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

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. In this paper, we propose a novel position 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. The proposed method is evaluated on one public TCGA dataset and two in-house datasets. Experimental results validate it is effective in the tasks of lung cancer subtyping and gene mutation status prediction for lung cancer and gastrointestinal stromal tumor (GIST), and superior to the state-of-the-art MIL-based methods.

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 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

*Corresponding author. yszheng@buaa.edu.cn (Yushan Zheng), wuhaibo@ustc.edu.cn (Haibo Wu).

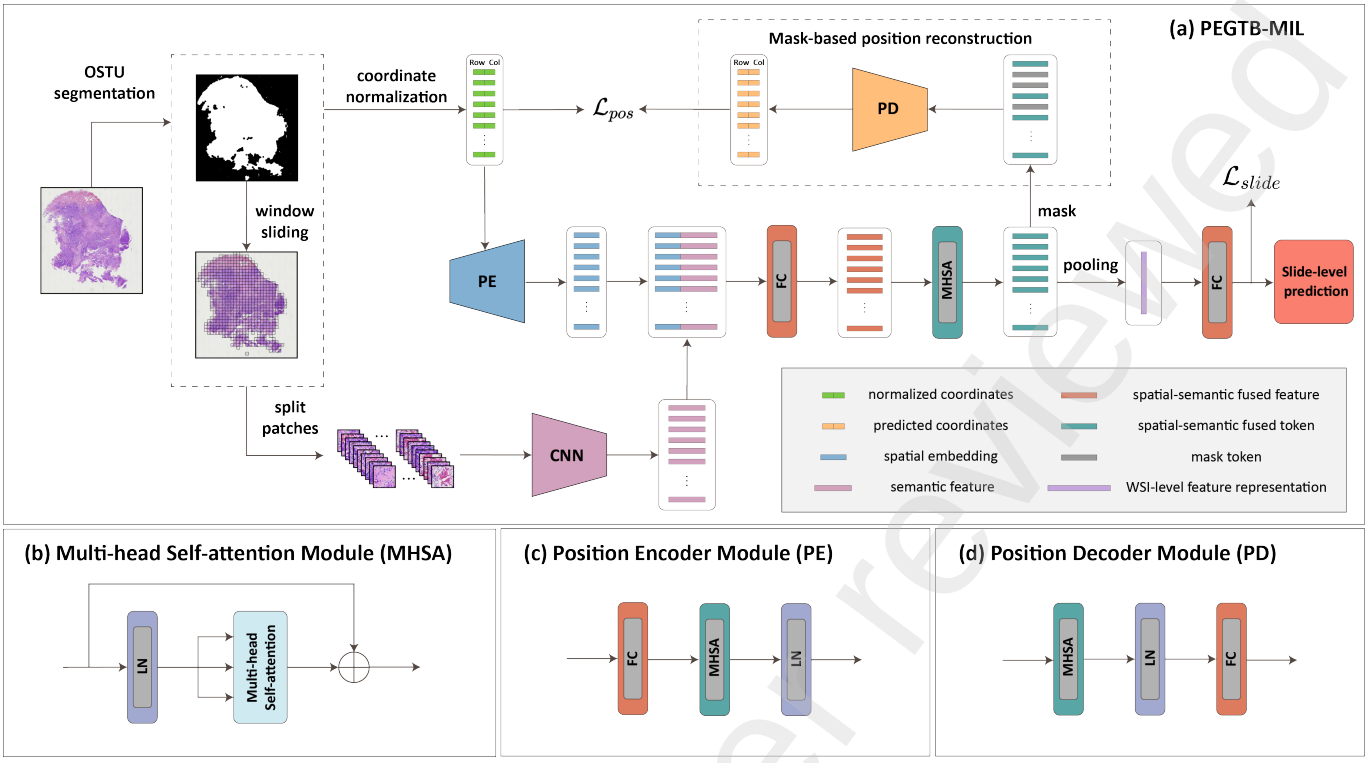


Fig. 1: The overview of the proposed PEGT-MIL for WSI classification, where (a) shows the entire workflow, (b) describes the structure of the multi-head self-attention module, (c) and (d) illustrate the position encoder module and position decoder module, respectively.

that ABMIL [24] 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. Recently, self-attention [28, 29] 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 [30–33] 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) [29] 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 and a gastrointestinal stromal tumor (GIST) dataset, and is compared with the state-of-the-art MIL-based methods [2, 24, 30–32]. 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 one public TCGA dataset and two in-house clinical datasets. The results prove our method has achieved the superior classification performance in lung 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 II reviews the MIL-related works. Section III introduces the proposed framework. Section IV shows the experimental results and analysis. Section V gives the conclusion.

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, 30–33] 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. ABMIL[24] 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. CLAM [2] 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[30–33] have designed the transformer-based MIL methods for WSI classification, which focus on the global correlation exploration of the patches within a WSI. TransMIL[30] 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. LAMIL[31] 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. SETMIL[32] leverages the spatial encoding transformer to update instance representations by simultaneously aggregating neighboring and globally correlated instances. MSPT[33] 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

Digital pathology WSIs contain rich morphological information and are therefore considered the gold standard for cancer diagnosis. In addition to cellular features such as nuclear-cytoplasmic ratio, morphology, and staining

intensity, the relative positions 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. In the early stage of MIL, the spatial relationships among the instances were often overlooked. Recently, ViT[29] 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[30–33] 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[30] 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[31] 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[32] 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[34] 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. Furthermore, to preserve the rotation semantic invariance of the WSIs and enhance the robustness of the model, the random rotation data augmentations are employed to transform the coordinates, including the clockwise rotations of 0°, 90°, 180°, and 270°.

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[30], the Nyströmformer[35] 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:

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

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, and d_m refers to the dimension of the spatial-semantic fused features.

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)[36], 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, LN layer, and a FC 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:

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

where $\mathrm{MaskIDs}(m, n)$ 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:

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

where $\mathbf{W}_{PD} \in \mathbb{R}^{d_m\times 2}$ is a learnable parameter matrix. The Mean Squared Error (MSE) loss $\mathcal{L}_{pos}$ is used to measure

Table 1
The WSI Distribution of the three Datasets

TCGA-LUNG	Normal	LUAD	LUSC		
Train	330	279	285		
Val	55	47	48		
Test	165	141	144		
USTC-EGFR	Neg	19del	L858R	Wild	Others
Train	103	71	114	86	86
Val	14	9	23	13	12
Test	48	38	47	47	43
USTC-GIST	Wild	Exon 9[1]	Exon 11[2]	Others	
Train	44	40	256	29	
Val	7	8	41	6	
Test	18	16	128	25	

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

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 in one public TCGA dataset (TCGA-LUNG) and two in-house clinical datasets (USTC-EGFR and USTC-GIST) which come from the First Affiliated Hospital of USTC (University of Science and Technology of China). TCGA-LUNG is used for lung 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.

• **TCGA-LUNG** contains 1345 cases from the Cancer Genome Atlas (TCGA) program of the National Cancer Institute (NCI). We selected 1494 slides while maintaining class balance. This dataset includes three categories: non-cancerous tissue (Normal), Lung Squamous Cell Carcinoma(LUSC), and Lung Adenocarcinoma (LUAD).

• **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[37]. 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 618 slides from 92 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[38]. Particularly, its incidence in China 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[39]. 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[40]. 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 two categories, the wild type (Wild) and other driver gene mutation types (Others).

A comprehensive summary of these datasets is presented in Table 1. All the labels of these two in-house clinical datasets have been confirmed by pathologists. We divided the three datasets into train, validation, and test sets with a ratio of 6:1:3 at the patient level. The train set is used to train the model, the validation set is used for model selection and hyper-parameter verification, and finally, we evaluate the performance 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[34] 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 ResNet50[41] pre-trained in the ImageNet. 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 set. All the experiments are conducted on one computer with an AMD Ryzen Threadripper 3960X 24-Core Processor CPU and a NVIDIA RTX 3090 GPU. The accuracy of classification

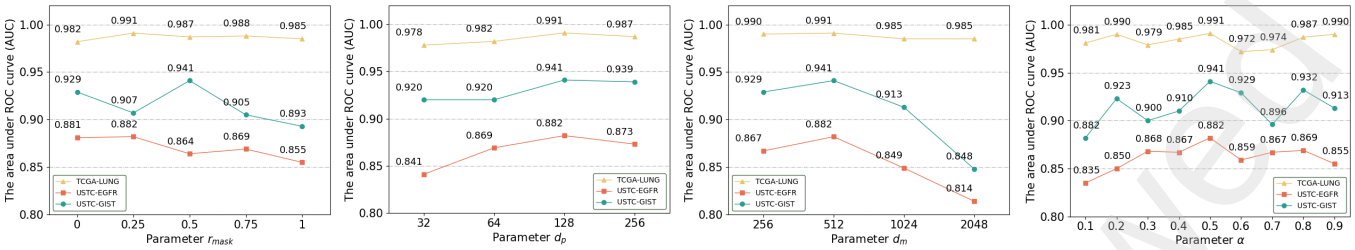


Fig. 2: The performance curves of tuning four hyper-parameters on the three datasets, which (a) is the mask ratio r_{mask}, (b) is the dimension of the spatial embedding d_p, (c) is the dimension d_m of the spatial-semantic fused feature, and (d) the weight α of $\mathcal{L}_{slide}$.

Table 2
Ablation study of PEGTB-MIL for TCGA-LUNG, USTC-EGFR, and USTC-GIST datasets.

Methods	TCGA-LUNG		USTC-EGFR		USTC-GIST	
	ACC	AUC	ACC	AUC	ACC	AUC
PEGTB-MIL w/o PE&PD	0.871	0.966	0.444	0.772	0.652	0.742
PEGTB-MIL w/o PD	0.827	0.945	0.507	0.820	0.561	0.716
PEGTB-MIL	**0.880**	**0.972**	**0.547**	**0.849**	**0.679**	**0.768**

(ACC) and the area under receiver operating characteristic curve (AUC) are served as the metric of performance evaluation. Our codes are available at https://github.com/HFUT-miaLab/PEGTB-MIL.

4.3. Hyper-parameter verification

We conduct experiments on the three datasets 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 in these three datasets are all selected according to the AUC results of the validation set. The detailed results are shown in Fig. 2.

4.3.1. 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 in the TCGA-LUNG dataset, 0.25 in the USTC-EGFR dataset, and 0.5 in the USTC-GIST dataset. It is worth noting that the validation results in the USTC-GIST dataset show relatively larger fluctuation as r_{mask} varies. This may be due to the fact that the class imbalance has an impact on feature representation, resulting in unstable classification performance.

4.3.2. 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. All the optimal values for d_p are 128 on the three datasets, 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).

4.3.3. The dimension d_m of the spatial-semantic fused features

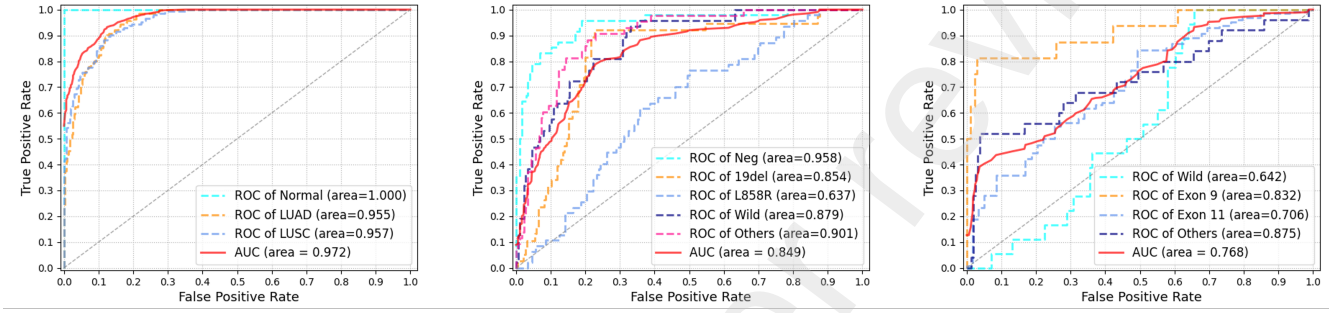
After incorporating the patch semantic features and their corresponding spatial embeddings, a transformation matrix $\mathbf{W}_{feat}$ is utilized to map the dimension of the concatenated features to d_m, resulting in the spatial-semantic fused features. We tune d_m within the range of [256, 512, 1024, 2048]. As presented in Fig. 2c, the optimal values of d_m in all these three datasets are 512. Note that the performance of PEGTB-MIL is sensitive to the parameter d_m in the USTC-EGFR and USTC-GIST datasets. When d_m = 2048, the performance shows a significant decline. This may be because a higher feature dimension leads to overfitting when the number of cases is insufficient.

4.3.4. 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 in the three datasets is 0.5, which also demonstrates that balanced $\mathcal{L}_{slide}$ and $\mathcal{L}_{pos}$ can better guide the model training.

Table 3
Results of various MIL methods on TCGA-LUNG, USTC-EGFR, and USTC-GIST datasets.

Methods	TCGA-LUNG		USTC-EGFR		USTC-GIST	
	ACC	AUC	ACC	AUC	ACC	AUC
ABMIL	0.851	0.944	0.444	0.754	0.658	0.714
CLAM	0.876	0.965	0.520	0.816	0.572	0.748
TransMIL	0.871	0.969	0.439	0.785	0.642	0.702
LAMIL	0.791	0.955	0.511	0.824	0.588	0.720
SETMIL	0.793	0.914	0.444	0.787	0.567	0.706
PEGTB-MIL	**0.880**	**0.972**	**0.547**	**0.849**	**0.679**	**0.768**



(a) (b) (c)

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

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 in the three datasets. The results are shown in Table 2. PEGTB-MIL w/o PE&PD represents the network without position encoding and masked reconstruction-based position decoding. PEGTB-MIL w/o PD indicates that the network utilizes the PE module for position encoding without the PD module for position reconstruction. We can observe that the performance of PEGTB-MIL w/o PD shows a significant AUC decline as follows: 0.027 on TCGA-LUNG (0.972 to 0.945), 0.029 on USTC-EGFR (0.849 to 0.820), and 0.052 on USTC-GIST (0.768 to 0.716). 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. Note that the performance of PEGTB-MIL w/o PE&PD is better than PEGTB-MIL w/o PD in the experiments conducted on the TCGA-LUNG and USTC-GIST datasets. It can be explained that without the PD module, the spatial embeddings generated by the PE module and the patch semantic features may not maintain great consistency. Therefore, it yields the training biases and performance decline.

4.5. Comparative experiments

We compare our method with five most representative MIL-based methods, including ABMIL[24], CLAM[2], TransMIL[30], LAMIL[31], and SETMIL[32]. The experiments for these methods are conducted using official code and the same experimental settings. The results of comparative experiments on the test set are presented in Table 3.

Overall, the proposed PEGTB-MIL achieves the best performance with ACC and AUC of 0.880 and 0.972 on the TCGA-LUNG dataset, 0.547 and 0.849 on the USTC-EGFR dataset, and 0.679 and 0.768 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 three datasets, with AUC gains of 0.021 on the TCGA-LUNG dataset, 0.062 on the USTC-EGFR dataset, and 0.034 on the USTC-GIST dataset.

For the transformer-based MIL methods, TransMIL achieves better performance than CLAM on the TCGA-LUNG dataset (AUC is 0.004 higher), but its AUC is lower than CLAM by 0.031 and 0.046 on the USTC-EGFR and USTC-GIST datasets, respectively. The reason may be that the two in-house datasets (USTC-EGFR and USTC-GIST) have a limited number of cases compared to the TCGA-LUNG dataset, especially USTC-GIST, which may lead to the performance decline. As portrayed in Table 3, LAMIL achieves an AUC of 0.039 and 0.018 higher than TransMIL on the USTC-EGFR and USTC-GIST datasets, respectively. Compared to the reshaping strategy employed by TransMIL,

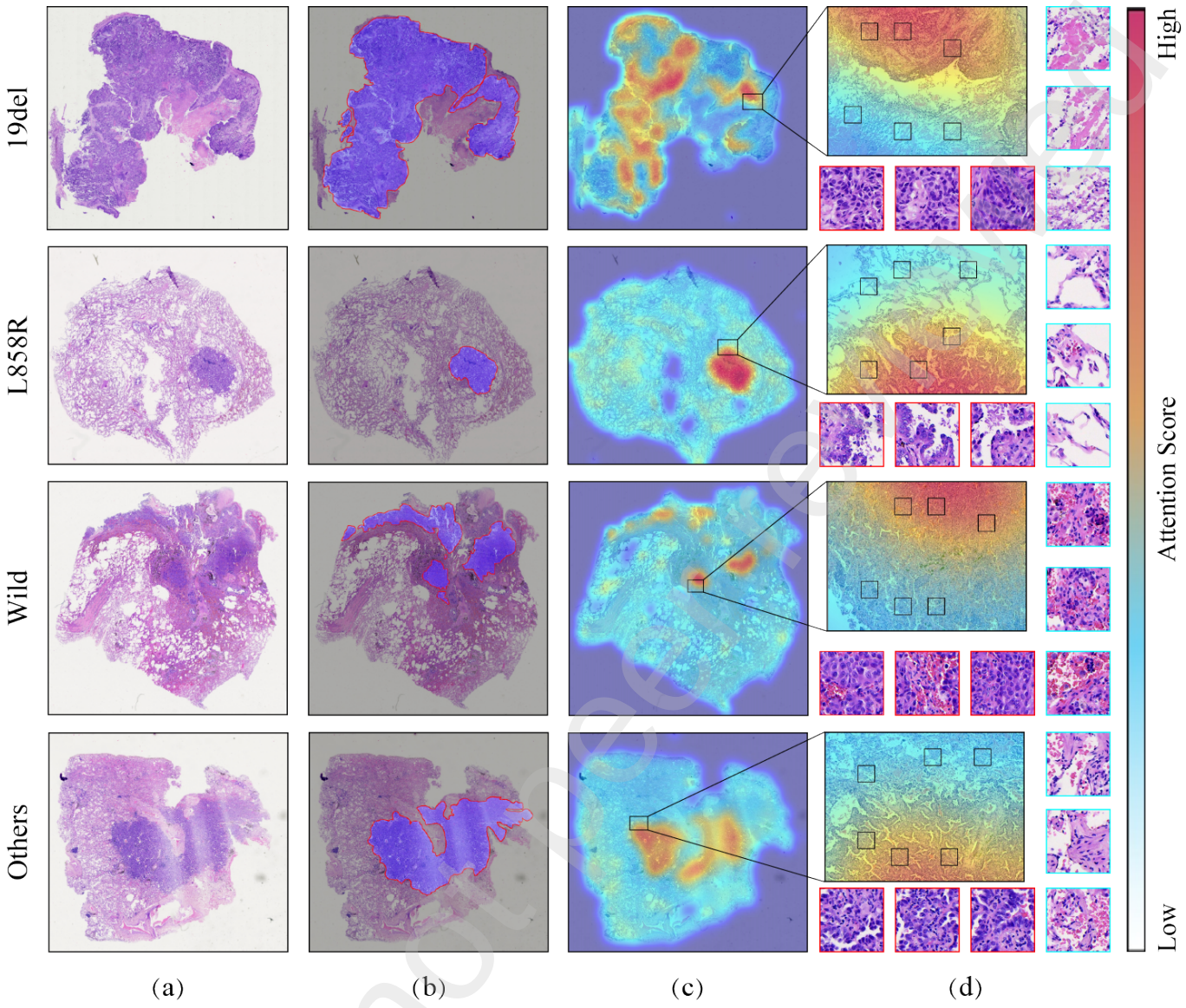


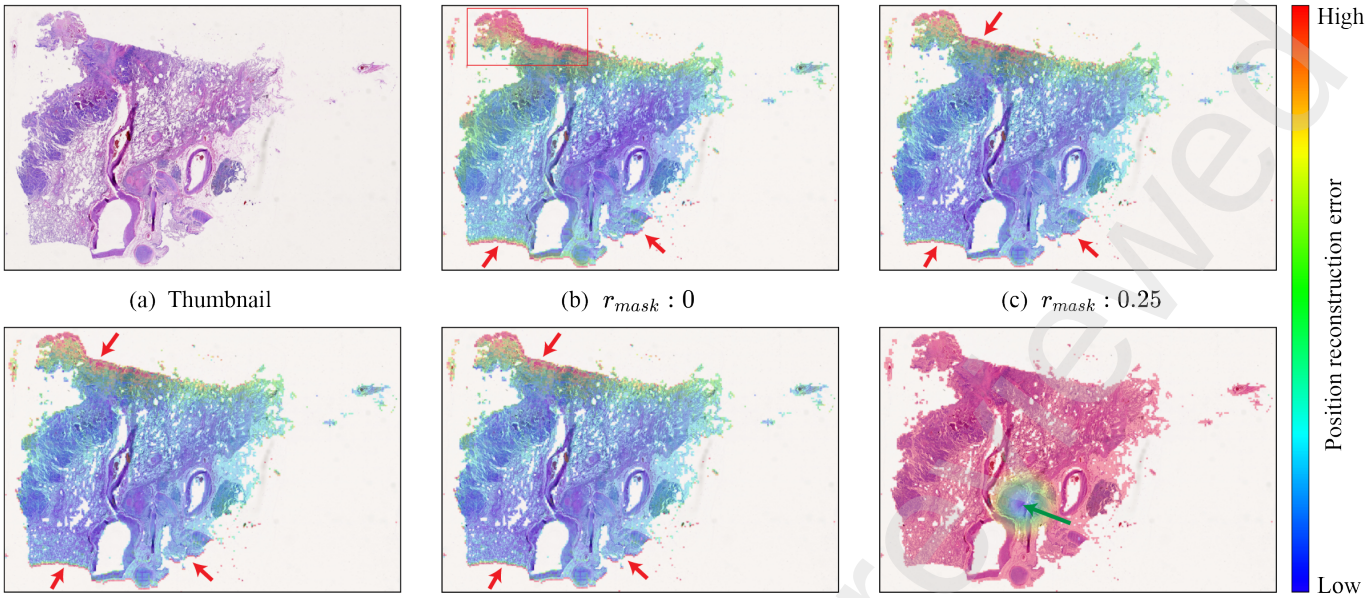
Fig. 4: Visualization based on the attention map in PEGTB-MIL, where (a) displays the thumbnails for each slide, (b) shows the lesion tissue images annotated by pathologists, (c) displays the attention heatmaps and (d) shows the ROI images of the attention heatmaps along with some representative patches.

LAMIL utilizes the 2D coordinates to build a KNN graph, which reflects the real adjacency relationships and facilitates the model to learn better contextual spatial relationships among the patches.

Note that SETMIL has a poor performance (AUC = 0.914) on the TCGA-LUNG dataset compared to other methods. It can be explained that SETMIL is sensitive and unsuitable for this dataset. SETMIL introduces 2D positional information as a bias term into the self-attention calculation to learn relative spatial relationships. It is difficult to maintain the spatial-semantic consistency of features for the fine-grained gene mutation prediction task, resulting in a decline in model performance. Therefore, SETMIL has an AUC of 0.037 lower than LAMIL on the USTC-EGFR dataset and 0.014 lower on the USTC-GIST dataset.

As indicated in Table 3, the proposed method achieves a 0.009 higher ACC and a 0.003 higher AUC than the second-best method TransMIL on the TCGA-LUNG dataset. On the USTC-EGFR dataset, it outperforms the second-best method LAMIL by 0.036 of ACC and 0.025 of AUC. On the USTC-GIST dataset, it exhibits a 0.107 higher ACC and a 0.020 higher AUC 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 three datasets in Fig. 3. Fig. 3a shows the ROC curves of PEGTB-MIL in the TCGA-LUNG



(a) Thumbnail (b) r_{mask} : 0 (c) r_{mask} : 0.25

(d) r_{mask} : 0.50 (e) r_{mask} : 0.75 (f) r_{mask} : 1.00

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.

dataset. We can observe that PEGTB-MIL exhibits the best performance in the Normal category (AUC = 1.000), and also shows excellent results in the other two tumor categories (LUAD and LUSC), with AUC greater than 0.95. Fig. 3b displays the ROC curves of PEGTB-MIL in the USTC-EGFR dataset. The proposed method has demonstrated the efficient performance in identifying the negative category (AUC = 0.958). In addition, PEGTB-MIL exhibits poorer performance in identifying the L858R mutation compared to other categories, with an AUC of 0.217 lower than the 19del category, 0.242 lower than the Wild category, and 0.264 lower than the Others category. It may be because the L858R category is more challenging to identify compared to other categories. Although PEGTB-MIL exhibits relatively lower performance in identifying the L858R category, it has demonstrated a notable AUC result (AUC = 0.854) for the 19del category. Fig. 3c exhibits the ROC curves of PEGTB-MIL in the USTC-GIST dataset. For GIST, a rare neoplasm, the overall AUC has gained an acceptable result, 0.768, and the AUC of each mutation type (Exon 9, Exon 11 and Others) is higher than 0.70. It has indicated that PEGTB-MIL has better classification ability under very limited cases.

4.6. Visualization and interpretation

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) Attention heatmaps generated by the attention score of each patch; and (2) Visualization of position reconstruction.

4.6.1. 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 thumbnail. According to the settings of the r_{mask}

in Section IV.C, we show the results of position coordinates reconstruction for $r_{mask} = 0, 0.25, 0.5, 0.75$, and 1.00 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 fewer correct coordinate (pointed by green arrow), as depicted in Fig. 5f. When r_{mask} is equal to 0.25, 0.50, or 0.75, it is evident that the reconstruction effects in Figs. 5(c)-(e) are generally similar, indicating that the position reconstruction is insensitive to changes in the r_{mask} parameter and has strong robustness. Note that the coordinate reconstruction of the patches located at the edges of tissue regions (e.g. the red arrows as shown in Figs. 5(c)-(e)) exhibits poorer accuracy compared to the patches in the central regions. The reason may be that the patches near the edge of the tissue have fewer neighboring patches to obtain positional information than the patches in the center of the tissue. Consequently, it leads to a decrease in the reconstruction performance for the patches near the edge of the tissue. Notably, when r_{mask} is equal to 0, the position reconstruction task removes the mask mechanism. From the red-boxed region in Fig. 5b, it can be observed that the model with $r_{mask} = 0$ exhibits larger position reconstruction errors. Therefore, it has indicated that the mask mechanism can enhance the generalization ability of the PD module in reconstructing coordinates.

5. 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 a 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, 42, 43] 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.

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:** Resources, Data Curation. **Wei Wang:** Resources, Data Curation. **Haibo Wu:** 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 the Fundamental Research Funds for the Central Universities of China (No. JZ2022HGTB0285), partly supported by Emergency Key Program of Guangzhou Laboratory (No. EKPG21-32), partly supported by Joint Fund for Medical Artificial Intelligence (No. MAI2023C014), and partly supported by National Key Research and Development Program of China (No. 2021YFF1201000).

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] 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).

[29] 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).

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] N. Otsu, A threshold selection method from gray-level histograms, IEEE transactions on systems, man, and cybernetics 9 (1) (1979) 62–66.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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).