Positional Encoding-Guided Transformer-Based Multiple Instance Learning for Histopathology Whole Slide Images Classification

Jun Shi[a], Dongdong Sun[b], Kun Wu[c], Zhiguo Jiang[c], Xue Kong[d,e], Wei Wang[d,e], Haibo Wu[d,e] and Yushan Zheng[f,*]

[a]School of Software, Hefei University of Technology, Hefei, 230601, Anhui Province, China
[b]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230601, Anhui Province, China
[c]Image Processing Center, School of Astronautics, Beihang University, Beijing, 102206, China
[d]Department of Pathology, the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, 230036, Anhui Province, China
[e]Intelligent Pathology Institute, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, 230036, Anhui Province, China
[f]School of Engineering Medicine, Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing, 100191, China

ARTICLE INFO

Keywords:
Digital pathology
Whole slide image
Multiple instance learning
Position encoding
Cancer subtyping
Gene mutation prediction

ABSTRACT

Background and objectives: Whole slide image (WSI) classification is of great clinical significance in computer-aided pathological diagnosis. Due to the high cost of manual annotation, weakly supervised WSI classification methods have gained more attention. As the most representative, multiple instance learning (MIL) generally aggregates the predictions or features of the patches within a WSI to achieve the slide-level classification under the weak supervision of WSI labels. However, most existing MIL methods ignore spatial position relationships of the patches, which is likely to strengthen the discriminative ability of WSI-level features.

Methods: In this paper, we propose a novel positional encoding-guided transformer-based multiple instance learning (PEGTB-MIL) method for histopathology WSI classification. It aims to encode the spatial positional property of the patch into its corresponding semantic features and explore the potential correlation among the patches for improving the WSI classification performance. Concretely, the deep features of the patches in WSI are first extracted and simultaneously a position encoder is used to encode the spatial 2D positional information of the patches into the spatial-aware features. After incorporating the semantic features and spatial embeddings, multi-head self-attention (MHSA) is applied to explore the contextual and spatial dependencies of the fused features. Particularly, we introduce an auxiliary reconstruction task to enhance the spatial-semantic consistency and generalization ability of features.

Results: The proposed method is evaluated on two public benchmark TCGA datasets (TCGA-LUNG and TCGA-BRCA) and two in-house clinical datasets (USTC-EGFR and USTC-GIST). Experimental results validate it is effective in the tasks of cancer subtyping and gene mutation status prediction. In the test stage, the proposed PEGTB-MIL outperforms the other state-of-the-art methods and respectively achieves 97.13±0.34%, 86.74±2.64%, 83.25±1.65%, and 72.52±1.63% of the area under the receiver operating characteristic (ROC) curve (AUC).

Conclusions: PEGTB-MIL utilizes positional encoding to effectively guide and reinforce MIL, leading to enhanced performance on downstream WSI classification tasks. Specifically, the introduced auxiliary reconstruction module adeptly preserves the spatial-semantic consistency of patch features. More significantly, this study investigates the relationship between position information and disease diagnosis and presents a promising avenue for further research.

1. Introduction

With the rapid development of digital pathology and Artificial Intelligence (AI), histopathology Whole Slide Image (WSI) classification based on deep learning has been widely used in cancer subtyping [1–3], tumor grading [4–6], prognosis analysis [7–9], gene mutation prediction [10–12], etc., which can promote the diagnosis efficiency and quality for pathologists. In the past decades, the WSI classification methods usually perform supervised deep learning models on the lesion regions or cells fully annotated by pathologists for feature learning [13–15]. Inevitably the annotation process is time-consuming and labour-intensive. Furthermore, the subjectivity and uncertainty of manual annotation can easily lead to the inconsistency of annotations and thus affect the quality of annotation data for AI model training. To deal with these problems, weakly supervised methods [16–19] have been widely applied in the field of computational pathology which generally use slide-level labels to supervise the feature learning of the patches within the WSI and further obtain the WSI-level feature representation for classification instead of region-level or cell-level fine-grained annotation. Multiple Instance Learning (MIL) [20–24] is the most representative weakly supervised method which generally regards a WSI as a bag and the divided patches within the WSI as the instances in the digital pathology community. It usually assumes that a WSI is considered positive in a binary classification problem if at least one instance is positive. Otherwise, the WSI is regarded as negative. Conventional MIL-based

*Corresponding author. yszheng@buaa.edu.cn (Yushan Zheng)

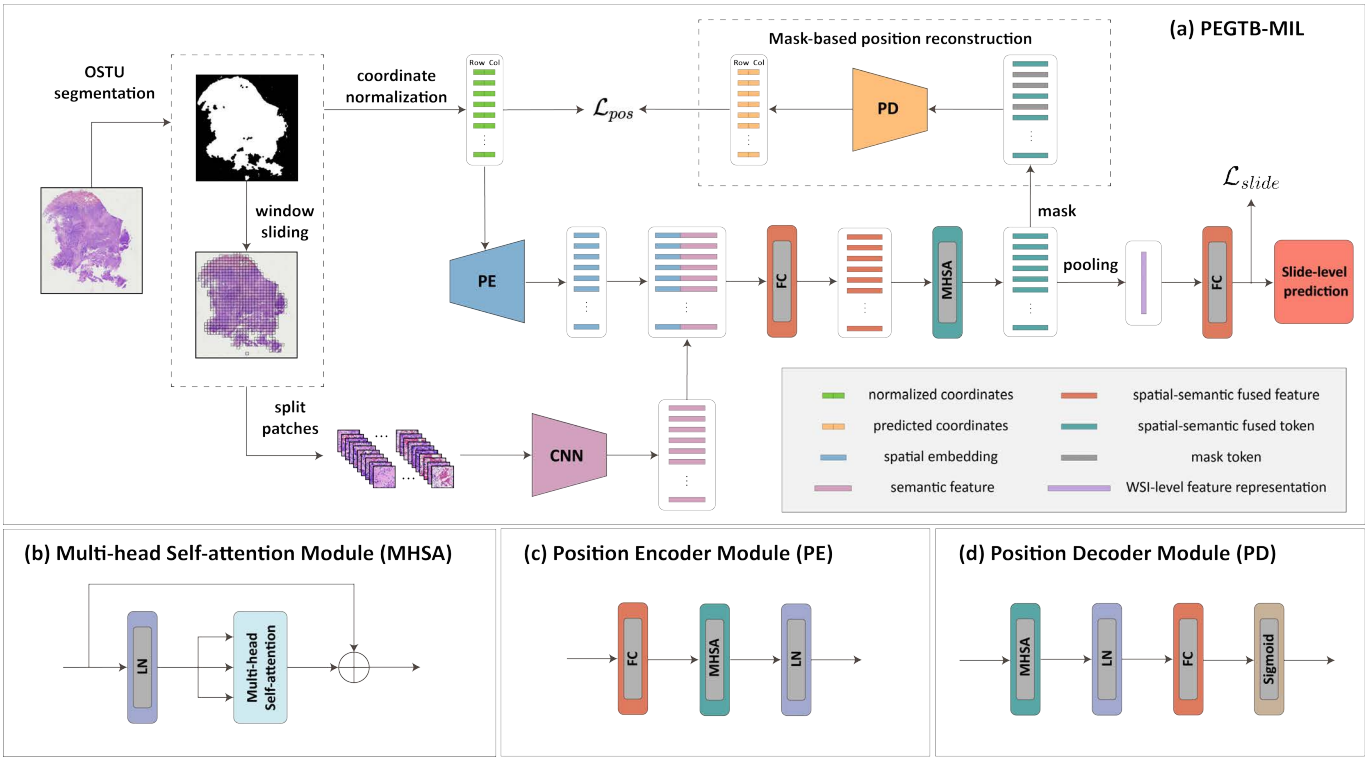


Fig. 1: The overview of the proposed PEGTB-MIL for WSI classification, where (a) shows the entire workflow, (b) describes the structure of the multi-head self-attention module, (c) and (d) illustrate the position encoder module and position decoder module, respectively.

WSI classification methods often perform pooling operation on the predictions of the instances to acquire the bag-level prediction [1, 25–27]. However, these methods rely on the instance-level classifier, and the performance is easily influenced by the pseudo labels of the instances. Instead, the current popular MIL-based WSI classification methods mostly aggregate instance-level feature embeddings extracted by Convolutional Neural Network (CNN) to learn bag-level representation and then achieve the WSI classification through a bag-level classifier [2, 24]. These methods obtain the superior classification performance, yet all the patches within a WSI are independently treated in these methods. Consequently, the global correlation among different patches is ignored and thus the classification performance and interpretability of the AI model is affected. Note that the attention-based method tries to use the attention mechanism to explore the contribution of each patch to the WSI label. However, it essentially calculates the attention scores for each patch independently and thus the intrinsic dependencies among the patches are not fully investigated.

Digital pathology WSIs contain rich morphological information and are therefore considered the gold standard for cancer diagnosis. In clinical, tumors represent spatially organized ecosystems comprised of diverse cell populations. The spatial distribution of cells serves as a tool for better understanding the tumor microenvironment, predicting outcomes, and potentially aiding in the selection of therapeutic interventions [28, 29]. The growth and progression of tumors involve spatial processes that encompass the destruction, invasion, and metastasis of normal tissue. For these reasons, spatial patterns are integral components of histological tumor grading and staging [30]. Similarly, several studies [29, 31] have demonstrated their close association with survival prognosis. The spatial information between different regions and their intrinsic semantic information can also help the AI model learn the morphological details and the tumor microenvironment-related patterns within the WSI structure, which are difficult for pathologists to directly identify, especially in the task of gene mutation prediction. Recently, self-attention [32, 33] has been successfully used in the communities of natural language processing and computer vision which can exploit the long-range dependencies among the tokens within the input sequence. Therefore, recent studies [34–37] apply self-attention for MIL-based histopathology WSI classification which can explore the global relationships among the patches within a WSI. However, most of these methods fail to consider the spatial position information of the patches and the spatial-semantic consistency of the patch features, which may help enhance the spatial-aware ability of the patch features and the representational ability of WSI-level features. Note that Vision Transformer (ViT) [33] introduces the position embeddings of the patches, yet it separately uses position encoding for horizontal and vertical coordinates of the patch and then concatenates the horizontal and vertical embeddings as the final position embeddings. Therefore, the spatial structural information

among the patches may be lost to some extent and thus be unsuitable for fine-grained WSI classification tasks (e.g. cancer subtyping and gene mutation status prediction).

Motivated by the above discussions, we propose a novel positional encoding-guided transformer-based multiple instance learning (PEGTB-MIL) method for histopathology WSI classification. The entire framework is shown in Fig. 1. Compared with the conventional transformed-based MIL methods for WSI classification, the proposed PEGTB-MIL uses a position encoder (PE) module to encode the normalized 2D positional coordinates of each tissue patch into the spatial embeddings. At the same time, the CNN features of the tissue patches are concatenated with their spatial embeddings as the final features of the patches. Then multi-head self-attention (MHSA) module is used to exploit the spatial and contextual correlation among the patches. Particularly, a position decoder (PD) module is designed to decode the patch features into the 2D position coordinates, which applies the mask-based position reconstruction for auxiliary guidance and thus improves the spatial-semantic consistency and generalization ability of the patch features. The proposed PEGTB-MIL is evaluated on two lung cancer datasets, a breast cancer dataset, and a gastrointestinal stromal tumor (GIST) dataset, and is compared with the state-of-the-art MIL-based methods [2, 24, 34–36]. Experimental results have validated that the proposed PEGTB-MIL has better WSI classification performance in the tasks of cancer subtyping and gene mutation status prediction.

The contributions of this paper can be summarized in three folds:

- We propose a novel transformer-based multiple instance learning framework for histopathology WSI classification. Different from the traditional transformer-based MIL methods, a position encoder module is used to uniformly encode the 2D position coordinates of the patches into the spatial embeddings. Then multi-head self-attention module is applied to explore the contextual and spatial dependencies among the patches within a WSI and thus the more discriminative WSI-level feature representation can be gained.
- We introduce mask-based position reconstruction as an auxiliary task to guide the model training. Unlike the MIL methods based on position encoding, a position decoder module is developed to guarantee the decoded spatial coordinates and the true coordinates of the patches are as consistent as possible. Consequently, the spatial-semantic consistency and generalization capability of the patch features can be greatly enhanced.
- We conduct experiments to validate the proposed method and existing state-of-the-art MIL-based methods on two public benchmark TCGA datasets and two in-house clinical datasets. The results prove that our method can achieve superior classification performance in lung and breast cancer subtyping. More importantly, it has also shown more promising results through directly using Hematoxylin and Eosin (H&E) histopathology WSIs to predict the epidermal growth factor receptor (EGFR) mutational status of lung cancer and KIT mutational status of GIST.

The rest of this paper is organized as follows. Section 2 reviews the MIL-related works. Section 3 introduces the proposed framework. Section 4 shows the experimental results and analysis. The discussion and conclusion are presented in Section 5 and Section 6, respectively.

2. Related works

In this section, we provide a brief overview of the related works on the MIL-based WSI classification methods and the MIL methods incorporating position encoding.

2.1. Multiple Instance Learning in WSI analysis

Due to the high cost of fine-grained annotations for lesion regions, WSI classification problem can be defined as a weakly supervised learning task. Currently, multiple instance learning (MIL) is a promising choice for weakly supervised WSI classification. It can be roughly classified into two categories: instance-based methods and bag-based methods.

For instance-based methods [1, 25–27], the simplest approach is to use max-pooling or average-pooling to aggregate the predicted probabilities of all the instances, thereby generating a bag-level prediction. Campanella et al. [1] proposed a MIL method based on Recurrent Neural Network (RNN), where the pseudo labels of the patches are used to train an instance-level classifier, and then the patches with top K positive probabilities generated by the classifier are selected as the input of RNN for the final WSI classification. Xu et al. [25] designed a weakly supervised learning framework CAMEL for histopathology image segmentation, which leverages a label enrichment strategy to dynamically refine the labels of the instances, and subsequently trains an instance classifier to achieve the instance-level classification and pixel-level segmentation. Qu et al. [27] introduced an instance-based MIL framework which combines contrastive learning and prototype learning to train an instance classifier for instance-level and bag-level classification. In short, these above methods mostly use the pseudo labels of the patches to train an instance classifier and then aggregate the predictions from all the instances for WSI classification. However, there exists the inherent noise within the pseudo labels due to the absence of true labels for the instances and thus the trained instance classifier may limit the WSI classification performance.

Bag-based methods [2, 24, 34–37] aggregate the patch features to obtain a WSI-level feature representation for classification through a bag classifier and they have been the mainstream of MIL-based WSI analysis methods. Ilse et al. [24] proposed an attention-base MIL (ABMIL) method, which calculates attention scores for each instance through an attention module, and then aggregates the features of all the instances into a bag embedding by treating the scores as weights. Lu et al. [2] presented a clustering-constrained-attention MIL (CLAM) method, which employs the attention module to identify critical regions for disease diagnosis, enabling accurate WSI classification. Different from ABMIL, CLAM optimizes the patch feature representation by

using an instance-level clustering module. This allows the attention module to better distinguish between positive and negative patches, resulting in more discriminative WSI-level representations. However, they have treated each patch as an independent entity and thus the global dependencies among the patches are not fully explored.

Recently the transformer architecture has achieved great success in various AI application scenarios, which aims to use self-attention to capture the long-range dependencies among the tokens for a given sequence. In the field of histopathology WSI analysis, recent works [34–37] have designed the transformer-based MIL methods for WSI classification, which focus on the global correlation exploration of the patches within a WSI. Shao et al. [34] firstly proposed a transformer-based MIL (TransMIL) method, which employs the multi-head self-attention module to learn the potential relationships among the patches and enhance the context-aware ability with a pyramid position encoding generator (PPEG) module. Reisenbuchler et al. [35] presented a local attention graph-based Transformer for MIL (LAMIL) method, which selects the K nearest neighbors for each instance and then calculates the self-attention scores with these neighboring instances to model the relative relationships among the patches. Zhao et al. [36] developed a novel spatial encoding transformer-based MIL (SETMIL) method, which leverages the spatial encoding transformer to update instance representations by simultaneously aggregating neighboring and globally correlated instances. Ding et al. [37] proposed a novel multi-scale prototypical transformer (MSPT) for WSI classification, which efficiently integrates multi-scale information into the prototypical transformer and thus achieves the multi-scale feature fusion. Obviously, the aforementioned methods generate more powerful WSI-level feature representation and thus have gained superior WSI classification results than the traditional bag-based methods, since the intrinsic relationships among the patches are explored.

2.2. Position encoding in MIL

In the early stage of MIL, the spatial relationships among the instances were often overlooked. Recently, ViT [33] separately performs position encoding on the horizontal and vertical coordinates of each patch and subsequently concatenates these embeddings to form the final positional embeddings. As a result, the spatial structural information among the patches may be partially lost, rendering it unsuitable for fine-grained WSI classification tasks. In addition, several transformer-based MIL methods [34–37] have been proposed, which implicitly or explicitly consider the spatial relationship exploration of the patches and design different position encoding strategies based on the transformer architecture. TransMIL [34] reshapes the feature sequence of the patches into a fixed-size square feature map. Consequently, it fails to describe the real positional relationships of the patches and does not consider the impact of the diverse shapes of the tissue regions within the WSI on the feature representation. LAMIL [35] integrates K-Nearest Neighbor (KNN) graph and transformer architecture for modeling the patch spatial relationships. However, it is difficult to define the optimal number of neighbors to characterize the local morphological structure for different fine-grained downstream tasks (e.g. cancer subtyping and gene mutation prediction). SETMIL [36] simultaneously considers the absolute and relative position encoding. Same with ViT, the absolute position encoding uses 1D sequential position which easily leads to the part spatial information lost. Besides, the relative spatial relationships are introduced as a bias term involved in the self-attention mechanism. However, the measurement of relative relationships is derived from the 2D coordinates in the feature map. Consequently, it may fail to reflect the original spatial morphological structure of the tissue regions and the spatial-semantic consistency may be influenced and limit the performance of fine-grained WSI classification.

Through the above discussions, we still use the transformer structure to explore the spatial relationships which is beneficial to improve the representational ability of WSI-level features and the performance of WSI classification. Different from the aforementioned works, we utilize a position encoding module to obtain the spatial-aware embeddings for each patch. The spatial-aware embeddings are then fused with the semantic features of the patches, and input into the multi-head self-attention module to learn spatial and semantic relationships among patches. More importantly, we adopt a mask-based position reconstruction auxiliary task to enhance the spatial-semantic consistency and generalization capability of the spatial-semantic fused features.

3. Methodology

3.1. Overview

The framework of the proposed method is shown in Fig. 1. Firstly, a given WSI is split into the patches and their corresponding features can be gained. Then, the normalized coordinates are inputted into the position encoding (PE) module to obtain the spatial embeddings. After that, the spatial embeddings are fused with the patch semantic features and fed into the multi-head self-attention (MHSA) module for learning the spatial and semantic dependencies among the patches. The spatial-semantic fused tokens generated by the MHSA module are pooled into the WSI-level representation for classification. Particularly, the position decoding (PD) module is applied to preserve the spatial-semantic consistency, which masks partial tokens for the spatial coordinate reconstruction. The position reconstruction loss performed on the true positions and decoded positions of the patches and the cross-entropy loss for WSI-level features are jointly used for model training.

3.2. Pre-processing and feature extraction

The high resolution of WSIs makes them unsuitable to be directly inputted into the neural network for training.

To address this problem, a window sliding strategy is employed to divide a single WSI into non-overlapping fixed-size patches. Then, the tissue mask image is generated using the OTSU [38] threshold segmentation method to remove background patches. Therefore, a WSI can be represented as $\mathbf{X} = \{(x_1, p_1), (x_2, p_2), \cdots, (x_n, p_n)\}$, where $x_i \in \mathbb{R}^{s\times s\times 3}$, $p_i \in \mathbb{R}^2$, n is the number of the patches within a WSI, x_i represents the i-th patch, s refers to the size of the patch, and p_i denotes the corresponding center position coordinates of x_i. All the patch features are extracted by a pre-trained CNN network. As a result, a WSI can be represented as a feature matrix $\mathbf{F} \in \mathbb{R}^{n\times d}$, where d is the feature dimension.

3.3. Position coordinate normalization

To facilitate network calculation and training convergence, the input position coordinate matrix $\mathbf{P}$ needs to be normalized, where $\mathbf{P} = [p_1, p_2, \cdots, p_n] = [(r_1, c_1), (r_2, c_2), \cdots, (r_n, c_n)] \in \mathbb{R}^{n\times 2}$. The maximum and minimum values of the row coordinate vector $\mathbf{P}_r$ and column coordinate vector $\mathbf{P}_c$ are calculated from all the coordinates, which can be used to obtain the height h and width w of the entire tissue. Then, the maximum value of h and w is used for the scale λ of coordinate transformation to normalize the coordinates. Finally, the normalized position coordinates $\mathbf{P}^{'} \in \mathbb{R}^{n\times 2}$ are used as the input for the PE module. The entire process of coordinate normalization is given in Algorithm 1. The normalized coordinates can enhance network convergence and mitigate biases caused by the size differences of tissue regions in WSIs.

Algorithm 1: Position coordinate normalization

Input: The original position coordinates $\mathbf{P}$, which can be decomposed into two vectors corresponding to the row and column, $\mathbf{P}_r = [r_1, r_2, \cdots, r_n] \in \mathbb{R}^n$ and $\mathbf{P}_c = [c_1, c_2, \cdots, c_n] \in \mathbb{R}^n$.

Output: The normalized position coordinates $\mathbf{P}^{'}$.

1) Calculate the maximum and minimum values of $\mathbf{P}_r$ and $\mathbf{P}_c$:

$r_{min}, r_{max}, c_{min}, c_{max} \leftarrow \min\mathbf{P}_r, \max\mathbf{P}_r, \min\mathbf{P}_c, \max\mathbf{P}_c$

2) Calculate the scale λ of coordinate transformation:

$h \leftarrow r_{max} - r_{min},\ w \leftarrow c_{max} - c_{min},\ \lambda \leftarrow \max\{h, w\}$

3) Normalize the coordinates:

for $i \longleftarrow 1$ **to** n **do**

$$r_i^{'}, c_i^{'} \leftarrow \frac{r_i - r_{min}}{\lambda}, \frac{c_i - c_{min}}{\lambda}$$

end

$\mathbf{P}^{'} = [(r_1^{'}, c_1^{'}), (r_2^{'}, c_2^{'}), \cdots, (r_n^{'}, c_n^{'})]$

return $\mathbf{P}^{'}$

3.4. Spatial position encoding

To effectively capture the spatial relationships among the patches, the normalized 2D position coordinates $\mathbf{P}^{'}$ are encoded into the spatial embeddings $\mathbf{F}_{pos}$ using the PE module, which consists of a fully connected (FC) layer, a MHSA module, and layer normalization (LN) as displayed in Fig. 1c. Similar to TransMIL [34], the Nyströmformer [39] is adopted as the MHSA module, as shown in Fig. 1b. It utilizes the Nystrom method to approximate self-attention and reduces the final computational complexity from $O(n^2)$ to $O(n)$. Through the MHSA module, the spatial embedding of each patch preserves its own positional information and simultaneously exploits the positional dependencies among all the patches. The operation can be described as follows:

$$\mathbf{F}_{pos} = \mathrm{LN}\left(\mathrm{MHSA}\left(\mathbf{P}^{'}\mathbf{W}_{PE}\right)\right) \tag{1}$$

where $\mathbf{W}_{PE} \in \mathbb{R}^{2\times d_p}$ denotes a learnable parameter matrix and d_p is the dimension of the spatial embedding.

3.5. WSI-level feature generation and classification

To embed spatial embeddings into the semantic features of patches, the spatial embeddings are concatenated with the patch semantic features to obtain the spatial-semantic fused features $\mathbf{F}^{'} \in \mathbb{R}^{n\times(d+d_p)}$. Here, the MHSA module is employed to learn the spatial and semantic relationships among the patches, resulting in the spatial-semantic fused tokens $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n] \in \mathbb{R}^{n\times d_m}$, where $\mathbf{h}_i$ represents the spatial-semantic fused token of the i-th patch within the WSI $\mathbf{X}$. The process is given as follows:

$$\begin{aligned} \mathbf{F}^{'} &= \mathrm{Concat}(\mathbf{F}, \mathbf{F}_{pos})\mathbf{W}_{feat} \\ \mathbf{A}, \mathbf{H} &= \mathrm{MHSA}(\mathrm{LN}(\mathbf{F}^{'})) \end{aligned} \tag{2}$$

where $\mathrm{Concat}(\cdot)$ denotes the concatenation operation, $\mathbf{W}_{feat} \in \mathbb{R}^{(d+d_p)\times d_m}$ is a learnable transformation matrix, d_m refers to the dimension of the spatial-semantic fused features, and $\mathbf{A} \in \mathbb{R}^n$ represents the attention scores of n patches and is used for visualization analysis of attention maps (Section 4.6.1 for more details).

Similar to conventional bag-based MIL methods, the spatial-semantic fused tokens are pooled to generate a WSI-level feature representation $\mathbf{h}_{slide} \in \mathbb{R}^{d_m}$ for classification, as shown in Eq. (3)

$$\begin{aligned} \mathbf{h}_{sldie} &= \mathrm{Pool}(\mathbf{H}) \\ \mathbf{p}_{slide} &= \sigma(\mathbf{h}_{slide}\mathbf{W}_{slide}) \end{aligned} \tag{3}$$

where $\mathrm{Pool}(\cdot)$ is a pooling operation, $\mathbf{W}_{slide} \in \mathbb{R}^{d_m\times c}$ is a learnable parameter matrix for linear transformation , c is the number of the categories, $\sigma(\cdot)$ is denoted as the softmax function, and $\mathbf{p}_{slide} \in \mathbb{R}^c$ is the predicted probabilities corresponding to each class. The cross-entropy loss is utilized for WSI-level classification, which is formulated as:

$$\mathcal{L}_{slide} = -\mathbf{y}\log\left(\mathbf{p}_{slide}\right) \tag{4}$$

where $\mathbf{y}$ is the one-hot ground truth of the WSI $\mathbf{X}$.

3.6. Mask-based position reconstruction

The conventional transformer-based MIL methods typically directly employ the position encoding but ignore the

spatial-semantic consistency of the patch features. Inspired by Masked AutoEncoder (MAE) [40], we introduce the mask-based position reconstruction as an auxiliary task, which aims to guarantee the decoded spatial coordinates and the true coordinates of the patches are as consistent as possible through the mask strategy and thus improve the spatial-semantic consistency and generalization of the patch features. As shown in Fig. 1a, the mask-based position reconstruction is achieved through the PD module, which consists of an MHSA module, a LN layer, a FC layer, and a Sigmoid layer, as displayed in Fig. 1d. Specifically, we mask the spatial-semantic fused tokens $\mathbf{H}$ by replacing them with a learnable token $\mathbf{h}_{mask}$ through a fixed mask ratio r_{mask}. The mask operation can be formatted as below:

$$\begin{aligned} \mathbf{H}_{mask} &= \left[\mathbf{h}_1^m, \mathbf{h}_2^m, \cdots, \mathbf{h}_n^m\right] \leftarrow \mathbf{H} = \left[\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n\right] \\ \mathbf{h}_i^m &= \begin{cases} \mathbf{h}_{mask} & \text{if } i \in \text{MaskIDs}(n * r_{mask}, n) \\ \mathbf{h}_i & \text{else} \end{cases} \end{aligned} \tag{5}$$

where MaskIDs(m, n) is a function used to randomly select $\lfloor m \rfloor$ indices from 1 to n and $\mathbf{H}_{mask} \in \mathbb{R}^{n \times d_m}$ denotes the tokens processed by mask operation. Then, $\mathbf{H}_{mask}$ is inputted into the PD module to predict the coordinates $\mathbf{P}_{pred} = \left[(r_1'', c_1''), (r_2'', c_2''), \cdots, (r_n'', c_n'')\right] \in \mathbb{R}^{n \times 2}$ of all the patches. The entire process can be depicted:

$$\mathbf{P}_{pred} = \text{Sigmoid}(\text{LN}\left(\text{MHSA}\left(\mathbf{H}_{mask}\right)\right)\mathbf{W}_{PD}) \tag{6}$$

where $\mathbf{W}_{PD} \in \mathbb{R}^{d_m \times 2}$ is a learnable parameter matrix and Sigmoid$(\cdot)$ denotes the sigmoid activation function. The Mean Squared Error (MSE) loss $\mathcal{L}_{pos}$ is used to measure the position coordinate reconstruction error between $\mathbf{P}'$ and $\mathbf{P}_{pred}$ in Eq. (7)

$$\mathcal{L}_{pos} = \tau \times \frac{1}{n} \sum_{i}^{n} \sqrt{(r_i'' - r_i')^2 + (c_i'' - c_i')^2} \tag{7}$$

where τ is a scaling parameter with a default value of 100. Finally, the entire framework is trained end-to-end based on the composite objective function:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{slide} + (1 - \alpha)\mathcal{L}_{pos} \tag{8}$$

where α is the weight of $\mathcal{L}_{slide}$.

4. Experiment and Result

4.1. Datasets

Our proposed method is evaluated on two public benchmark TCGA datasets (TCGA-LUNG and TCGA-BRCA)[1] and two in-house clinical datasets (USTC-EGFR and USTC-GIST)[2] which come from the First Affiliated Hospital of USTC (University of Science and Technology of China). TCGA-LUNG and TCGA-BRCA are used for lung and breast cancer subtyping. USTC-EGFR is applied for gene mutation prediction of the EGFR gene in non-small cell lung cancer (NSCLC). USTC-GIST is employed for gene mutation prediction in gastrointestinal stromal tumor (GIST). Notably, the gene mutation task in our experiment aims to predict the fine-grained gene mutation status for EGFR and GIST through H&E-stained WSIs, which is more convenient and can effectively reduce costs compared with the traditional gene sequencing methods. The detailed profiles of the three datasets are presented below.

Table 1
The WSI Distribution of the four Datasets.

TCGA-LUNG	Normal	LUAD	LUSC		
Train	385	326	333		
Test	165	141	144		
TCGA-BRCA	IDC	ILC			
Train	555	142			
Test	239	62			
USTC-EGFR	Neg	L858R	19del	Wild	Others
Train	117	80	137	99	98
Test	48	38	47	47	43
USTC-GIST	Neg	Wild	Exon 9[1]	Exon 11[2]	Others
Train	17	52	53	296	44
Test	9	16	11	124	21

[1] Exon 9: KIT gene exon 9.
[2] Exon 11: KIT gene exon 11.

- **TCGA-LUNG** contains 1494 slides from 1345 cases for lung cancer subtyping from the TCGA program. This dataset includes three categories: non-cancerous tissue (Normal), Lung Squamous Cell Carcinoma(LUSC), and Lung Adenocarcinoma (LUAD).
- **TCGA-BRCA** contains 998 cases and each case corresponds to a single slide. It has two categories of breast cancer: Invasive ductal (IDC) and Invasive lobular carcinoma (ILC).
- **USTC-EGFR** contains 754 slides from 521 cases. This dataset has two common EGFR mutation types: EGFR exon 19 deletion (19del) and a missense mutation in exon 21 (L858R). The 19del and L858R mutations are the most prevalent actionable alterations, accounting for 40% and 45% of NSCLC driver mutations, respectively [42]. Accurate classification for these two categories is of great clinical significance for guiding personalized therapeutic strategies. Besides, we collect another three categories, negative (Neg), the wild type (Wild), and other driver mutation types (Others).
- **USTC-GIST** contains 643 slides from 116 cases. GIST is the most common sarcoma of the gastrointestinal (GI) tract. It is a rare neoplasm and the reported incidence varies significantly ranging from 0.4 to 2 cases per 100,000 individuals per year [43]. Particularly, its incidence in China

[1] TCGA-LUNG and TCGA-EGFR were obtained from The Cancer Genome Atlas portal (https://portal.gdc.cancer.gov/) [41].

[2] The study was approved by the Medical Research Ethics Committee of the First Affiliated Hospital of the University of Science and Technology of China (Anhui Provincial Hospital) under the protocols No. 2022-RE-454 and No. 2024KY-009.

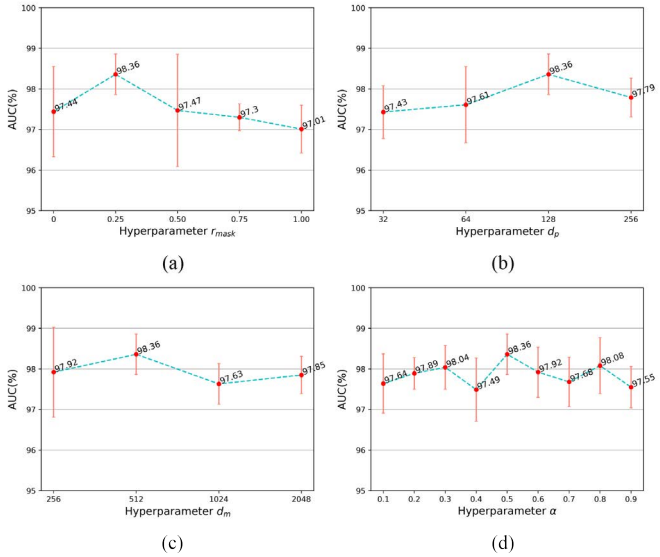


Fig. 2: The performance curves of tuning four hyper-parameters on the TCGA-LUNG dataset, which (a) is the mask ratio r_{mask}, (b) is the dimension of the spatial embedding d_p, (c) is the dimension d_m of the spatial-semantic fused feature, and (d) the weight α of $\mathcal{L}_{slide}$.

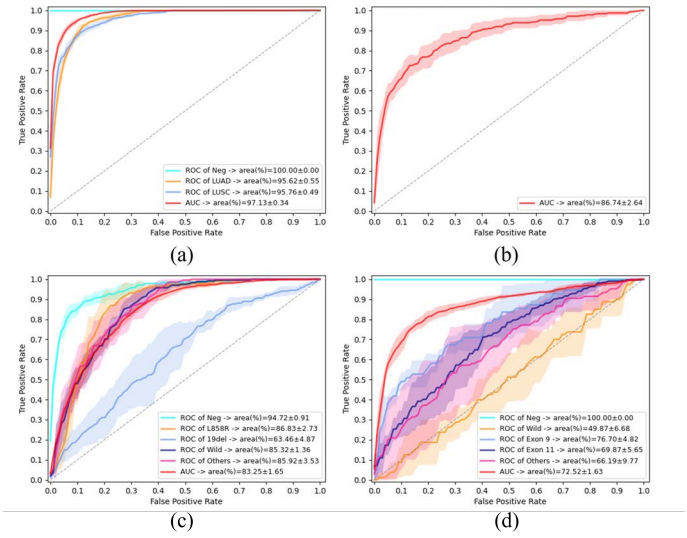


Fig. 3: The ROC curves of PEGTB-MIL on the TCGA-LUNG, TCGA-BRCA, USTC-EGFR, and USTC-GIST datasets are shown in (a), (b), (c), and (d), respectively.

has been increasing year by year, which is only lower than gastric cancer and colorectal cancer in gastrointestinal tumors. Approximately 75-80% of GIST have mutations in the KIT gene [44]. Imatinib, a multi-targeted tyrosine kinase inhibitor (TKI) against KIT, PDGFRA, and BCR-ABL, is the advanced therapy for unresectable or metastatic GIST. However, different imatinib dosage strategies are crucial for the specific KIT gene mutations, such as KIT gene exon 9 and KIT gene exon 11 mutations [45]. Therefore, the ability to identify KIT mutation status in advance through H&E-stained WSIs has clinical importance for guiding optimal treatment pathways in GIST patients. This dataset has two common GIST mutation types: KIT gene exon 9 and KIT gene exon 11. Besides, we collect another three categories, negative (Neg), the wild type (Wild), and other driver gene mutation types (Others).

A comprehensive summary of these datasets is presented in Table 1. TCGA-LUNG and TCGA-BRCA are publicly available datasets with labels confirmed from the analysis of official clinical information. The mutation types in USTC-EGFR and USTC-GIST datasets are obtained through molecular sequencing and organized by pathologists. We divide the four datasets into train and test sets with a ratio of 7:3 at the patient level. The train set is used for model training and hyper-parameter verification (Section 4.3 for more details). Finally, we evaluate the performance of the model in the test set.

4.2. Implementation details

Before training, all the slides are divided into patches of size 256×256 using the sliding window strategy at 20× magnification. The OTSU [38] segmentation method is employed to generate a tissue mask image to extract foreground patches and remove background patches. Then, the 1024-dimensions features of the patches are extracted by a standard ResNet50 [46] pre-trained in the ImageNet, which leverages the representations learned from a large number of images, enhancing its ability to capture complex patterns and features relevant to the specific tasks. During the training stages of PEGTB-MIL, the Adam optimizer is employed with an initial learning rate of 5e-4 and weight decay of le-5. The size of the mini-batch is 1. The initial values of r_{mask}, d_p, d_m, and α are set to 0.25, 128, 512, and 0.5, respectively. These four hyper-parameters are selected in the validation stage. The accuracy of classification (ACC), the area under the receiver operating characteristic curve (AUC), and F1 score are served as the metric of performance evaluation. The following experimental results are reported by five-fold cross-validation. All the experiments are conducted on one computer with an AMD Ryzen Threadripper 3960X 24-Core Processor CPU and a NVIDIA RTX 3090 GPU. Our codes are available at https://github.com/HFUT-miaLab/PEGTB-MIL.

4.3. Hyper-parameter verification

We conduct experiments on the TCGA-LUNG dataset to investigate the effects of four hyper-parameters on the performance of PEGTB-MIL. The four hyper-parameters are as follows: (1) the mask ratio r_{mask}, (2) the dimension d_p of spatial embedding, (3) the dimension d_m of the spatial-semantic fused features, and (4) the weight α of $\mathcal{L}_{slide}$. Here, we sequentially tune r_{mask}, d_p, d_m, and α. The optimal values of these four hyper-parameters on the TCGA-LUNG dataset are all selected according to the AUC results in the validation stage. The detailed results are shown in Fig. 2.

Table 2
Ablation study of PEGTB-MIL on the TCGA-LUNG dataset.

Settings	Position encoding strategy	Evaluation metrics		
		ACC (%)	AUC (%)	F1 (%)
(A)	None	85.11 ± 2.43	95.89 ± 0.36	83.94 ± 3.10
(B)	1D-Embedding	83.33 ± 4.03	96.00 ± 0.59	81.42 ± 5.75
(C)	2D-Embedding	87.96 ± 2.49	96.88 ± 0.62	87.13 ± 2.88
(D)	PE w/o Normalization	86.53 ± 2.63	96.56 ± 0.24	85.52 ± 3.09
(E)	PE	88.58 ± 1.58	97.04 ± 0.55	87.94 ± 1.67
(F)	PE+PD (ours)	$\mathbf{89.16 \pm 1.00}$	$\mathbf{97.13 \pm 0.34}$	$\mathbf{88.55 \pm 1.07}$

Table 3
Results of various MIL methods on two public benchmark TCGA-LUNG and TCGA-BRCA datasets.

Methods	**TCGA-LUNG** ($cls = 3$)			**TCGA-BRCA** ($cls = 2$)		
	ACC (%)	AUC (%) [2.5% CI, 97.5% CI]	F1 (%)	ACC (%)	AUC (%) [2.5% CI, 97.5% CI]	F1 (%)
ABMIL	$87.82 \pm 0.45^{\dagger}$ [86.49, 89.11]	$95.66 \pm 0.30^{\ddagger}$ [94.95, 96.27]	$87.16 \pm 0.48^{\dagger}$ [85.76, 88.42]	$79.53 \pm 2.81^{\ddagger}$ [77.41, 81.66]	$75.91 \pm 6.11^{\dagger}$ [72.56, 79.03]	$66.19 \pm 5.11^{\ddagger}$ [63.07, 69.08]
CLAM	88.18 ± 0.90 [86.80, 89.64]	96.83 ± 0.25 [96.25, 97.40]	87.44 ± 1.08 [85.96, 88.87]	$82.92 \pm 1.83^{\dagger}$ [80.93, 84.65]	83.80 ± 3.81 [81.06, 86.35]	74.83 ± 2.28 [72.02, 77.19]
TransMIL	87.24 ± 1.79 [85.82, 88.67]	$96.16 \pm 0.43^{\ddagger}$ [95.56, 96.75]	86.48 ± 1.99 [85.02, 87.85]	$82.46 \pm 2.31^{\dagger}$ [80.33, 84.25]	86.07 ± 0.97 [84.03, 88.97]	74.16 ± 0.94 [71.12, 76.68]
LAMIL	$85.91 \pm 2.05^{\dagger}$ [84.49, 87.24]	$96.07 \pm 0.28^{\ddagger}$ [95.43, 96.66]	$84.88 \pm 2.44^{\dagger}$ [83.34, 86.24]	$79.14 \pm 5.58^{\dagger}$ [77.08, 81.06]	84.61 ± 3.56 [81.99, 87.25]	72.94 ± 4.78 [70.11, 75.43]
SETMIL	$83.87 \pm 0.95^{\ddagger}$ [82.36, 85.38]	$94.54 \pm 0.20^{\ddagger}$ [93.86, 95.25]	$82.96 \pm 1.72^{\ddagger}$ [81.39, 84.39]	$70.10 \pm 5.36^{\ddagger}$ [67.77, 72.49]	$80.74 \pm 1.59^{\ddagger}$ [78.04, 83.32]	$64.57 \pm 3.87^{\dagger}$ [61.76, 67.22]
PEGTB-MIL	$\mathbf{89.16 \pm 1.00}$ [87.91, 90.44]	$\mathbf{97.13 \pm 0.34}$ [96.63, 97.61]	$\mathbf{88.55 \pm 1.07}$ [87.20, 89.82]	$\mathbf{86.31 \pm 1.80}$ [84.52, 88.11]	$\mathbf{86.74 \pm 2.64}$ [84.01, 89.13]	$\mathbf{76.71 \pm 3.31}$ [73.79, 79.41]

†: p-value < 0.05.
‡: p-value < 0.01.

(1) *Mask ratio* r_{mask}: r_{mask} controls the ratio of the masked tokens in the spatial-semantic fused tokens $\mathbf{H}$. We tune the value of r_{mask} within the range of [0, 0.25, 0.5, 0.75, 1]. As shown in Fig. 2a, we can observe that the optimal value of r_{mask} is 0.25 on the TCGA-LUNG dataset. Notably, there is a significant decrease when $r_{mask} = 1$. It has demonstrated that the PD module ($r_{mask} \leq 1$) can have a positive impact on the model performance.

(2) *The dimension* d_p *of spatial embedding*: d_p controls the dimensionality of the spatial embedding. Experimentally, we tune $d_p \in [32, 64, 128, 256]$ for verification. The optimal value for d_p is 128 on the TCGA-LUNG dataset, as can be observed in Fig. 2b. The lower dimensions (e.g.,d_p = 32 or 64) result in a loss of representational capacity for complex spatial relationships, while higher dimensions exhibit redundancy and overfitting issues (e.g.,$d_p = 256$).

(3) *The dimension* d_m *of the spatial-semantic fused features*: After incorporating the patch semantic features and their corresponding spatial embeddings, a transformation matrix $\mathbf{W}_{feat}$ is utilized to map the dimension of the concatenated features to d_m, resulting in the spatial-semantic fused features. We tune d_m within the range of $[256, 512, 1024, 2048]$. As presented in Fig. 2c, the optimal value of d_m on the TCGA-LUNG dataset is 512. This explanation indicates that a lower dimension of fused features (e.g., $d_m = 256$) can lead to decreased learning ability of the model, while higher dimensions (e.g., d_m= 1024 or 2048) can make the model more difficult to converge.

(4) *The weight* α *of* $\mathcal{L}_{slide}$: α controls the contributions of $\mathcal{L}_{slide}$ and $\mathcal{L}_{pos}$. We test it in the range of [0.1, 0.9] with a step of 0.1. As depicted in Fig. 2d, it is clear that the optimal value of α on the TCGA-LUNG dataset is 0.5, which also demonstrates that balanced $\mathcal{L}_{slide}$ and $\mathcal{L}_{pos}$ can better guide the model training.

4.4. Ablation study

To verify the effectiveness of the proposed position encoding (PE) module and mask-based position reconstruction (PD module), we conduct the ablation study on the TCGA-LUNG dataset. The results are shown in Table 2. In detail, the settings are as follows: (A) represents the network without position encoding and decoding; (B) indicates the network only with Transformer-1D [32] position encoding; (C) infers the network only with ViT-2D [33] position encoding; (D) symbolizes the network with the proposed PE module, the coordinates are not normalized; (E) denotes the network only with the proposed PE module, the coordinates

Table 4
Results of various MIL methods on two in-house clinical USTC-EGFR and USTC-GIST datasets.

Methods	USTC-EGFR ($cls = 5$)			USTC-GIST ($cls = 5$)		
	ACC (%)	AUC (%) [2.5% CI, 97.5% CI]	F1 (%)	ACC (%)	AUC (%) [2.5% CI, 97.5% CI]	F1 (%)
ABMIL	$44.84 \pm 1.96^{\ddagger}$ [42.24, 47.62]	$76.20 \pm 0.75^{\ddagger}$ [74.49, 77.88]	$40.01 \pm 2.84^{\dagger}$ [37.49, 42.21]	$61.66 \pm 7.93^{\ddagger}$ [57.77, 63.71]	70.89 ± 2.35 [66.39, 73.10]	$36.84 \pm 10.89^{\dagger}$ [33.14, 40.29]
CLAM	$47.17 \pm 0.59^{\dagger}$ [44.39, 49.96]	$79.23 \pm 1.01^{\ddagger}$ [77.63, 80.84]	$44.75 \pm 1.33^{\dagger}$ [41.96, 47.14]	$60.22 \pm 3.86^{\dagger}$ [58.40, 64.39]	70.94 ± 1.68 [67.98, 74.06]	42.36 ± 1.70 [39.01, 45.47]
TransMIL	48.16 ± 2.58 [45.29, 51.12]	81.45 ± 1.05 [79.99, 82.94]	45.46 ± 5.57 [42.52, 47.85]	$53.15 \pm 13.77^{\ddagger}$ [50.14, 56.39]	70.57 ± 3.70 [67.53, 73.54]	41.86 ± 4.77 [38.61, 45.16]
LAMIL	$46.82 \pm 2.76^{\ddagger}$ [43.95, 49.78]	$80.82 \pm 0.96^{\dagger}$ [79.43, 82.14]	$42.20 \pm 3.88^{\dagger}$ [39.61, 44.48]	$44.09 \pm 10.83^{\ddagger}$ [41.20, 48.17]	$65.57 \pm 3.35^{\dagger}$ [62.84, 68.11]	36.26 ± 8.80 [34.41, 38.19]
SETMIL	$48.16 \pm 2.48^{\ddagger}$ [45.20, 51.21]	81.79 ± 1.71 [80.41, 83.15]	$42.74 \pm 2.73^{\ddagger}$ [40.11, 44.97]	$53.92 \pm 5.95^{\ddagger}$ [50.38, 57.24]	69.39 ± 2.71 [66.73, 72.02]	39.05 ± 1.85 [36.20, 41.36]
PEGTB-MIL	**52.47 ± 3.77** [49.68, 55.52]	**83.25 ± 1.65** [81.93, 84.62]	**51.40 ± 4.10** [48.43, 54.00]	**65.52 ± 3.76** [63.75, 69.83]	**72.52 ± 1.63** [69.71, 74.97]	**45.28 ± 2.63** [41.96, 47.00]

†: p-value < 0.05.
‡: p-value < 0.01.

are normalized; (F) depicts complete PEGTB-MIL and no ablation is performed.

For the analysis of detailed results, (A) has an obvious decline in terms of three metrics compared to (C) and (E). It has indicated that learnable position encoding can effectively enhance the performance of the model. (B) exhibits poorer performance, possibly because the Transformer-1D encoding method is not learnable and cannot be optimized. Note that (E) surpasses (C) by 0.62% in ACC, 0.16% in AUC, and 0.81% in F1. It can be explained that the proposed PE module can express richer position information compared to ViT-2D, which encodes rows and columns independently. To validate the effectiveness of the proposed PD module, we compare the experimental results of (E) and (F). When the PD module is removed, we can observe that the performance of the model has declined, with ACC decreasing by 0.58%, AUC by 0.09%, and F1 by 0.61%. It has been demonstrated that the mask-based position reconstruction task effectively preserves the spatial-semantic consistency of features and improves the classification performance of the model. Additionally, in the proposed PEGTB-MIL, the input of the PE module needs to be normalized. Comparing with the experimental results of (D) and (E), (D) shows a significant performance drop, with ACC dropping by 2.05%, AUC by 0.48%, and F1 by 2.42%. The reason may be that the unnormalized input prevents the model from converging properly in the training stage.

To strengthen above conclusions, we present the results of the ablation studies conducted on other three datasets: TCGA-BRCA, USTC-EGFR, and USTC-GIST. The results can be seen in Table A1 in the appendix. The conclusions obtained from the results on the other three datasets are highly consistent with those from the TCGA-LUNG dataset. Specifically, compared to the better-performing setting (B) method, our method achieves increases of 4.33% in ACC, 3.58% in AUC, and 3.07% in F1 on the TCGA-BRCA dataset. On the USTC-EGFR dataset, we observe increases of 2.27% in ACC, 3.06% in AUC, and 5.75% in F1. For the USTC-GIST dataset, the improvements are even more notable, with ACC increasing by 7.96%, AUC by 2.04%, and F1 by 4.29%. These results demonstrate that the proposed positional encoding method is critical for WSI analysis and outperforms other positional encoding techniques. It also highlights the importance and effectiveness of the proposed position reconstruction auxiliary task in maintaining spatial-semantic consistency.

4.5. Comparative experiments

We compare our method with five most representative MIL-based methods, including ABMIL [24], CLAM [2], TransMIL [34], LAMIL [35], and SETMIL [36]. The experiments for these methods are conducted using official code and the same experimental settings. The mean and standard deviation results on the test set are presented in Tables 2-3, where the special indicators † and ‡ represent p-value < 0.05 and p-value < 0.01, respectively. Additionally, we evaluate the uncertainty of the results through including the 2.5 and 97.5 percentile confidence intervals (CI), which are obtained from 1000 bootstrapping iterations.

Overall, the proposed PEGTB-MIL achieves the best performance with ACC, AUC, and F1 of 89.16%, 97.13%, and 88.55% on the TCGA-LUNG dataset, 86.31%, 86.74%, and 76.71% on the TCGA-BRCA dataset, 52.47%, 83.25%, and 51.40% on the USTC-EGFR dataset, and 65.52%, 72.52%, and 45.28% on the USTC-GIST dataset.

In these comparison methods, ABMIL and CLAM are both classic attention-based methods. CLAM uses the instance clustering module to optimize the feature space of patches, which enhances the attention module with improved discriminative capability. Therefore, CLAM outperforms ABMIL on all four datasets, with AUC gains of 1.17% on the TCGA-LUNG dataset, 7.89% on the TCGA-BRCA dataset,

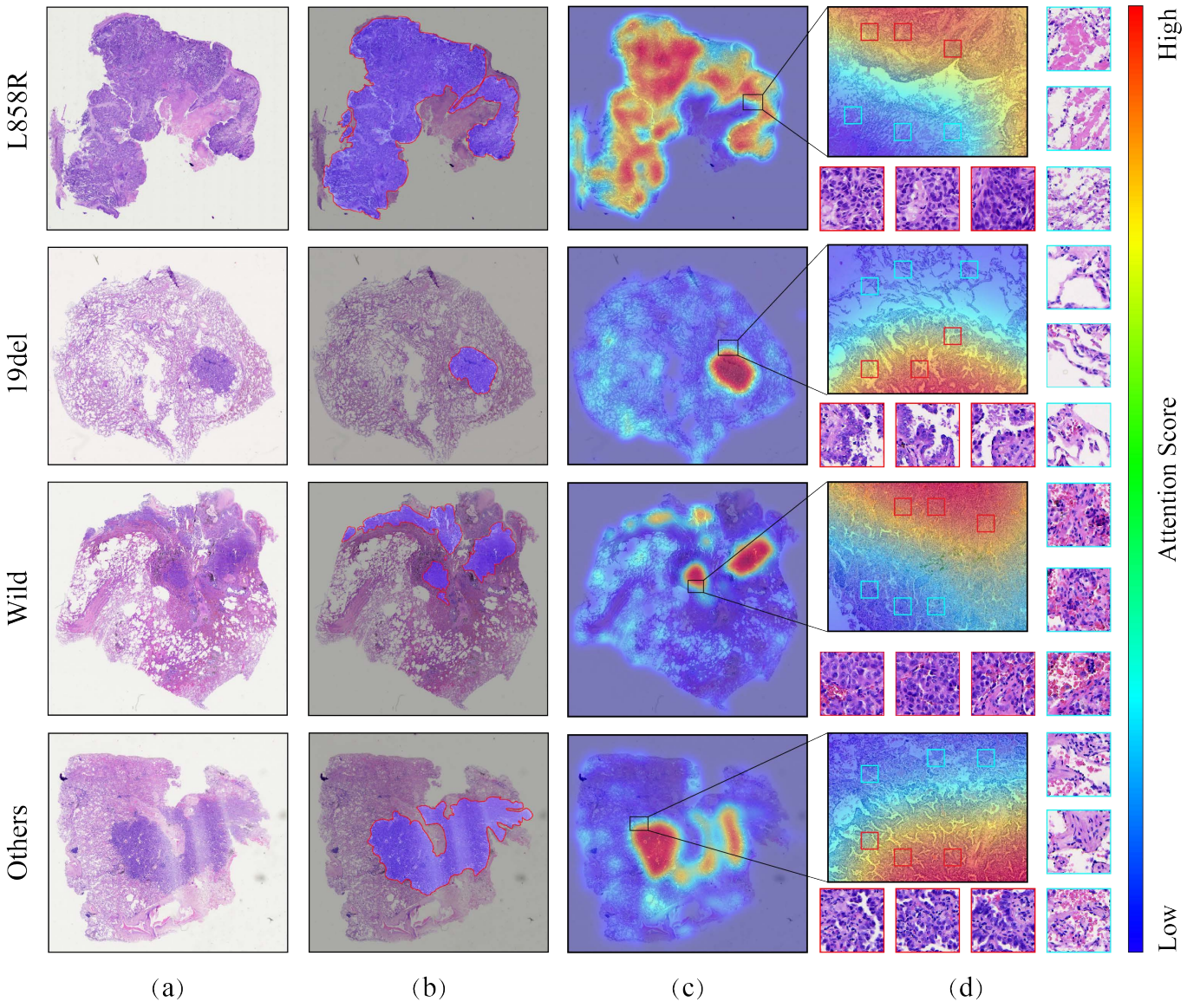


Fig. 4: Visualization based on the attention map in PEGTB-MIL, where (a) displays the thumbnails for each slide, (b) shows the lesion tissue images annotated by pathologists, (c) displays the attention heatmaps and (d) shows the ROI images of the attention heatmaps along with some representative patches.

3.03% on the USTC-EGFR dataset, and 0.05% on the USTC-GIST dataset.

For the transformer-based MIL methods, TransMIL achieves better performance than CLAM on the TCGA-BRCA dataset (AUC is 2.27% higher) and USTC-EGFR dataset (AUC is 2.22% higher), but its AUC is lower than CLAM by 0.67% and 0.37% on the TCGA-LUNG and USTC-GIST datasets, respectively. The reason may be that TransMIL exhibits overfitting on the two datasets (TCGA-LUNG and USTC-GIST). As portrayed in Tables 3-4, LAMIL achieves the moderate performance on the TCGA-LUNG, TCGA-BRCA, and USTC-EGFR datasets. However, LAMIL performs poorly on the USTC-GIST dataset, possibly because this method is sensitive to or unsuitable for this dataset.

Note that SETMIL has a poor performance (ACC, AUC, and F1 are 83.87%, 94.54%, and 82.96%) on the TCGA-LUNG dataset compared to other methods. On the TCGA-BRCA dataset, the results of SETMIL are also unsatisfactory. Compared to the proposed PEGTB-MIL, ACC, AUC, and F1 are lower by 16.21%, 6.00%, and 12.14% respectively. SETMIL introduces 2D positional information as a bias term into the self-attention calculation to learn relative spatial relationships, resulting in an improvement in model performance compared to LAMIL. Therefore, on the USTC-EGFR dataset, SETMIL outperforms LAMIL in ACC, AUC, and F1 by 1.34%, 0.97%, and 0.54% respectively. On the USTC-GIST dataset, LAMIL is lower than SETMIL in ACC, AUC, and F1 by 9.83%, 3.82%, and 2.79% respectively.

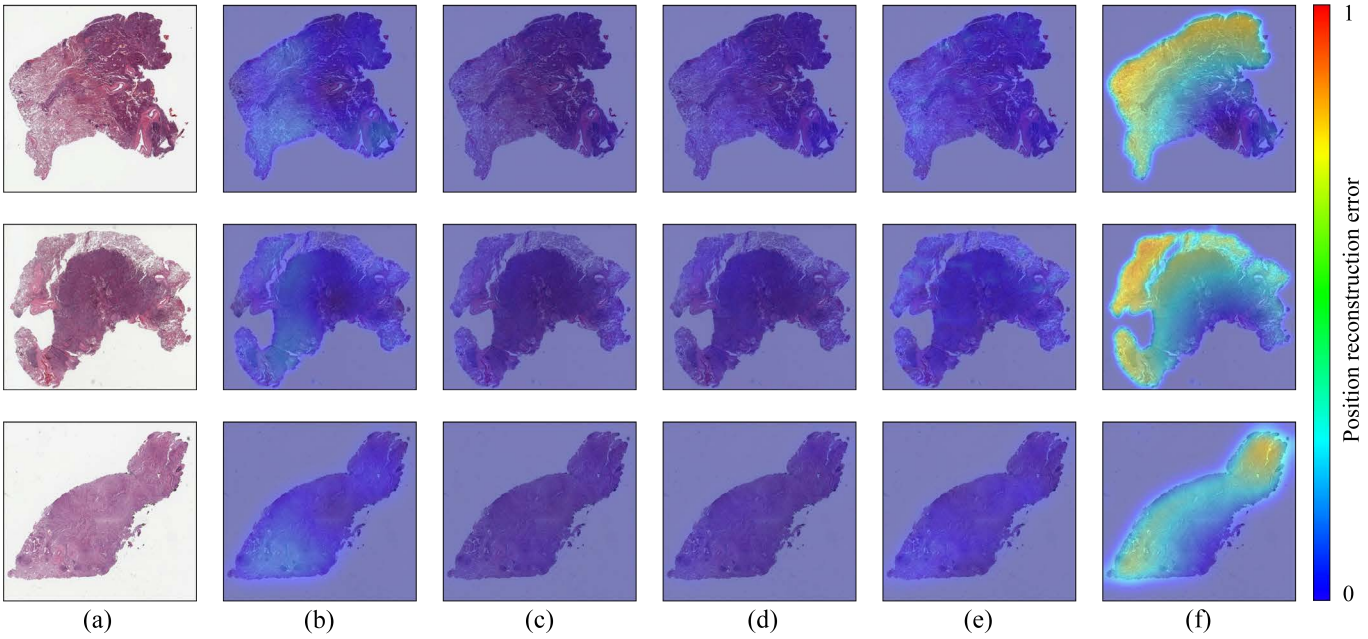


Fig. 5: The visualization of the position reconstruction is as follows: (a) represents the thumbnail of the slide, and (b)-(f) represent the visualizations at r_{mask} of 0, 0.25, 0.50, 0.75, and 1.00 respectively.

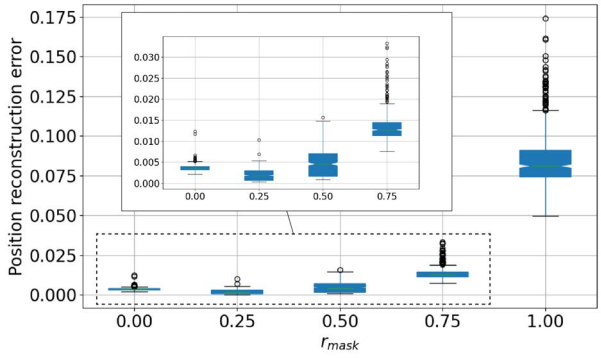


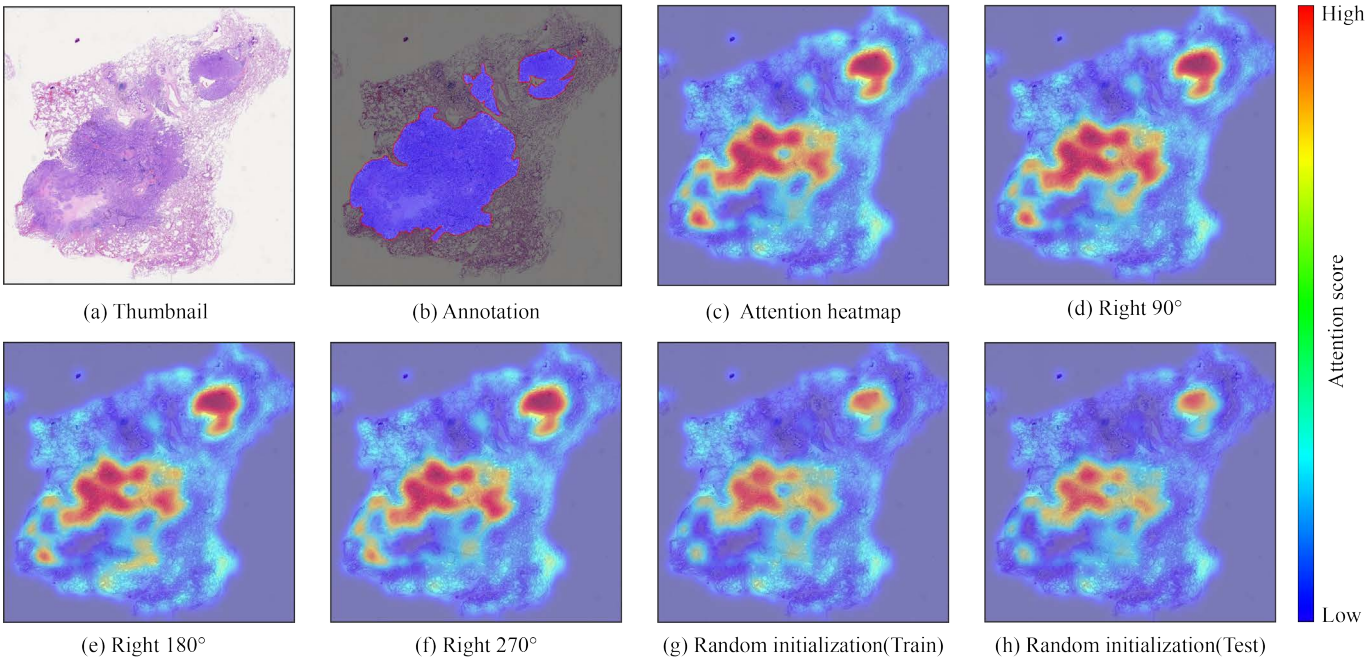
Fig. 6: Box plot of position reconstruction errors on the TCGA-LUNG dataset.

As indicated in Tables 3-4, the proposed PEGTB-MIL achieves 0.98% higher ACC, 0.30% higher AUC, and 1.11% higher F1 than the second-best method CLAM on the TCGA-LUNG dataset. On the TCGA-BRCA dataset, compared to the second-best method TransMIL, PEGTB-MIL improves ACC, AUC, and F1 by 3.85% , 0.67%, and 2.55% respectively. On the USTC-EGFR dataset, it outperforms the second-best method SETMIL by 4.31% (p-value < 0.01) of ACC, 1.46% of AUC, and 8.66% (p-value < 0.01) of F1. On the USTC-GIST dataset, PEGTB-MIL exhibits a 5.30% (p-value < 0.05) higher ACC, a 1.58% higher AUC, and a 2.92% higher F1 compared to the second-best method CLAM. Compared to the transformer-based MIL methods such as TransMIL, LAMIL, and SETMIL, PEGTB-MIL leverages the proposed position encoding and mask-based position reconstruction to effectively enhance the spatial-semantic consistency and the classification performance, especially for predicting gene mutation status.

Additionally, we conduct the ROC curves of PEGTB-MIL in each category of these four datasets in Fig. 3. Fig. 3a shows the ROC curves of PEGTB-MIL on the TCGA-LUNG dataset. We can observe that PEGTB-MIL exhibits the best performance in the Normal category (AUC = 100%), and also shows excellent results in the other two tumor categories (LUAD and LUSC), with AUC greater than 95%. Fig. 3b depicts the ROC curves of PEGTB-MIL on the TCGA-BRCA dataset. For a binary classification task, the proposed PEGTB-MIL achieves an overall AUC of 86.74%. Fig. 3c displays the ROC curves of PEGTB-MIL on the USTC-EGFR dataset. The proposed method has demonstrated efficient performance in identifying the negative category (AUC = 94.72%). In addition, PEGTB-MIL exhibits poorer performance in identifying the 19del mutation compared to other categories, with an AUC of 23.37% lower than the L858R category, 21.86% lower than the Wild category, and 22.46% lower than the Others category. It may be because the 19del category is more challenging to identify compared to other categories. Although PEGTB-MIL exhibits relatively lower performance in identifying the 19del category, it has demonstrated a notable AUC result (AUC = 86.83%) for the L858R category. Fig. 3d exhibits the ROC curves of PEGTB-MIL on the USTC-GIST dataset. For GIST, a rare neoplasm, the overall AUC has gained an acceptable result, 72.52%, and the AUC of each mutation type (Exon 9, Exon 11, and Others) is higher than 65%. It has indicated that PEGTB-MIL has better classification ability under very limited cases.

Table 5
The anti-interference experimental results of PEGTB-MIL on the TCGA-LUNG dataset.

Settings	Transformation	ACC(%)	$\triangle_{ACC}$(%)	AUC(%)	$\triangle_{AUC}$(%)	F1(%)	$\triangle_{F1}$(%)
(A)	RIGHT 90°	**89.60 ± 0.68**	+0.44	96.99 ± 0.33	−0.14	**89.03 ± 0.73**	+0.48
(B)	RIGHT 180°	89.51 ± 0.63	+0.35	97.02 ± 0.29	−0.11	88.95 ± 0.66	+0.40
(C)	RIGHT 270°	88.93 ± 0.74	−0.23	96.92 ± 0.39	−0.21	88.30 ± 0.78	−0.25
(D)	Random initialization (Train)	87.73 ± 1.96	−1.43	96.47 ± 0.60	−0.66	86.98 ± 2.23	−1.57
(E)	Random initialization (Test)	88.58 ± 1.58	−0.58	96.68 ± 0.57	−0.45	87.94 ± 1.67	−0.61
(F)	None	89.16 ± 1.00	——	**97.13 ± 0.34**	——	88.55 ± 1.07	——



(a) Thumbnail (b) Annotation (c) Attention heatmap (d) Right 90°
(e) Right 180° (f) Right 270° (g) Random initialization(Train) (h) Random initialization(Test)

Fig. 7: Visualization based on the attention map in PEGTB-MIL under various transformations, where (a) displays the thumbnail for the slide, (b) shows the lesion tissue image annotated by pathologists, (c) displays the attention heatmap, (d)-(f) represent the attention heatmaps corresponding to respectively rotating the slide 90 degrees to the right, 180 degrees, and 270 degrees, (g) denotes the attention heatmap with random initialization coordinates, and (h) refers the attention heatmap generated by the model with random initialization coordinates in training stage.

4.6. Visualization and discussion

For the WSI classification task, the interpretability analysis can help pathologists better understand the AI-generated results. In this section, we analyze and discuss the interpretability of the proposed PEGTB-MIL. We design two parts: (1) Interpretation with attention heatmaps generated by the attention score of each patch; and (2) Visualization of position reconstruction. In addition, we conduct anti-interference experiments on the proposed PEGTB-MIL, followed by comprehensive analysis and discussion.

4.6.1. Interpretation with attention heatmaps

To investigate the interpretability of the proposed PEGTB-MIL, we generate attention heatmaps based on the attention scores of each patch and its corresponding position information, as shown in Fig. 4. Note that precisely delineating mutation-related tissue regions directly from H&E WSI remains more challenging. Therefore, we show the lesion tissue annotated by pathologists and try to explore the model interpretability for identifying the potential mutation-related sites within the lesion regions. Fig. 4a shows the thumbnails of these slides. Fig. 4b displays the annotations of the tumor region (within the red curve) of each slide. Fig. 4c shows the attention heatmaps for each slide. It can be seen that the regions with high attention scores correspond to the tumor regions and are consistent with the regions annotated by pathologists. It has demonstrated that our method can identify the tumor regions under the weak supervision. Moreover, the model flags these red sites within the regions as potential candidates with driver gene mutations, potentially offering valuable insights for targeted therapies. Fig. 4d shows the regions of interest (ROIs) selected from the attention heatmaps, with some representative patches. The patches with a red border potentially exhibit the cells with gene mutations compared to the normal patches with a cyan border.

4.6.2. Visualization of position reconstruction

To validate the spatial-semantic consistency of the features, we design a position reconstruction visualization as shown in Fig. 5. Specifically, we use the squared error to quantify the position reconstruction error between the predicted coordinates and ground truth coordinates. Fig. 5a shows the thumbnails of the three slides selected from the test set on the TCGA-LUNG dataset. According to the settings of the r_{mask} in Section 4.3, we show the results of position coordinates reconstruction for r_{mask} = 0, 0.25, 0.5, 0.75, and 1 in Figs. 5(b)-(f) respectively. When r_{mask} is set to 1, all the tokens are masked in the spatial-semantic fused tokens $\mathbf{H}_{mask}$, which makes the PD module is difficult to accurately reconstruct the coordinates of the patches and thus randomly predict invalid coordinates, as depicted in Fig. 5f. When r_{mask} is equal to 0, 0.25, 0.50, or 0.75, it is evident that the reconstruction error in Figs. 5(b)-(e) are generally lower, indicating that the position reconstruction is insensitive to changes in the r_{mask} parameter and has strong robustness. Notably, when r_{mask} is equal to 0, the position reconstruction task removes the mask mechanism. From Fig. 5b, it can be observed that the model with $r_{mask} = 0$ exhibits larger position reconstruction errors compared to Fig. 5c. Therefore, it has demonstrated that the mask mechanism can enhance the generalization ability of the PD module in reconstructing coordinates.

4.6.3. Anti-interference ability

To quantitatively analyze the position reconstruction error across the entire dataset, we measure the error at different r_{mask} using the TCGA-LUNG dataset. The results are shown in Fig. 6. Overall, when the r_{mask} is 0, 0.25, 0.50, and 0.75, the error remains relatively low. However, at r_{mask} of 1, the reconstruction error becomes significant. It has indicated that the proposed PD module is effective and stable in terms of position reconstruction performance. Besides, it can be observed that when $r_{mask} = 0$, overfitting leads to more outliers in the reconstruction results compared to when r_{mask} is 0.25 or 0.5.

To verify the robustness of our proposed PEGTB-MIL, we rotate the WSIs in the test set and then perform testing on the TCGA-LUNG dataset. The results are shown in Table 5. We conduct three groups of rotation tests, labeled (A)-(C), representing right rotations of 90 degrees, 180 degrees, and 270 degrees, respectively. The results from these experiments indicate that the differences($\triangle$) in the three performance metrics, compared to the conventional test results (F), are consistently less than 0.5%, demonstrating a notable degree of robustness.

Furthermore, to assess the importance of accurate position information, we randomly initialize the position information (2D coordinates) during either the training or test stages. Specifically, this involves shuffling the original patch-coordinate pairs, resulting in each patch is assigned an incorrect coordinate. We design two experiments, labeled (D) and (E), described as follows: (D) random initialization of position information during the training stage, with correct position information provided during the test stage; (E) correct position information provided during the training stage, with random initialization of position information during the test stage. Compared to the experiment with accurate position information (F), settings (D) and (E) show significant performance declines. It has underscored the critical importance of accurate position information for WSI classification and highlighted the necessity of maintaining spatial-semantic consistency of patch features. To facilitate more extensive validation, we also conduct anti-interference studies on TCGA-BRCA, USTC-EGFR, and USTC-GIST datasets, with results presented in Table A2 in the appendix. Overall, the conclusions analyzed from these results are generally consistent with those on the TCGA-LUNG dataset. It can be found that after coordinate rotation, the performance of the model fluctuates between -2% and 2%. The settings (D) and (E) exhibit a more significant decline in classification performance. It has further emphasized the importance of maintaining spatial-semantic consistency.

To investigate the impact of the aforementioned transformations on the attention mechanism of the model, we set up a series of attention heatmaps for analysis, as shown in Fig. 7. Fig. 7a displays the thumbnail of slide. Fig. 7b shows the annotated image. Fig. 7c illustrates heatmaps for regular testing. Figs. 7(d)-(h) corresponds to the attention heatmaps under the settings (A)-(E) as listed in Table 5. Overall, the rotation transformations have minimal impact on the attention mechanism of PEGTB-MIL. Even after applying the transformations of random initialization coordinates, the model can still identify tumor regions reasonably well. However, compared to using correct positional information, these transformations disrupt the spatial-semantic consistency of features, leading to decreased attention. This can also explain the significant decrease in model performance observed in Table 5 under settings (D) and (E).

4.7. Computational efficiency

We calculate the model parameters (Params) and floating point operations (FLOPs) for model inference to demonstrate the computational efficiency of the compared methods. The dimensionality of input features is set to 1024, while all other settings are based on the official implementation. The results are shown in Table 6. ABMIL and CLAM exhibit lower computational costs because they do not require modeling the dependencies among patches. LAMIL decreases the number of patch pairs in self-attention calculations by utilizing positional information, which further reduces computational costs. Consequently, LAMIL has fewer FLOPs than TransMIL. SETMIL contains more Transformer modules, leading to higher Params and FLOPs. Although the proposed PEGTB-MIL is not as computationally efficient as ABMIL and CLAM, it can achieve better performance. It is worth noting that PEGTB-MIL is more computationally efficient than other Transformer-based methods (TransMIL, LAMIL, and SETMIL) while also delivering better classification performance.

Table 6
Comparison of model parameters (Params) and floating point operations (FLOPs) between PEGTB-MIL and compared methods.

Methods	Params (MB)	$n_p = 500$* FLOPs ($\times 10^9$)	$n_p = 5000$ FLOPs ($\times 10^9$)	$n_p = 12000$ FLOPs ($\times 10^9$)	$n_p = 20000$ FLOPs ($\times 10^9$)
ABMIL	0.53	0.26	2.62	6.29	10.49
CLAM	0.79	0.39	3.93	9.44	15.73
TransMIL	2.67	1.92	13.78	33.07	54.56
LAMIL	2.63	1.31	13.11	31.46	52.43
SETMIL	14.70	1.79	18.09	44.67	74.44
PEGTB-MIL	1.71**	0.88	8.78	20.80	34.85

*: n_p denotes the number of patches.
**: The complete model has a parameter number of 2.76 MB. During inference, the PD module is removed. Therefore, the parameter number for inference is 1.71 MB.

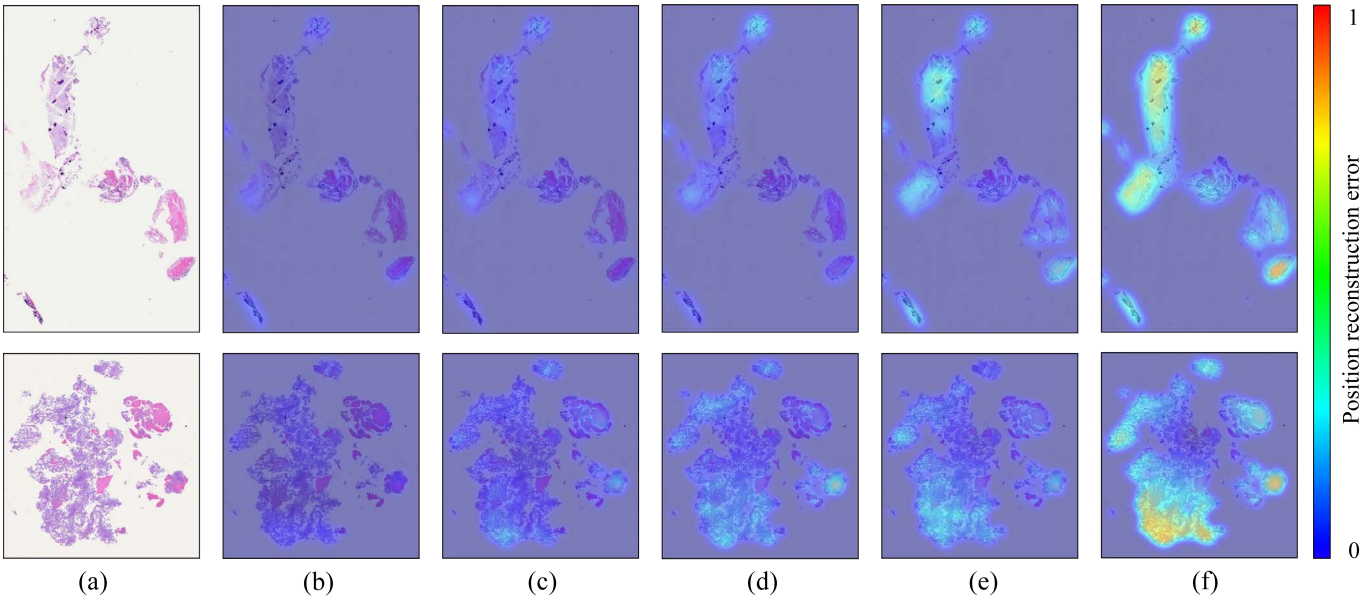


Fig. 8: The visualization of poor position reconstruction is as follows: (a) represents the thumbnail of the slide, and (b)-(f) represent the visualizations at r_{mask} of 0, 0.25, 0.50, 0.75, and 1.00 respectively.

5. Discussion

The proposed PEGTB-MIL is evaluated on two kinds of tasks (i.e. cancer subtyping and gene mutation prediction), specifically focusing on three kinds of cancers (i.e. lung, breast, and gastrointestinal cancers). In the future, we plan to improve our method and evaluate it on a broader range of multi-organ and multi-cancer datasets. In the field of computational pathology, while the capabilities of WSI analysis have shown promising advancements, there is a significant gap in clinical application. To achieve clinical application, more powerful AI models are needed. We propose a novel positional encoding method and a position reconstruction auxiliary task, achieving better performance compared to the previous MIL method. Notably, predicting gene mutations only from H&E slides can effectively reduce time and costs, assisting clinical decision-making. In this study, the improvements of our method on the USTC-EGFR and USTC-GIST datasets enhance model accuracy, providing more reliable results and reducing the gap with practical application requirements. Meanwhile, it also provides a new insight for research on MIL.

There are two potential limitations of the proposed PEGTB-MIL: (1) since position information is crucial for our method, it may not be well-suited for certain patch sampling strategies; (2) compared to image feature learning, the integration of position information needs a larger number of samples for effective positional pattern learning. However, obtaining sufficient samples presents a significant bottleneck in the medical field. Additionally, the size and distribution of tissue also impacts the results of position reconstruction. To validate this, we select and analyze two representative slides: (1) the first slide has small tissue regions; (2) the second slide has a more dispersed tissue distribution. The position reconstruction results for both slides are shown in Fig. 8. It can be observed that the small tissue regions may lead to insufficient reconstruction information at high mask ratios, resulting in poorer quality. For the second slide, tissue dispersion may increase the distances between regions, leading to insufficient reconstruction information.

6. Conclusion

In this paper, we propose a novel position encoding-guided transformer-based multiple instance learning (PEGTB-MIL) method for histopathology WSI classification. The proposed position encoding (PE) module is used to encode the 2D positional coordinates into the spatial-aware embeddings. Then, the spatial-aware embedding and the semantic features of the patches are incorporated to learn the spatial and semantic relationship among the patches by the multi-head self-attention (MHSA) module. In particular, a mask-based position reconstruction auxiliary task is proposed to enhance the spatial-semantic consistency and generalization capability of the patch features. The proposed method is validated on two publicly available TCGA and two in-house datasets. Experimental results demonstrate the effectiveness of PEGTB-MIL in cancer subtyping and gene mutation status prediction tasks.

In the future, we will try to incorporate multi-scale into position information reconstruction and explore the magnification-spatial coordinate-semantic feature representation. Furthermore, several multi-modal studies [8, 9, 47, 48] have already demonstrated the benefits of utilizing multiple modalities to improve model performance. We will consider extending our proposed position encoder-decoder modules into a multi-modal framework to enhance gene mutation detection performance from H&E slides.

Research data for this article

The Ethics Committee approvals were obtained for the private USTC-EGFR and USTC-GIST datasets. However, as we lack the necessary permissions for the USTC-EGFR and USTC-GIST datasets, the data remains undisclosed. Regarding the public datasets TCGA-LUNG and TCGA-BRCA, we have appropriately cited the required references.

Ethics statement

Data used in this study includes two publicly available TCGA-LUNG and TCGA-BRCA datasets, which have been ethically approved for research purposes. Besides, the WSIs from the USTC-EGFR and USTC-GIST datasets are protected by privacy and approved by the Medical Research Ethics Committee of the First Affiliated Hospital of the University of Science and Technology of China (Anhui Provincial Hospital) under protocols No. 2022-RE-454 and No. 2024KY-009. No animals were involved in this study and all experiments utilizing human data were conducted in accordance with the applicable ethical guidelines and regulations.

CRediT authorship contribution statement

Jun Shi: Conceptualization, Methodology, Data Curation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Dongdong Sun:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Kun Wu:** Software, Data Curation. **Zhiguo Jiang:** Supervision, Project administration, Funding acquisition. **Xue Kong:** Validation, Resources, Data Curation. **Wei Wang:** Validation, Resources, Data Curation. **Haibo Wu:** Validation, Resources, Data Curation, Funding acquisition. **Yushan Zheng:** Conceptualization, Methodology, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partly supported by the National Natural Science Foundation of China (No. 61906058, 61901018, 62171007, and 61771031), partly supported by Beijing Natural Science Foundation (Grant No. 7242270), partly supported by the Fundamental Research Funds for the Central Universities of China (No. YWF-23-Q-1075 and JZ2022HGTB0285), partly supported by Emergency Key Program of Guangzhou Laboratory (No. EKPG21-32), partly supported by Joint Fund for Medical Artificial Intelligence (No. MAI2023C014), partly supported by National Key Research and Development Program of China (No. 2021YFF1201004), partly supported by Research Funds of Centre for Leading Medicine and Advanced Technologies of IHM (No. 2023IHM01043), and partly supported by Anhui Provincial Health and Medical Research Project (Grant No. AHWJ2023A10143).

References

[1] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, T. J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature medicine 25 (8) (2019) 1301–1309.

[2] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nature biomedical engineering 5 (6) (2021) 555–570.

[3] Y. Zheng, J. Li, J. Shi, F. Xie, J. Huai, M. Cao, Z. Jiang, Kernel attention transformer for histopathology whole slide image analysis and assistant cancer diagnosis, IEEE Transactions on Medical Imaging (2023).

[4] A. Raju, J. Yao, M. M. Haq, J. Jonnagaddala, J. Huang, Graph attention multi-instance learning for accurate colorectal cancer staging, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23, Springer, 2020, pp. 529–539.

[5] J. Xu, H. Lu, H. Li, C. Yan, X. Wang, M. Zang, D. G. de Rooij, A. Madabhushi, E. Y. Xu, Computerized spermatogenesis staging (css) of mouse testis sections via quantitative histomorphological analysis, Medical image analysis 70 (2021) 101835.

[6] W. Bulten, K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. van Boven, R. Vink, et al.,

Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge, Nature medicine 28 (1) (2022) 154–163.
[7] Y. Fu, A. W. Jung, R. V. Torne, S. Gonzalez, H. Vöhringer, A. Shmatko, L. R. Yates, M. Jimenez-Linan, L. Moore, M. Gerstung, Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis, Nature cancer 1 (8) (2020) 800–810.
[8] R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, F. Mahmood, Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis, IEEE Transactions on Medical Imaging 41 (4) (2020) 757–770.
[9] R. J. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo, et al., Pan-cancer integrative histology-genomic analysis via multimodal deep learning, Cancer Cell 40 (8) (2022) 865–878.
[10] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, A. Tsirigos, Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning, Nature medicine 24 (10) (2018) 1559–1567.
[11] R. Yamashita, J. Long, T. Longacre, L. Peng, G. Berry, B. Martin, J. Higgins, D. L. Rubin, J. Shen, Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study, The Lancet Oncology 22 (1) (2021) 132–141.
[12] R. Yan, Y. Shen, X. Zhang, P. Xu, J. Wang, J. Li, F. Ren, D. Ye, S. K. Zhou, Histopathological bladder cancer gene mutation prediction with hierarchical deep multiple-instance learning, Medical Image Analysis 87 (2023) 102824.
[13] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, N. Rajpoot, Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, Medical image analysis 58 (2019) 101563.
[14] H. Sharma, N. Zerbe, I. Klempert, O. Hellwich, P. Hufnagl, Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology, Computerized Medical Imaging and Graphics 61 (2017) 2–13.
[15] C.-W. Wang, S.-C. Huang, Y.-C. Lee, Y.-J. Shen, S.-I. Meng, J. L. Gaol, Deep learning for bone marrow cell detection and classification on whole-slide images, Medical Image Analysis 75 (2022) 102270.
[16] Y. Xu, J.-Y. Zhu, I. Eric, C. Chang, M. Lai, Z. Tu, Weakly supervised histopathology cancer image segmentation and classification, Medical image analysis 18 (3) (2014) 591–604.
[17] J. van der Laak, F. Ciompi, G. Litjens, No pixel-level annotations needed, Nature biomedical engineering 3 (11) (2019) 855–856.
[18] H. Pinckaers, W. Bulten, J. van der Laak, G. Litjens, Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels, IEEE Transactions on Medical Imaging 40 (7) (2021) 1817–1826.
[19] W. Lu, M. Toss, M. Dawood, E. Rakha, N. Rajpoot, F. Minhas, Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer, Medical Image Analysis 80 (2022) 102486.
[20] T. G. Dietterich, R. H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artificial intelligence 89 (1-2) (1997) 31–71.
[21] J. Amores, Multiple instance classification: Review, taxonomy and comparative study, Artificial intelligence 201 (2013) 81–105.
[22] M. Oquab, L. Bottou, I. Laptev, J. Sivic, et al., Weakly supervised object recognition with convolutional neural networks, in: Proc. of NIPS, Vol. 2014, Citeseer, 2014, pp. 1545–5963.
[23] X. Wang, Y. Yan, P. Tang, X. Bai, W. Liu, Revisiting multiple instance neural networks, Pattern Recognition 74 (2018) 15–24.
[24] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International conference on machine learning, PMLR, 2018, pp. 2127–2136.
[25] G. Xu, Z. Song, Z. Sun, C. Ku, Z. Yang, C. Liu, S. Wang, J. Ma, W. Xu, Camel: A weakly supervised learning framework for histopathology image segmentation, in: Proceedings of the IEEE/CVF International Conference on computer vision, 2019, pp. 10682–10691.
[26] P. Courtiol, E. W. Tramel, M. Sanselme, G. Wainrib, Classification and disease localization in histopathology using only global labels: A weakly-supervised approach, arXiv preprint arXiv:1802.02212 (2018).
[27] L. Qu, Y. Ma, X. Luo, M. Wang, Z. Song, Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need, arXiv preprint arXiv:2307.02249 (2023).
[28] L. Keren, M. Bosse, D. Marquez, R. Angoshtari, S. Jain, S. Varma, S.-R. Yang, A. Kurian, D. Van Valen, R. West, et al., A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging, Cell 174 (6) (2018) 1373–1387.
[29] E. R. Parra, J. Zhang, M. Jiang, A. Tamegnon, R. K. Pandurengan, C. Behrens, L. Solis, C. Haymaker, J. V. Heymach, C. Moran, et al., Immune cellular patterns of distribution affect outcomes of patients with non-small cell lung cancer, Nature communications 14 (1) (2023) 2364.
[30] Z. Seferbekova, A. Lomakin, L. R. Yates, M. Gerstung, Spatial biology of cancer evolution, Nature Reviews Genetics 24 (5) (2023) 295–313.
[31] C. M. Schürch, S. S. Bhate, G. L. Barlow, D. J. Phillips, L. Noti, I. Zlobec, P. Chu, S. Black, J. Demeter, D. R. McIlwain, et al., Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front, Cell 182 (5) (2020) 1341–1359.
[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[34] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al., Transmil: Transformer based correlated multiple instance learning for whole slide image classification, Advances in neural information processing systems 34 (2021) 2136–2147.
[35] D. Reisenbüchler, S. J. Wagner, M. Boxberg, T. Peng, Local attention graph-based transformer for multi-target genetic alteration prediction, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II, Springer, 2022, pp. 377–386.
[36] Y. Zhao, Z. Lin, K. Sun, Y. Zhang, J. Huang, L. Wang, J. Yao, Setmil: spatial encoding transformer-based multiple instance learning for pathological image analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 66–76.
[37] S. Ding, J. Wang, J. Li, J. Shi, Multi-scale prototypical transformer for whole slide image classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 602–611.
[38] N. Otsu, A threshold selection method from gray-level histograms, IEEE transactions on systems, man, and cybernetics 9 (1) (1979) 62–66.
[39] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, V. Singh, Nyströmformer: A nyström-based algorithm for approximating self-attention, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 14138–14148.
[40] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
[41] D. A. Gutman, J. Cobb, D. Somanna, Y. Park, F. Wang, T. Kurc, J. H. Saltz, D. J. Brat, L. A. Cooper, J. Kong, Cancer digital slide archive: an informatics resource to support integrated in silico analysis of tcga pathology data, Journal of the American Medical Informatics Association 20 (6) (2013) 1091–1098.
[42] Y.-L. Zhang, J.-Q. Yuan, K.-F. Wang, X.-H. Fu, X.-R. Han, D. Threapleton, Z.-Y. Yang, C. Mao, J.-L. Tang, The prevalence of egfr mutation in patients with non-small cell lung cancer: a systematic review and meta-analysis, Oncotarget 7 (48) (2016) 78985.

[43] P. G. Casali, J.-Y. Blay, N. Abecassis, J. Bajpai, S. Bauer, R. Biagini, S. Bielack, S. Bonvalot, I. Boukovinas, J. Bovee, et al., Gastrointestinal stromal tumours: Esmo–euracan–genturis clinical practice guidelines for diagnosis, treatment and follow-up, Annals of oncology 33 (1) (2022) 20–33.

[44] A. Jakhetiya, P. K. Garg, G. Prakash, J. Sharma, R. Pandey, D. Pandey, Targeted therapy of gastrointestinal stromal tumours, World Journal of Gastrointestinal Surgery 8 (5) (2016) 345.

[45] G. D. Demetri, M. Von Mehren, C. D. Blanke, A. D. Van den Abbeele, B. Eisenberg, P. J. Roberts, M. C. Heinrich, D. A. Tuveson, S. Singer, M. Janicek, et al., Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors, New England Journal of Medicine 347 (7) (2002) 472–480.

[46] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[47] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, J. Zou, A visual–language foundation model for pathology image analysis using medical twitter, Nature medicine 29 (9) (2023) 2307–2316.

[48] Z. Wang, L. Yu, X. Ding, X. Liao, L. Wang, Shared-specific feature learning with bottleneck fusion transformer for multi-modal whole slide image analysis, IEEE Transactions on Medical Imaging (2023).

Appendix

Table A1
Ablation study of PEGTB-MIL on the TCGA-BRCA, USTC-EGFR, and USTC-GIST datasets.

TCGA-BRCA				
Settings	**Position encoding strategy**	**Evaluation metrics**		
		ACC (%)	AUC (%)	F1 (%)
(A)	None	77.06 ± 2.71	80.47 ± 0.90	66.07 ± 3.23
(B)	1D-Embedding	77.21 ± 5.34	81.86 ± 4.30	69.42 ± 7.88
(C)	2D-Embedding	81.98 ± 2.08	83.16 ± 2.19	73.64 ± 4.88
(D)	PE w/o Normalization	80.33 ± 4.19	83.33 ± 3.38	73.07 ± 7.30
(E)	PE	83.21 ± 3.66	84.12 ± 2.07	76.24 ± 3.55
(F)	PE+PD (ours)	$\mathbf{86.31 \pm 1.80}$	$\mathbf{86.74 \pm 2.64}$	$\mathbf{76.71 \pm 3.31}$
USTC-EGFR				
Settings	**Position encoding strategy**	**Evaluation metrics**		
		ACC (%)	AUC (%)	F1 (%)
(A)	None	50.84 ± 1.06	79.27 ± 1.45	51.95 ± 1.79
(B)	1D-Embedding	49.13 ± 1.85	78.66 ± 0.99	47.02 ± 3.06
(C)	2D-Embedding	50.20 ± 2.88	80.19 ± 0.74	45.65 ± 1.37
(D)	PE w/o Normalization	51.66 ± 4.35	82.49 ± 1.74	49.65 ± 3.31
(E)	PE	51.66 ± 3.10	82.61 ± 2.09	49.75 ± 5.25
(F)	PE+PD (ours)	$\mathbf{52.47 \pm 3.77}$	$\mathbf{83.25 \pm 1.65}$	$\mathbf{51.40 \pm 4.10}$
USTC-GIST				
Settings	**Position encoding strategy**	**Evaluation metrics**		
		ACC (%)	AUC (%)	F1 (%)
(A)	None	53.56 ± 6.46	67.45 ± 3.31	37.16 ± 3.64
(B)	1D-Embedding	51.60 ± 14.20	66.87 ± 4.08	38.78 ± 4.91
(C)	2D-Embedding	57.56 ± 4.56	70.48 ± 3.93	40.99 ± 2.60
(D)	PE w/o Normalization	58.56 ± 5.26	68.33 ± 3.37	41.40 ± 1.64
(E)	PE	56.02 ± 6.98	69.13 ± 2.63	42.83 ± 1.01
(F)	PE+PD (ours)	$\mathbf{65.52 \pm 3.76}$	$\mathbf{72.52 \pm 1.63}$	$\mathbf{45.28 \pm 2.63}$

Table A2
The anti-interference experimental results of PEGTB-MIL on the TCGA-BRCA, USTC-EGFR, and USTC-GIST datasets.

TCGA-BRCA							
Settings	Transformation	ACC(%)	$\triangle_{ACC}$(%)	AUC(%)	$\triangle_{AUC}$(%)	F1(%)	$\triangle_{F1}$(%)
(A)	RIGHT 90°	86.23 ± 0.97	−0.08	$\mathbf{87.10 \pm 1.93}$	+0.36	76.33 ± 1.44	−0.38
(B)	RIGHT 180°	$\mathbf{86.49 \pm 1.03}$	+0.20	85.72 ± 1.28	−1.02	76.55 ± 2.90	−0.16
(C)	RIGHT 270°	85.79 ± 0.64	−0.52	86.97 ± 2.30	+0.23	76.68 ± 1.32	−0.03
(D)	Random initialization (Train)	82.38 ± 1.17	−3.93	83.44 ± 0.98	−3.30	70.72 ± 2.19	−5.99
(E)	Random initialization (Test)	84.02 ± 1.92	−2.29	83.63 ± 0.22	−3.11	71.92 ± 1.56	−4.79
(F)	None	86.31 ± 1.80	——	86.74 ± 2.64	——	$\mathbf{76.71 \pm 3.31}$	——
USTC-EGFR							
Settings	Transformation	ACC(%)	$\triangle_{ACC}$(%)	AUC(%)	$\triangle_{AUC}$(%)	F1(%)	$\triangle_{F1}$(%)
(A)	RIGHT 90°	52.93 ± 2.90	+0.46	82.17 ± 2.59	−1.08	51.17 ± 3.20	−0.23
(B)	RIGHT 180°	$\mathbf{52.98 \pm 3.87}$	+0.51	83.49 ± 1.76	+0.24	$\mathbf{52.73 \pm 1.77}$	+1.33
(C)	RIGHT 270°	51.82 ± 1.69	−0.65	$\mathbf{84.92 \pm 3.92}$	+1.67	50.64 ± 2.04	−0.76
(D)	Random initialization (Train)	47.77 ± 2.57	−4.70	78.40 ± 2.47	−4.85	43.93 ± 2.12	−7.47
(E)	Random initialization (Test)	45.98 ± 3.53	−6.49	80.44 ± 3.70	−2.81	42.24 ± 3.74	−9.16
(F)	None	52.47 ± 3.77	——	83.25 ± 1.65	——	51.40 ± 4.10	——
USTC-GIST							
Settings	Transformation	ACC(%)	$\triangle_{ACC}$(%)	AUC(%)	$\triangle_{AUC}$(%)	F1(%)	$\triangle_{F1}$(%)
(A)	RIGHT 90°	64.98 ± 3.32	−0.54	70.73 ± 1.93	−1.79	45.11 ± 1.80	−0.17
(B)	RIGHT 180°	64.33 ± 2.75	−1.19	$\mathbf{73.41 \pm 1.22}$	+0.89	43.40 ± 2.36	−1.88
(C)	RIGHT 270°	63.75 ± 2.05	−1.77	70.86 ± 2.19	−1.66	$\mathbf{45.38 \pm 2.64}$	+0.10
(D)	Random initialization (Train)	57.46 ± 4.70	−8.06	67.29 ± 4.61	−5.23	41.56 ± 2.97	−3.72
(E)	Random initialization (Test)	60.29 ± 4.47	−5.23	69.07 ± 3.53	−3.45	40.69 ± 1.44	−4.59
(F)	None	$\mathbf{65.52 \pm 3.76}$	——	72.52 ± 1.63	——	45.28 ± 2.63	——