N_0 \times N_0 \times 3}, P_h \in \mathbb{R}^{N_h \times N_0 \times N_0 \times 3}\} \in \mathbb{R}^{(N_l+N_h) \times N_0 \times N_0 \times 3}$. $N_l$ and $N_h$ denote the numbers of the low and high-scale of patches. $N_0$ is the patch size. $W_l$ and $W_h$ represent the slides at low and high magnification. $P_l$ and $P_h$ denote the corresponding patch sets to $W_l$ and $W_h$. Following the current embedding-based MIL methods [16], we utilize the non-overlapping sliding window method to crop the patches $P$ from the WSI. Then, a feature extractor $h(\cdot)$ maps the patches $P$ into a feature vector $H = \{H_l \in \mathbb{R}^{N_l \times d}, H_h \in \mathbb{R}^{N_h \times d}\} \in \mathbb{R}^{(N_l+N_h) \times d}$, where $d$ is the dimension of the patch feature. To obtain the bag-level representation $Z$, the feature vector $H$ is aggregated by a pooling function $p(\cdot)$. Then the bag-level prediction result $Y'$ is acquired by passing $Z$ into a classifier $f(\cdot)$:

$$Y' = f\left(p\left(\{h(P_l^1), \ldots, h(P_l^{N_l}), h(P_h^1), \ldots, h(P_h^{N_h})\}\right)\right). \tag{2}$$

The patches (*i.e.*, $P_l$ and $P_h$) describe the same regions in different scales. To identify the corresponding spatial relations

[2]We take two resolutions as an example to illustrate our method.

between multi-scale patches, we define an alignment matrix $A \in \{0,1\}^{N_l \times N_h}$ to align the patches from two scales. Specifically, $A_{i,j} = 1$ if the $j$-th patch $P_h^j$ in $P_h$ subordinates to the $i$-th patch $P_l^i$ in $P_l$, otherwise $A_{i,j} = 0$.

### C. Patch Anchoring Module

To filter out massive redundant and irrelevant patches in the input space, we design a patch anchoring module (PAM), which generates the class attention map (CA-Map) from the multi-head self-attention in the Transformer block. Recently, some work [25] in the field of natural images has proven the ability to utilize the multi-head self-attention to locate informative regions. The core idea of PAM is to project a large number of input patches with redundancy into a fixed and smaller amount of patches by selecting the most informative ones.

Specifically, the patch features $H_l$ are fed into a linear layer to reduce the feature dimension from $d$ to $d'$ first. Then, following the setting of ViT [24], we add a learnable class token $x_l^{cls} \in \mathbb{R}^{1\times d'}$ into the patch features $H_l$ to generate the features $H_l^{cls} = \{x_l^{cls} || H_l\} \in \mathbb{R}^{(N_l+1)\times d'}$, where $||$ denotes the concatenation operation. Then, $H_l^{cls}$ is fed into a Transformer block to generate the CA-Map. Suppose the Transformer block has $M$ heads, $Q$ and $K$ are $d'/M$-dimensional query vectors and key vectors of all tokens, then the self-attention weight for each head can be calculated by:

$$Att_l = \text{softmax}(\frac{QK^T}{\sqrt{d'/M}}). \tag{3}$$

where $Att_l \in \mathbb{R}^{(N_l+1)\times(N_l+1)}$, the output features of the Transformer block are denoted as $H_l^{cls'}$ and the class token is updated as $x_l^{cls'}$. As shown in Fig. 3, for the $m$-th head, the class attention map $C_m \in \mathbb{R}^{1\times N_l}$ between the class token and all the other patch tokens can be extracted from $Att_l$. Then, the CA-Map can be generated by averaging the $C_m$ of each head as $C = \sum_{m=1}^{M} C_m/M$. Based on the CA-Map $C \in \mathbb{R}^{1\times N_l}$, $N_l'$ patches with top highest attention values are selected from $H_l^{cls'}$.

Furthermore, we introduce a cross-attention layer to enhance the interactions between the selected local high-response patches and all the input patches. Specifically, we take the $N_l'$ top highest response patches as query $Q^l$ vectors, all the input patch tokens in $H_l^{cls'}$ as key $K^g$ and value $V^g$ vectors, the cross attention between the local query and the global key-value pairs as below:

$$H_l^{local} = \text{softmax}(\frac{Q^l {K^g}^T}{\sqrt{d'}})V^g. \tag{4}$$

To introduce more fine-grained information from the high-scale patches, we also select corresponding high-response patch features from $H_h$. Specifically, based on the alignment matrix $A$ and the selected $N_l'$ top high-response patch features in $H_l^{local} \in \mathbb{R}^{N_l'\times d'}$, the patch features $H_h^{local} \in \mathbb{R}^{N_h'\times d}$ from high-scale can be sampled from $H_h$ ($N_h' \ll N_h$).

For the selected patch features $H_h^{local}$ from the high scale, we also first feed it into a linear layer to reduce its dimension from $d$ to $d'$. Then, we concatenate a learnable class token $x_h^{cls} \in \mathbb{R}^{1\times d'}$ to it. Following the setting of ViT [24], to introduce the absolute position information, learnable position embeddings $E_p^l \in \mathbb{R}^{(N_l'+1)\times d'}$ and $E_p^h \in \mathbb{R}^{(N_h'+1)\times d'}$ are added to the patch features $H_l^{local}$ and $H_h^{local}$, respectively. At this point, we acquire the input features in different scales for the Transformer encoder $E(\cdot)$ as:

$$H_l^0 = [x_l^{cls'} || H_l^{local}] + E_p^l, \tag{5}$$

$$H_h^0 = [x_h^{cls} || H_h^{local}] + E_p^h. \tag{6}$$

The Transformer encoder $E(\cdot)$ is composed of $T+S$ Transformer blocks, and the last $S$ blocks are equipped with our dynamic structure information learning module. Each Transformer block includes a multi-head self-attention (MHSA) layer and a multi-layer perceptron (MLP) layer with two fully connected layers. The output of the $t$-th Transformer block for the low-scale input is calculated as follows:

$$H_l^{t'} = \text{MHSA}(\text{LN}(H_l^{(t-1)})) + H_l^{(t-1)}, \tag{7}$$

$$H_l^t = \text{MLP}(\text{LN}(H_l^{t'})) + H_l^{t'}, \tag{8}$$

where $\text{LN}(\cdot)$ denotes the layer normalization. After passing through the first $T$ Transformer blocks, the input patch feature $H_l^0$ is updated as $H_l^T \in \mathbb{R}^{(N_l'+1)\times d'}$.

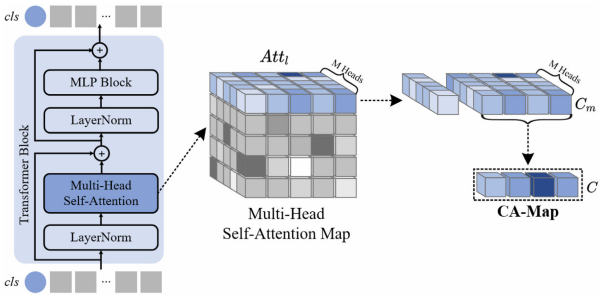



Fig. 3. The calculation of CA-Map. The multi-head self-attention map is first extracted from the multi-head self-attention layer in the Transformer block, then the attention map for each head between the class token and all the other patch tokens is selected. Total $M$ attention maps are merged by a mean pooling operator to generate the CA-Map.

### D. Dynamic Structure Information Learning Module

The Transformer block captures the complicated spatial relation among patches by leveraging the self-attention mechanism to obtain a global receptive field. However, it does not explicitly model the local tissue structures, which is essential for accurately identifying subtypes with subtle differences. To address this limitation, we propose the dynamic structure information learning module (SILM) that explicitly integrates the local tissue structure information into the Transformer block.

The input of the first SILM for the low scale is the feature vector $H_l^{T'}$, which is obtained by passing $H_l^T$ through a Transformer block. To explicitly model the local tissue structure, the first step is to construct a tissue graph that represents the local spatial relations among patches. Specifically, all the patch tokens in $H_l^{T'}$ are chosen as nodes in the tissue graph. Then, the node feature matrix is denoted as

$V_l \in \mathbb{R}^{N_l' \times d'}$. We utilize the nearest neighbors (NN) algorithm [33] to build the topological structures between nodes and use the Euclidean distances between patch centroids to quantify distances between patches. Formally, for each node pair $(v, u)$, edge $e_{vu}$ is built if $u \in \{w|\text{dist}(v, w) \leq d_{edge}, \forall w, v \in V_l\}$. The tissue graph topology is represented by a binary adjacency matrix $E_l \in \{0, 1\}^{N_l' \times N_l'}$. Then, the tissue graph is formulated as $TG_l = \{V_l, E_l\}$. Note that the structure of the tissue graph is dynamic along with the training process, as the PAM module selects different patches at each step. To extract the local structure information, the tissue graph is passed through $L$ GCN layers with dense connections. The $L$ different-hop structure features are concatenated and passed through a $1 \times 1$ convolution layer to align the feature dimension to $d'$. To integrate the local structure information into the Transformer block, the mean pooling operator is utilized to combine all the updated patch features. The resulting feature is then added to the class token. Specifically, the message passing of the $l$-th GCN layer is formulated as:

$$h_{ij}^l = \text{ReLU}(v_i^l + v_j^l),\ j \in \mathcal{N}(i), \tag{9}$$

$$v_i^{l+1} = \sum_{j \in \mathcal{N}(i)} \frac{\exp(h_{ij}^l)}{\sum_{j' \in \mathcal{N}(i)} \exp(h_{ij'}^l)} \cdot h_{ij}^l, \tag{10}$$

where $v_*^l \in \mathbb{R}^{d'}$ for $l \in [1, L]$ denote the input node features of the $l$-th GCN layer ($v_*^0 \in V_l$), $\mathcal{N}(i)$ is the neighboring nodes of the $i$-th node. $h_{ij}^j \in \mathbb{R}^{d'}$ denotes a summation of features of node $i$ and $j$ in $l$-th layer. Similar to the attention mechanism [3], $v_i^{l+1}$ is the fused feature by considering the importance of different neighboring nodes. After learning by the last $S$ Transformer blocks equipped with SILM, the feature vector $H_l^T$ from the low-scale is updated as $H_l^E \in \mathbb{R}^{(N_l'+1) \times d'}$.

The same with low-scale feature $H_l^0$, the high-scale feature $H_h^0$ is also passed through the same Transformer encoder $E(\cdot)$, then the updated high-scale feature vector is denoted as $H_h^E \in \mathbb{R}^{(N_h'+1) \times d'}$.

### E. Multi-Scale Information Bottleneck Module

After learned by the Transformer encoder $E(\cdot)$, two class tokens denoted as $Z_l$ and $Z_h$ are selected from $H_l^E$ and $H_h^E$ to generate the bag-level representation $Z$ for final classification. The bag-level representation $Z$ is usually obtained by summing or concatenating $Z_l$ and $Z_h$. However, such linear aggregation operations neglect to remove the redundant information in the bag-level representation that may be detrimental to the model predictions. To minimize the redundancy in the bag-level representation, we propose the multi-scale information bottleneck module (MIBM) to obtain a compact and sufficient bag-level representation $Z$ for classification. Specifically, the feature $Z_l$ and $Z_h$ are concatenated together first and then feed forward a fusion network to generate the bag-level representation $Z = f_\theta(Z_l||Z_h)$. The fusion network is a fully connected layer to scale the dimension of $Z$ and $Z_l$ to be consistent. Then, the bag-level prediction $\hat{Y}$ is obtained by feeding $Z$ into a classifier $f_c(\cdot)$. The overall objective of MIBM is formulated as follows:

$$\underset{\phi,\theta,\omega}{\text{argmax}}(I(Y; Z) - \beta_l I(X_l; Z_l) - \beta_h I(X_h; Z_h)), \tag{11}$$

where $\beta_l$ and $\beta_h$ are the regularization parameters of the low and high-scale features. $\phi, \theta$, and $\omega$ are the parameters of the Transformer encoder $E(\cdot)$, the fusion network $f_\theta(\cdot)$, and the classifier $f_c(\cdot)$, respectively. $X_l$ and $X_h$ denote the input of the Transformer encoder $E(\cdot)$ (*i.e.*, $H_l^0$ and $H_h^0$). For the maximization of $I(Y; Z)$, it can be calculated by the risk associated with $Z$ to the prediction performance on ground truth $Y$ based on the cross-entropy loss,

$$\mathcal{L}_{CE} = \text{CE}(Y, \hat{Y}). \tag{12}$$

Therefore, Eq (11) can be rewritten as:

$$\underset{\phi,\theta,\omega}{\text{argmin}}(\mathcal{L}_{CE} + \beta_l I(X_l; Z_l) + \beta_h I(X_h; Z_h)). \tag{13}$$

There are many deep IB methods [34], [35] to estimate $I(X_*; Z_*)$, such as variational approximation [32] and mutual information neural estimator [36]. However, these kinds of methods require an additional auxiliary network to approximate a lower bound of mutual information values. Inspired by [37], we utilize the matrix-based Rényi's $\alpha$-order entropy functional to estimate information-theoretic quantities directed on the reproducing kernel Hilbert space formed by the projected samples. Specifically, given $N$ pairs of samples $\{X_l^m, Z_l^m\}_{m=1}^N$ from a mini-batch in the low-scale feature. $X_l^m$ and $Z_l^m$ are both viewed as random vectors. According to [37], the entropy of $X_l$ can be defined over the eigenspectrum of a Gram matrix $K_{X_l} \in \mathbb{R}^{N_l' \times N_l'}$ ($K_{X_l}(m, n) = k(X_l^m, X_l^n)$ and $k$ is a Gaussian kernel) as:

$$\begin{aligned} H_\alpha(A_{X_l}) &= \frac{1}{1-\alpha}\log_2((\text{tr}(A_{X_l}^\alpha))) \\ &= \frac{1}{1-\alpha}\log_2(\sum_{m=1}^{N} \lambda_m(A_{X_l})^\alpha), \end{aligned} \tag{14}$$

where $\alpha \in (0, 1) \cup (1, \infty)$. We set order $\alpha$=1.01 to approximate the Shannon entropy. $A_{X_l}$ is the normalized version of $K_{X_l}$ (*i.e.*, $A_{X_l} = K_{X_l}/\text{tr}(K_{X_l})$). $\lambda_m(A_{X_l})$ denotes the $m$-th eigenvalue of $A_{X_l}$. The entropy of $Z_l$ can be measured on the eigenspectrum of another normalized Gram matrix $A_{Z_l}$. Then the joint entropy for $X_l$ and $Z_l$ can be calculated as:

$$H_\alpha(A_{X_l}, A_{Z_l}) = H_\alpha(\frac{A_{X_l} \circ A_{Z_l}}{\text{tr}(A_{X_l} \circ A_{Z_l})}), \tag{15}$$

where $\circ$ denotes element-wise product. Finally, the matrix-based Rényi's $\alpha$-order mutual information $I_\alpha(X_l; Z_l)$ in an analogy of Shannon's mutual information is defined as:

$$I_\alpha(X_l; Z_l) = H_\alpha(A_{X_l}) + H_\alpha(A_{Z_l}) - H_\alpha(A_{X_l}, A_{Z_l}). \tag{16}$$

Note that $I(X_h; Z_h)$ can be calculate as the same way with $I(X_l; Z_l)$.

TABLE I
DATA STATISTICS

| Task | | Cancer Subtyping | | | Gene Mutation Detection | | |
|---|---|---|---|---|---|---|---|
| Cancer Type | | Breast | | Kidney | Breast | | Gastric |
| Dataset | | BRIGHT | TCGA-BRCA | TCGA-RCC | MAP3K1 / GATA3 / TP53 PIK3CA / CDH1 | KMT2C | MSI |
| Number of WSI | | 501 | 877 | 640 | 559 | 559 | 263 |
| Dataset Split | Training | 301 | 525 | 256 | 335 | 337 | 157 |
| | Validation | 100 | 176 | 192 | 112 | 111 | 53 |
| | Test | 100 | 176 | 192 | 112 | 111 | 53 |
| Number of Patches | 5× | 325,768 | 555,304 | 611,858 | 393,220 | 393,220 | 219,452 |
| | 10× | 1,235,527 | 2,122,570 | 2,375,351 | 1,506,230 | 1,506,230 | 839,347 |

### F. Training Strategy

A major challenge in applying MIL for WSI analysis is that the training process utilizes vast and multi-scale input patches but is only supervised by a weak bag-level label, which may decrease the training stability. We propose a semantic consistency loss $\mathcal{L}_{SC}$ to help the model learn more semantic consistency information between two-scale features. The low-scale cls token $Z_l$ and high-scale cls token $Z_h$ typically contain more tissue structure information and details, respectively. These two cls tokens should have the same prediction category when they are classified separately. The main idea of $\mathcal{L}_{SC}$ is to minimize the semantic discrepancy between the two-scale prediction results. Specifically, two class tokens $Z_l$ and $Z_h$ are separately fed into a multi-layer perception (MLP) layer to get the prediction logits $p_l$ and $p_h$. Their difference is denoted as $\Delta_p = |p_l - p_h|_1$. Then, the semantic prediction $\Delta_Y$ depicting the semantic disagreement between two scales can be acquired through $\Delta_p$. This semantic disagreement should be minimized by:

$$\mathcal{L}_{SC} = \mathrm{CE}(Y, \Delta_Y). \tag{17}$$

Finally, the model is trained end-to-end, and the overall training loss is as:

$$\mathcal{L} = \mathcal{L}_{CE} + \beta_l I(X_l; Z_l) + \beta_h I(X_h; Z_h) + \beta_{sc}\mathcal{L}_{SC}, \tag{18}$$

where $\beta_{SC}$ is the regularization parameter of the semantic consistency loss.

## IV. EXPERIMENTS AND RESULTS

### A. Datasets

We evaluate our MG-Trans on two WSI classification tasks: cancer subtyping and gene mutation detection. The cancer subtyping task includes three datasets (*i.e.*, BRIGHT, TCGA-BRCA and TCGA-RCC). The gene mutation detection task includes two kinds of cancers (*i.e.*, breast and gastric) with seven types of mutated genes. The data statistics are reported in Table I. The specific description for each dataset is as follows:

- **BRIGHT**[3]: There are two kinds of data catalog methods for all the breast slides in the BRIGHT dataset. The coarse-grained method catalogs the dataset into three subtypes (*i.e., non-cancerous, precancerous, and cancerous*), and the fine-grained method includes six subtypes (*i.e., pathological benign (PB), usual ductal hyperplasia (UDH), flat epithelial atypia (FEA), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), and invasive carcinoma (IC)*). Note that we denote the coarse-grained catalog method as BRIGHT3 and the fine-grained method as BRIGHT6. All the slides were acquired at the Fondazione G.Pascale, Italy, and scanned by an Aperio AT2 scanner at 40×. The data split ratio for training, validation, and test dataset is 6:2:2.
- **TCGA-BRCA**[4]: Total 877 breast slides are collected from the TCGA-BRCA project, which includes two subtypes (*i.e., invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC)*). The data split ratio for training, validation, and test dataset is 6:2:2.
- **TCGA-RCC**: A total of 640 renal cell carcinoma (RCC) slides are collected from the TCGA-RCC project, including three subtypes (*i.e., clear cell renal cell carcinoma (CCRCC), chromophore renal cell carcinoma (CRCC), and papillary renal cell carcinoma (PRCC)*). The data split ratio for training, validation, and test dataset is 4:3:3.
- **Gene Mutation Detection**: The slides including at least one of the top six most prevalently mutated genes (*i.e., MAP3K1, GATA3, TP53, PIK3CA, CDH1, and KMT2C*) are collected from the TCGA-BRCA project. The slides that have the microsatellite instability (*MSI*) mutation are also collected from the TCGA-STAD project. For each gene, all the slides are split into two categories (*i.e.*, normal or having this kind of gene mutation), and the data split ratio for training, validation, and test dataset is 6:2:2.

Totally, 2,839 slides are processed into small patches with the size of 256 × 256 pixels in two scales (*i.e.*, 5× and 10×) separately for the experiments.

### B. Implementation Details

The original WSI is processed by Otsu's binarization algorithm first to filter out the blank background. The stain of all the patches extracted from the slides is normalized by the z-score formulation. The cropped patch size $N_0$ is 256. $W_l$ and $W_h$ represent the slides at 5× and 10× magnifications. The patch feature extractor $h(\cdot)$ is a truncated ResNet50

[3]https://research.ibm.com/haifa/Workshops/BRIGHT/

[4]TCGA is a public cancer data consortium that contains diagnostic WSIs and corresponding pathological reports.

TABLE II

RESULT (PRESENT IN %) ON BRIGHT, TCGA-BRCA, AND TCGA-RCC SUBTYPING DATASETS. ± REPRESENTS MEAN ± STANDARD DEVIATION. THE TOP BEST RESULTS ARE MARKED IN BOLD, AND THE COMPARABLE PERFORMANCES ARE DENOTED BY SUPERSCRIPT * BASED ON THE PAIRED T-TEST (P-VALUE>0.05)

| Dataset | BRIGHT3 | | | BRIGHT6 | | | TCGA-BRCA | | | TCGA-RCC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metric | AUC | F1 | ACC | AUC | F1 | ACC | AUC | F1 | ACC | AUC | F1 | ACC |
| MIL-based | | | | | | | | | | | | |
| ABMIL | 77.8±3.0 | 47.6±3.3 | 64.8±4.6 | 68.2±2.4 | 21.0±2.3 | 44.0±2.2 | 80.6±4.2 | 67.8±8.3 | 81.7±3.2 | 93.8±1.2 | 83.4±2.5 | 85.8±1.7 |
| CLAM | 78.2±3.9 | 53.2±5.4 | 67.4±3.9 | 72.1±4.1 | 35.2±4.3 | 46.6±5.9 | 85.6±1.8 | 75.7±2.6 | 83.8±1.7 | 96.1±0.7 | 88.1±1.8 | 87.3±1.7 |
| DTMIL | 79.2±5.0 | 55.3±5.9 | 70.2±3.5 | 72.4±2.6 | 27.7±2.6 | 46.4±1.9 | 86.2±1.7 | 77.2±1.7 | 84.9±1.6 | 97.9±1.1 | 87.3±1.7 | 88.0±1.4 |
| Multi-scale-based | | | | | | | | | | | | |
| DSMIL | 79.4±3.0 | 57.6±4.4 | 69.8±4.0 | 73.0±4.1 | 26.1±4.2 | 45.8±1.6 | 86.5±3.0 | 76.7±4.9 | 83.2±3.0 | 97.4±0.4 | 88.5±1.5 | 89.7±1.2 |
| ZoomMIL | 82.8±2.6 | 59.5±8.1 | 71.0±4.1 | 72.4±4.2 | 30.8±5.4 | 45.8±4.5 | 87.7±3.9 | 76.4±4.2 | 86.1±1.8 | 98.1±0.4 | 87.8±1.7 | 89.6±1.7 |
| $H^2$MIL | 84.6±4.8 | 60.2±9.4 | 71.6±6.6 | 74.8±3.4 | 35.9±5.8 | 47.0±4.8 | 89.2±3.2 | 78.7±3.1 | 86.9±2.8 | 98.5±0.6* | 89.1±2.1 | 91.5±1.9 |
| Transformer-based | | | | | | | | | | | | |
| TransMIL | 77.1±4.7 | 60.9±6.4 | 70.2±3.1 | 69.4±2.7 | 26.3±2.5 | 46.4±2.8 | 86.1±4.5 | 75.6±4.9 | 85.0±2.2 | 98.1±0.7 | 89.7±1.0 | 91.4±1.1 |
| GT | 84.2±3.6* | 62.0±4.3 | 70.6±3.9 | **77.7±3.6** | 37.2±5.6 | 51.8±3.5 | 90.7±1.7 | 79.6±2.9 | 87.5±2.5* | 98.7±0.5* | 90.8±2.6 | 92.1±2.3* |
| **MG-Trans** | **84.8±3.6** | **66.3±5.8** | **74.6±3.8** | 77.1±2.9* | **38.5±4.8** | **52.4±3.4*** | **92.1±1.6** | **81.3±1.8** | **88.0±1.4** | **99.0±0.3** | **91.8±1.5** | **92.9±1.2** |

model [8] pre-trained on ImageNet [38] with a 1024 output dimension ($d = 1024$). The number of the sampled patches in the low-scale $N_l'$ is 512. The reduced feature dimension $d'$ is 192. The Transformer encoder $E(\cdot)$ includes $T = 1$ Transformer block and $S = 1$ deformed Transformer block with SILM. The head number $M$ in the Transformer block of the Transformer encoder $E(\cdot)$ is 8. In SILM, the GCN model has $L = 3$ layers, and $d_{edge}$ is 256. In MIBM, the fusion network $f_\theta(\cdot)$ is a fully connected layer. The classifier $f_c(\cdot)$ is also a fully connected layer. We adopt Adam optimization [39] with a learning rate of $1 \times 10^{-4}$ and weight decay of $1 \times 10^{-5}$. The regularization parameter $\beta_l$, $\beta_h$, and $\beta_{sc}$ are all set as $1 \times 10^{-4}$. Note that all the hyperparameters with the highest performance are selected on the BRIGHT3 dataset and applied to all other datasets in the experiment. The minimum training epoch number is 80, and we adopt the early stop strategy if the accuracy does not continuously increase for 20 epochs. The batch size is 1. MG-Trans and all the baselines are implemented with the PyTorch and PyG library [40] on a workstation with eight NVIDIA 2080Ti GPUs. Note that MG-Trans is just trained on one GPU. The code of our framework is available at https://github.com/Jiangbo-Shi/MG-Trans.

## C. Evaluation Metrics

The area under the curve (AUC) score, f1 score (F1), and accuracy (ACC) are reported for the experiments. We adopt the five-fold cross-validation to evaluate the model's performance. Specifically, in each fold, we randomly split the dataset into the training, validation, and test dataset based on the prescribed proportion. Each metric's mean and standard deviation values are computed based on the five-fold results with the same random seed. We have also utilized the paired t-test as a statistical method to determine the comparability of two paired results. Whenever the obtained P-value is larger than 0.05, we conclude that the two results are comparable.

## D. Comparisons With State-of-the-Art

We adopt three types of state-of-the methods for whole slide image classification as the comparison baselines: 1) MIL-based methods: ABMIL [3], CLAM [16], and DTMIL [18]. 2) Multi-scale-based methods: DSMIL [4], ZoomMIL [5], and $H^2$MIL [6]. 3) Transformer-based methods: TransMIL [7] and GT [11]. According to their original settings, the MIL-based and the Transformer-based methods both use the patches cropped from the slide at 10×. The multi-scale-based methods and our MG-Trans use the patches cropped from the slides at 5× and 10×. Note that the patch feature for all methods is extracted in the same manner.

*1) Cancer Subtyping:* We compare our model against baseline methods on three cancer subtyping datasets, and the experiment results are reported in Table II. The traditional MIL-based methods (*i.e.*, ABMIL, CLAM, and DTMIL) exhibit inferior performance compared to other methods. This is because these MIL-based methods take all the patches as the input without taking into account the impact of irrelevant and redundant patches on the model performance. Most of the multi-scale-based methods achieve better performance than the traditional MIL-based methods. $H^2$MIL achieves the best performance compared to the other two multi-scale methods (*i.e.*, DSMIL, and ZoomMIL). Because the spatial relations are modeled between multi-scale patches by $H^2$MIL. But the multi-scale patch features are linearly aggregated together, ignoring to reduce the redundant information in the bag-level representation. Moreover, the Transformer-based methods are proposed to fully model the complicated spatial relations between patches, but the local spatial relations are not explicitly modeled. MG-Trans almost achieves the best results compared with all the other methods in all three datasets. Specifically, MG-Trans achieves an improvement of 4.3% and 3% on F1 and ACC in the BRIGHT3 dataset compared with GT. This indicates that MG-Trans can accurately capture the most critical discriminative features for the subtyping tasks. BRIGHT6 is a more challenging task because it includes more fine-grained catalogs that lead to higher inter-class ambiguities. For example, the morphological difference between subtypes *ADH* and *DCIS* is small in clinical practice. In BRIGHT6, the experiment results still indicate its ability to discriminate the more fine-grained and subtle features. The best experiment results on the TCGA-BRCA and TCGA-RCC datasets also show that MG-Trans is suitable for multiple cancer subtyping tasks. Moreover, MG-Trans almost demonstrates a lower variance across all metrics, indicating its capacity to

TABLE III

RESULT (PRESENT IN %) ON SEVEN GENE MUTATION DETECTION DATASETS. ± REPRESENTS MEAN ± STANDARD DEVIATION. THE TOP RESULTS ARE MARKED IN BOLD, AND THE COMPARABLE PERFORMANCES ARE DENOTED BY SUPERSCRIPT * BASED ON THE PAIRED T-TEST (P-VALUE>0.05)

| Metric | F1 | | | | | | | ACC | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gene | MAP3K1 | KMT2C | GATA3 | TP53 | PIK3CA | CDH1 | MSI | MAP3K1 | KMT2C | GATA3 | TP53 | PIK3CA | CDH1 | MSI |
| MIL-based | | | | | | | | | | | | | | |
| ABMIL | 47.9±0.4 | 47.6±0.4* | 47.3±1.0 | 58.5±3.8 | 50.4±4.7 | 48.1±2.1 | 49.9±4.3 | 92.0±1.1* | 90.8±1.3 | 84.3±4.1 | 63.0±3.8 | 67.9±1.3 | 88.4±1.6 | 79.7±2.1* |
| CLAM | 48.1±0.2 | 47.6±0.5* | 46.5±0.3 | 62.8±6.3 | 48.2±7.0 | 51.7±3.6 | 49.4±7.4 | **92.5±0.7** | 91.0±1.8 | 85.0±1.1 | 65.3±4.6 | 68.9±0.4 | 87.6±0.7 | 78.5±4.1 |
| Multi-scale-based | | | | | | | | | | | | | | |
| DSMIL | 47.7±0.3 | 47.4±0.6* | 48.8±4.4 | 60.5±7.2 | 48.0±4.0 | 49.5±2.7 | 54.6±9.6 | 91.4±1.1 | 90.3±2.0 | 85.9±2.4* | 63.6±6.8 | 68.8±1.4 | 87.7±3.7 | **80.9±1.5** |
| $H^2$MIL | 49.2±0.5 | 47.5±0.3* | 47.2±6.9 | 67.4±3.4 | 54.7±2.0 | 51.9±4.3 | 56.9±5.2 | 91.7±0.5 | 90.9±0.8 | 86.0±1.3* | 66.3±2.2 | 69.2±1.1* | 88.1±1.1 | 79.6±1.6* |
| Transformer-based | | | | | | | | | | | | | | |
| GT | 48.0±0.1 | 47.6±0.1* | 46.3±5.1 | 60.7±3.8 | 48.1±3.6 | 49.8±3.5 | 49.3±6.4 | 92.1±0.4* | 91.5±0.4* | 86.1±0.9* | 64.6±4.9 | 67.9±1.6 | 88.9±1.2* | 78.9±1.0 |
| **MG-Trans** | **51.1±0.6** | **47.8±0.1** | **49.8±4.7** | **68.1±2.3** | **58.2±3.8** | **52.5±3.4** | **61.4±4.8** | 92.0±1.0* | **91.7±0.4** | **86.6±0.8** | **71.1±1.7** | **69.5±1.8** | **89.5±1.0** | 79.7±1.1* |

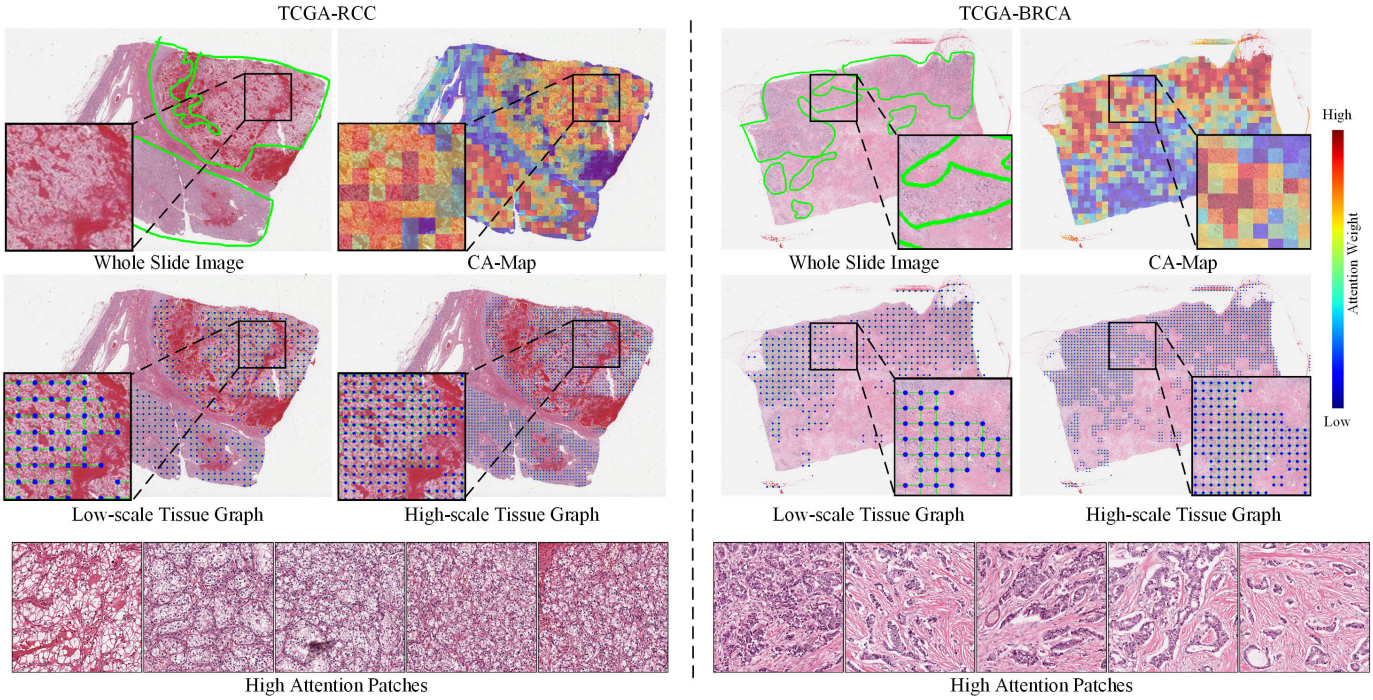



Fig. 4. Interpretability Analysis. The tumor regions within the original image are demarcated by professional pathologists using green lines. The attention value of each patch in CA-Map is also visualized. In the low-scale and high-scale tissue graphs, the nodes (patches) are denoted as blue dots, and the edges are represented as green lines. Five critical patches with the highest attention values of each case are also visualized to directly provide diagnosis-related evidence.

generate a robust bag-level representation and enhance the model's generalization.

*2) Gene Mutation Detection:* To evaluate the performance of MG-Trans on the gene mutation detection task, we employ the F1, ACC, and AUC as evaluation metrics. Compared to the cancer subtyping task, the gene mutation detection task can be more challenging due to the absence of distinct mutational patterns. As shown in Table III and Table IV, MG-Trans almost achieves all the best results in three metrics compared to all the current SOTA methods. Specifically, MG-Trans outperforms them by an average of 2.0% and 3.2% on F1 and AUC among all the mutated genes. The experimental results demonstrate the superior discriminative ability of the MG-Trans on the mutated gene-related critical features.

## E. Interpretability Analysis

In this study, we present a visual analysis of our proposed MG-Trans framework to demonstrate its interpretability ability. Specifically, we visually present the CA-Map, two-scale of tissue graphs, and critical patches to showcase the framework's interpretability ability. As illustrated in Fig. 4, the CA-Map identifies patches with high attention values that exhibit good spatial consistency with the annotated tumor regions identified by professional pathologists. This indicates

TABLE IV

RESULT (PRESENT IN %) ON SEVEN GENE MUTATION DETECTION DATASETS.± REPRESENTS MEAN ± STANDARD DEVIATION. THE TOP BEST RESULTS ARE MARKED IN BOLD, AND THE COMPARABLE PERFORMANCES ARE DENOTED BY SUPERSCRIPT * BASED ON THE PAIRED T-TEST (P-VALUE>0.05)

| Metric | AUC | | | | | | |
|---|---|---|---|---|---|---|---|
| Gene | MAP3K1 | KMT2C | GATA3 | TP53 | PIK3CA | CDH1 | MSI |
| MIL-based | | | | | | | |
| ABMIL | 42.7±3.1 | 51.0±2.6 | 51.2±3.1 | 63.8±5.3 | 53.2±4.2 | 57.3±3.9 | 59.6±4.1 |
| CLAM | 43.8±2.5 | 50.1±1.1 | 50.7±1.2 | 65.8±3.9 | 51.1±5.1 | 59.1±4.7 | 61.6±4.6 |
| Multi-scale-based | | | | | | | |
| DSMIL | 42.1±1.7 | 51.1±2.7* | 51.8±4.1 | 63.0±6.9 | 51.7±3.2 | 58.4±7.3 | 65.5±8.7 |
| $H^2$MIL | 53.2±2.1 | 50.7±1.3 | 51.3±5.7 | 67.1±4.4 | 56.9±5.4 | 60.2±7.8 | 66.8±4.4 |
| Transformer-based | | | | | | | |
| GT | 52.8±1.9 | 51.1±1.8* | 50.5±4.2 | 63.2±3.9 | 52.3±2.9 | 58.5±4.2 | 62.3±5.5 |
| **MG-Trans** | **55.2±1.3** | **51.3±1.2** | **53.3±3.7** | **73.4±3.5** | **60.8±2.7** | **64.0±3.7** | **71.7±5.1** |

TABLE V

ABLATION EXPERIMENT (PRESENT IN %) ON THE BRIGHT3 DATASET. ± REPRESENTS MEAN ± STANDARD DEVIATION. THE COMPARABLE PERFORMANCES ARE DENOTED BY SUPERSCRIPT * BASED ON THE PAIRED T-TEST (P-VALUE>0.05)

| Setting | | | | | Metric | | | Computation | |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | PAM | SILM | MIBM | SC | AUC | F1 | ACC | GFLOPs | Average Test Time |
| ✓ | | | | | 82.3 ± 2.1 | 60.5 ± 4.6 | 70.2 ± 1.2 | 3.87 | 0.07s |
| ✓ | ✓ | | | | 83.0 ± 2.8 | 61.6 ± 5.3 | 71.3 ± 2.3 | 3.11 | 0.04s |
| ✓ | ✓ | ✓ | | | 84.2 ± 2.9* | 63.0 ± 4.2 | 72.7 ± 3.2 | 3.38 | 0.5s |
| ✓ | ✓ | ✓ | ✓ | | 84.8 ± 3.5 | 65.4 ± 5.3 | 74.3 ± 3.5* | 3.38 | 0.5s |
| ✓ | ✓ | ✓ | ✓ | ✓ | 84.8 ± 3.6* | 66.3 ± 5.8 | 74.6 ± 3.8 | 3.38 | 0.5s |

that our proposed CA-Map has good localization ability of tumor regions and can help the framework filter out a large number of redundant and irrelevant input patches. Furthermore, we visualize the tissue graphs from low- and high-scale to capture the local tumor tissue structures in two granularities. This is critical to enhancing the model's discriminative ability to detect subtypes with subtle differences. We also present five critical patches with the highest attention values for each case to provide direct diagnostic evidence to the pathologists. These selected key patches are then examined by three professional pathologists on our team. According to the consensus among pathologists, the selected patches effectively reflect the critical features of each subtype. For example, the selected patches exhibit features of transparent cell morphology in tumor cells, which is a typical histological characteristic of the subtype of CCRCC.

### F. Ablation Studies

*1) Effects of Each Module in MG-Trans:* To study the impact of each module in MG-Trans, we conduct a series of ablation studies on the BRIGHT3 dataset. The floating point operations (FLOPs) and average prediction time for each module are also computed to evaluate their computation cost. The experiment results are reported in Table V. The specific module ablation settings are described as follows:

- Baseline: two cls tokens are separately concatenated with the 5× and 10× patch features extracted by a pre-trained ResNet50 on ImageNet. These two features are inputted into a Transformer model that consists of two general blocks. The updated two cls tokens are concatenated and fed into a fusion network and a classifier for classification.
- Baseline + PAM: a total of 512 patches of the low-scale are selected by PAM, which is the same as the setting of MG-Trans. The high-scale patches, which have space corresponding relations with the low-scale patches, are also selected. After passing through the Transformer encoder, the high- and low-scale cls tokens are concatenated and fed into a fusion network and a classifier for classification.
- Baseline + PAM + SILM: compared with "Baseline + PAM", only the SILM is introduced into the last general block of the Transformer encoder.
- Baseline + PAM + SILM + MIBM: compared with "Baseline + PAM + SILM", the high- and low-scale cls tokens are fused by our proposed MIBM.
- Baseline + PAM + SILM + SC: it's the framework MG-Trans proposed by this work.

Firstly, in terms of classification performance, after introducing the PAM, the model achieves better performance as it filters the irrelevant and redundant input patches out and helps the model focus on the critical regions. After introducing the SILM, the model's performance has been improved further. The explicit capture of local tissue structures by SILM has improved the model's ability to discern hard regions. The MIBM module brings a performance gain on all three metrics suggesting that it effectively fuses multi-scale instance features and generates a more robust bag-level representation for the classification task. Furthermore, we conducted statistical testing to assess the significance of including each module. After introducing the three primary modules (*i.e.*, PAM, SILM, and MIBM), the model's performance exhibits a significant improvement across three metrics. Additionally, the semantic consistency (SC) loss contributes to a gain in performance on the F1 and ACC metrics by helping the model learn consistent semantic information between the two scales.

TABLE VI

RESULT (PRESENT IN %) OF THE PAM ON BASELINES ON THE BRIGHT3 DATASET. ± REPRESENTS MEAN ± STANDARD DEVIATION

| Method | AUC | F1 | ACC |
|---|---|---|---|
| ABMIL | 77.8 ± 3.0 | 47.6 ± 3.3 | 64.8 ± 4.6 |
| ABMIL + PAM | 79.3 ± 4.6 | 52.0 ± 4.0 | 67.0 ± 4.9 |
| Δ | 1.5 ↑ | 4.4 ↑ | 2.2 ↑ |
| TransMIL | 77.1 ± 4.7 | 60.9 ± 6.4 | 70.2 ± 3.1 |
| TransMIL + PAM | 78.5 ± 4.3 | 62.5 ± 5.7 | 71.3 ± 3.3 |
| Δ | 1.4 ↑ | 1.6 ↑ | 1.1 ↑ |
| GT | 84.2 ± 3.6 | 62.0 ± 4.3 | 70.6 ± 3.9 |
| GT + PAM | 84.3 ± 3.5 | 63.1 ± 4.2 | 71.9 ± 3.4 |
| Δ | 0.1 ↑ | 1.1 ↑ | 1.3 ↑ |

Secondly, in terms of computation cost, after introducing the PAM, the model's computation cost decreases from 3.87 GFLOPs to 3.11 GFLOPs. After introducing the SILM, the computation cost has a certain degree of increment and the average test time is approximately 10 times slower. As SILM builds the tissue graph in an online way, it inevitably demands a slight increase in computation time. Note that MIBM and SC are the loss constraints that are solely used during training and their implementation does not incur any additional computation cost or testing time in practical deployment.

*2) Effect of PAM and Multi-Scale Learning:* To evaluate the effect of PAM and the Multi-scale Learning method on the baseline model, we select three baseline methods (*i.e.*, ABMIL [3], TransMIL [7], and GT [11]). For the PAM setting, we add PAM to the front of the baseline model and select 512 critical instances to ensure consistency with the comparable experiments. The experiment results are reported in Table VI. The introduction of PAM led to varying degrees of improvement across all three metrics, suggesting that PAM can also help filter out redundant and irrelevant input instances in the baseline. For the multi-scale learning method, we combine

TABLE VII
RESULT (PRESENT IN %) OF THE MULTI-SCALE FEATURES ON BASELINES ON THE BRIGHT3 DATASET. ± REPRESENTS MEAN ± STANDARD DEVIATION

| Method | AUC | F1 | ACC |
|---|---|---|---|
| ABMIL | 77.8 ± 3.0 | 47.6 ± 3.3 | 64.8 ± 4.6 |
| ABMIL + Multi-scale | 79.1 ± 3.7 | 51.2 ± 3.7 | 66.5 ± 4.7 |
| Δ | 1.3 ↑ | 3.6 ↑ | 1.7 ↑ |
| TransMIL | 77.1 ± 4.7 | 60.9 ± 6.4 | 70.2 ± 3.1 |
| TransMIL + Multi-scale | 78.3 ± 4.6 | 62.4 ± 5.8 | 71.5 ± 3.4 |
| Δ | 1.2 ↑ | 1.5 ↑ | 1.3 ↑ |
| GT | 84.2 ± 3.6 | 62.0 ± 4.3 | 70.6 ± 3.9 |
| GT + Multi-scale | 84.5 ± 3.9 | 63.4 ± 4.8 | 71.7 ± 3.5 |
| Δ | 0.3 ↑ | 1.4 ↑ | 1.1 ↑ |

TABLE VIII
RESULT (PRESENT IN %) OF DIFFERENT PATCH SIZES ON THE BRIGHT3 DATASET. ± REPRESENTS MEAN ± STANDARD DEVIATION

| Patch Size | AUC | F1 | ACC |
|---|---|---|---|
| 128 | 85.1 ± 3.2 | 66.7 ± 5.5 | 74.9 ± 3.5 |
| 256 | 84.8 ± 3.6 | 66.3 ± 5.8 | 74.6 ± 3.8 |
| 512 | 82.5 ± 4.1 | 63.6 ± 5.9 | 71.5 ± 4.3 |

TABLE IX
RESULT (PRESENT IN %) OF DIFFERENT ENCODER SETTINGS FOR PATCH FEATURE EXTRACTION ON THE BRIGHT3 DATASET. ± REPRESENTS MEAN ± STANDARD DEVIATION

| Setting | Method | AUC | AC | F1 |
|---|---|---|---|---|
| ResNet50 + ImageNet | $H^2$MIL | 84.6±4.8 | 60.2±9.4 | 71.6±6.6 |
| | GT | 84.2±3.6 | 62.0±4.3 | 70.6±3.9 |
| | **MG-Trans** | **84.8±3.6** | **66.3±5.8** | **74.6±3.8** |
| ResNet18 + SimCLR | $H^2$MIL | 85.1±3.4 | 62.1±7.5 | 72.1±5.8 |
| | GT | 84.9±2.9 | 64.1±3.6 | 71.8±3.7 |
| | **MG-Trans** | **85.4±2.5** | **66.5±4.7** | **75.2±4.1** |
| ViT + DINO | $H^2$MIL | 85.2±2.9 | 62.3±4.1 | 72.3±4.9 |
| | GT | 85.1±3.1 | 64.0±3.9 | 72.0±4.0 |
| | **MG-Trans** | **85.6±2.7** | **66.9±3.6** | **75.5±3.2** |

the multi-scale instance features as the input of the baseline model. Specifically, the spatially overlapping 5× and 10× patch features are concatenated as the baseline input. The results are reported in Table VII. After introducing the multi-scale patch features as the input, all three baselines show a certain degree of performance improvement across three metrics. This indicates that the multi-scale patch features are complementary. The low-scale patch features can help the baseline models learn more tissue structure information, while the high-scale patch features contain more tissue detail information.

*3) Effect of Patch Size:* To evaluate the effect of patch size on the model performance, we select three kinds of patch sizes (*i.e.*, 128, 256, and 512) to conduct the ablation experiment. The experiment results are reported in Table VIII. The best patch size for MG-Trans is 256. When the patch size is larger (*i.e.*, 512), each patch may contain various types of tissues, leading to ineffective modeling of the fine-grained spatial tissue structure information. When the patch size is smaller (*i.e.*, 128), the model performance has slightly improved. However, the computation cost associated with the patch size 128 is 4× higher than the patch size of 256.

*4) Effect of Patch Feature Extraction:* To evaluate the effect of different feature encoders $h(\cdot)$ on model performance, we select three self-supervised methods: 1) ResNet50 [8] + ImageNet [38]: the encoder uses the ResNet50 model truncated after the third residual block, which is pretrained on the ImageNet dataset by transfer learning; 2) ResNet18 [8] + SimCLR [41]: the encoder uses ResNet18 pretrained by the contrastive self-supervised method SimCLR on 57 histopathology dataset including multiple multi-organ datasets with different types of staining and multiple resolutions, which was released by [42]; 3) ViT [24] + DINO [43]: the encoder is the vision Transformer (ViT) using the DINO-based knowledge distillation strategy pretrained on 33 cancer types in 10,678 WSIs, which was released by [27]. The experiments are reported in Table IX. Specifically, fine-tuning the feature encoder with the domain-specific data can further enhance the model performance on downstream tasks. MG-Trans still achieves the best results on all three metrics, indicating its robustness and superiority to multiple feature extraction methods.

TABLE X
RESULT (PRESENT IN %) OF THE SAMPLING PATCH NUMBERS IN PAM ON THE BRIGHT3 DATASET. ± REPRESENTS MEAN ± STANDARD DEVIATION

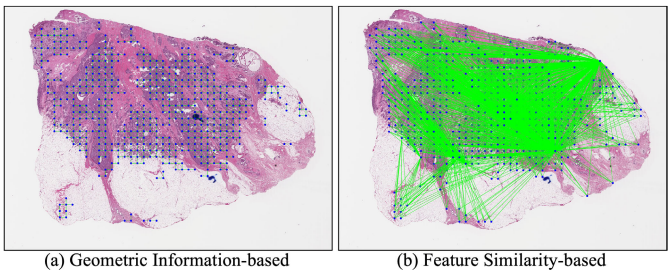
| Patch Number | AUC | F1 | ACC |
|---|---|---|---|
| 64 | 82.3 ± 2.3 | 58.9 ± 7.2 | 71.6 ± 4.3 |
| 128 | 84.2 ± 2.6 | 63.7 ± 2.6 | 73.2 ± 2.9 |
| 256 | 84.5 ± 4.0 | 64.6 ± 4.2 | 73.6 ± 5.4 |
| 512 | **84.8 ± 3.6** | **66.3 ± 5.8** | **74.6 ± 3.8** |
| 640 | 84.1 ± 3.6 | 65.4 ± 4.7 | 73.0 ± 5.0 |

TABLE XI
RESULT (PRESENT IN %) OF DIFFERENT SAMPLING METHODS ON THE BRIGHT3 DATASET. ± REPRESENTS MEAN ± STANDARD DEVIATION

| Method | AUC | F1 | ACC |
|---|---|---|---|
| Random | 83.0 ± 3.4 | 61.9 ± 5.6 | 72.2 ± 3.7 |
| Clustering | 83.5 ± 2.3 | 63.2 ± 4.2 | 73.8 ± 2.0 |
| Attention | 83.6 ± 3.7 | 62.9 ± 4.7 | 71.6 ± 4.7 |
| PAM | **84.8 ± 3.6** | **66.3 ± 5.8** | **74.6 ± 3.8** |

*5) Effect of Sampling Patch Numbers in PAM:* To study the effect of sampling patch numbers on model performance, we select the different numbers of patches based on the CA-Map. As shown in Table X, the relatively optimal sampling patch number is 512. When the sampling number is too low (*i.e.*, 64), PAM may drop too many informative patches, and the selected patches are insufficient to capture better class discrimination. When the sampling patch number is too high (*i.e.*, 640), redundant information may be introduced by selecting uninformative patches, degenerating the model performance.

*6) Effect of Patch Sampling Methods:* To study the effect of commonly used sampling methods on the model performance, we compare PAM with three types of sampling methods (*i.e.*, random-based, clustering-based, and attention-based). The random-based method randomly selected the same number (*i.e.*, 512) of patches as the PAM. The clustering-based method utilizes the K-means method to cluster all the input patches into eight clusters first and average sampling patches from each cluster. This is due to the fact that for a whole slide image (WSI), there are typically eight common tissue types present, namely tumor, inflammatory, connective, dead, epithelial, fat, fibrous stroma, and muscular. By dividing these tissue types into different clusters, we are able to select a diverse range of tissue samples for analysis. The attention-based method utilizes the gated attention mechanism [3] as



(a) Geometric Information-based (b) Feature Similarity-based

Fig. 5. Graph building methods. (a) Each node will connect its nearest several nodes based on the Euclidean distance and the max distance between two nodes is the patch size; (b) Each node will calculate the cosine similarity and connect the four nodes with the highest cosine similarity among the others.

TABLE XII
RESULT (PRESENT IN %) OF DIFFERENT GRAPH BUILDING METHODS ON BRIGHT3 DATASET. ± REPRESENTS MEAN ± STANDARD DEVIATION. ± REPRESENTS MEAN ± STANDARD DEVIATION

| Method | AUC | F1 | ACC |
|---|---|---|---|
| Feature Similarity | 82.3 ± 3.9 | 63.7 ± 4.7 | 72.2 ± 3.9 |
| Geometric Information | **84.8 ± 3.6** | **66.3 ± 5.8** | **74.6 ± 3.8** |

the metric to select the patches with high attention values. Note that the total sampling patch number for each method is identical. As shown in Table XI, PAM achieves the best results compared to the other sampling methods, which illustrates its good location ability of informative patches.

*7) Effect of Graph Building Methods:* To evaluate the impact of different graph-building methods on model performance, we conducted an ablation study comparing geometric information-based and feature similarity-based graph construction methods. The geometric information-based method has been described in the section of methodology. The feature similarity-based method calculates the cosine similarity between two nodes and connects the four nodes with the highest similarity among others. As shown in Fig. 5, the geometric-based method effectively captures the local tissue structure. The feature similarity-based method can also model the spatial relations between nodes. However, long-distance edges in the feature similarity-based method may limit the ability to focus on fine-grained local tissue structures. The experiment results are reported in Table XII. Specifically, the geometric information-based method achieves superior performances across all three metrics on the BRIGHT3 dataset, further indicating its efficacy in capturing local tissue structure information.

*8) Effect of Feature Fusion Methods:* To study the effect of different multi-scale feature fusion methods on the model performance, we select two commonly used feature fusion methods (*i.e.*, concatenation, and summation). As shown in Table XIII, our MIBM outperforms other methods in all three metrics indicating that it generates a more compact and robust bag-level representation and improves the accuracy and generalization of the model.

## V. DISCUSSION

Although MG-Trans achieves state-of-the-art results on three cancer subtyping datasets and seven gene mutation datasets, it still exhibits some limitations. Firstly, PAM is designed to filter out irrelevant and redundant input patches, which enables the model to focus on the most critical instances and reduce computation costs. However, this filtering process can also cause the model to fail to locate all cancerous regions in the slide, resulting in inaccurate quantification of tumor regions. This is because some cancerous instances may also be filtered out as redundant parts. Nonetheless, as shown in Fig. 4, our proposed CA-Map has been proven to be effective in locating most of the cancerous regions. Secondly, SILM builds the tissue graph in an online way, which facilitates the dynamic creation of local tissue structure information. However, this approach may lead to a minor reduction in the model's inference speed. This issue could potentially be resolved by designing local location encoding methods that are well-suited for pathological images. In the future, we will extend our model to more cancer diagnostic tasks and further improves its interpretability and inference speed.

TABLE XIII
RESULT (PRESENT IN %) OF DIFFERENT MULTI-SCALE FEATURE FUSION METHODS ON THE BRIGHT3 DATASET. ± REPRESENTS MEAN ± STANDARD DEVIATION

| Method | AUC | F1 | ACC |
|---|---|---|---|
| Concat | 84.0 ± 3.9 | 64.8 ± 6.0 | 74.0 ± 4.1 |
| Sum | 84.1 ± 4.5 | 63.4 ± 6.7 | 73.8 ± 3.9 |
| MIBM | **84.8 ± 3.6** | **66.3 ± 5.8** | **74.6 ± 3.8** |

## VI. CONCLUSION

In this work, we proposed a multi-scale graph Transformer (MG-Trans) with information bottleneck for whole slide image classification. The patch anchoring module projects a large number of input patches into a smaller and fixed number of informative patches. The dynamic structure information learning module explicitly introduces the local tissue structure into the Transformer block, improving the discriminative ability to the subtle difference between subtypings. MIBM fuses the multi-scale features by generating a compact yet sufficient bag-level representation for the classification task. Besides, a semantic consistency loss is proposed to stabilize the training of the model. Extensive comparative and ablation experiments on three cancer subtypings datasets and seven gene mutation detection datasets show that MG-Trans achieved state-of-the-art results for whole slide image classification.

## REFERENCES

[1] M. Cui and D. Y. Zhang, "Artificial intelligence and computational pathology," *Lab. Invest.*, vol. 101, no. 4, pp. 412–422, Apr. 2021.
[2] O. Iizuka, F. Kanavati, K. Kato, M. Rambeau, K. Arihiro, and M. Tsuneki, "Deep learning models for histopathological classification of gastric and colonic epithelial tumours," *Sci. Rep.*, vol. 10, no. 1, p. 1504, Jan. 2020.
[3] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.
[4] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14318–14328.

[5] K. Thandiackal et al., "Differentiable zooming for multiple instance learning on whole-slide images," in *Proc. Eur. Conf. Comput. Vis.*, vol. 13681. Cham, Switzerland: Springer, 2022, pp. 699–715.

[6] W. Hou et al., "$H^2$-MIL: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 933–941.

[7] Z. Shao et al., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2136–2147.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[9] W. Shao et al., "Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 99–110, Jan. 2020.

[10] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, "A survey on graph-based deep learning for computational histopathology," *Computerized Med. Imag. Graph.*, vol. 95, Jan. 2022, Art. no. 102027.

[11] Y. Zheng et al., "A graph-transformer for whole slide image classification," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3003–3015, Nov. 2022.

[12] P. Pati et al., "Hierarchical graph representations in digital pathology," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102264.

[13] H. Li et al., "DT-MIL: Deformable transformer for multi-instance learning on histopathological image," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 206–216.

[14] Y. Zhao et al., "SETMIL: Spatial encoding transformer-based multiple instance learning for pathological image analysis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 66–76.

[15] H. Sun, X. He, and Y. Peng, "SIM-Trans: Structure information modeling transformer for fine-grained visual categorization," in *Proc. 30th ACM Int. Conf. Multimedia (ACM MM)*, Oct. 2022, pp. 5853–5861.

[16] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, Mar. 2021.

[17] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, Aug. 2014.

[18] H. Zhang et al., "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18802–18812.

[19] Z. Wang, L. Yu, X. Ding, X. Liao, and L. Wang, "Lymph node metastasis prediction from whole slide images with transformer-guided multiinstance learning and knowledge transfer," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2777–2787, Oct. 2022.

[20] J. Yao, X. Zhu, and J. Huang, "Deep multi-instance learning for survival prediction from whole slide images," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 496–504.

[21] Y. Sharma, A. Shrivastava, L. Ehsan, C. A. Moskaluk, S. Syed, and D. Brown, "Cluster-to-Conquer: A framework for end-to-end multi-instance learning for whole slide image classification," in *Proc. Med. Imag. Deep Learn.*, 2021, pp. 682–698.

[22] J. Zhang et al., "A joint spatial and magnification based attention framework for large scale histopathology classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3771–3779.

[23] A. BenTaieb and G. Hamarneh, "Predicting cancer with a recurrent visual attention model for histopathology images," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 129–137.

[24] A. Dosovitskiy et al., "An image is worth $16 \times 16$ words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[25] J. He et al., "TransFG: A transformer architecture for fine-grained recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 852–860.

[26] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4692–4702.

[27] R. J. Chen et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16144–16155.

[28] Z. Lv, R. Yan, Y. Lin, Y. Wang, and F. Zhang, "Joint region-attention and multi-scale transformer for microsatellite instability detection from whole slide images in gastrointestinal cancer," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 293–302.

[29] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Annu. Allerton Conf. Communnication, Control Comput.*, 1999, pp. 368–377.

[30] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2015, pp. 1–5.

[31] A. Zhmoginov, I. Fischer, and M. Sandler, "Information-bottleneck approach to salient region discovery," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2021, pp. 531–546.

[32] S. Bang, P. Xie, H. Lee, W. Wu, and E. Xing, "Explaining a black-box by using a deep variational information bottleneck approach," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 13, pp. 11396–11404.

[33] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[34] Q. Wang, C. Boudreau, Q. Luo, P. Tan, and J. Zhou, "Deep multi-view information bottleneck," in *Proc. SIAM Int. Conf. Data Mining*, 2019, pp. 37–45.

[35] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–26.

[36] M. I. Belghazi et al., "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 531–540.

[37] L. G. S. Giraldo, M. Rao, and J. C. Principe, "Measures of entropy from data using infinitely divisible kernels," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 535–548, Jan. 2015.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[40] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch geometric," 2019, *arXiv:1903.02428*.

[41] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.

[42] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Mach. Learn. Appl.*, vol. 7, Mar. 2022, Art. no. 100198.

[43] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.