






# Multi-Scale Efficient Graph-Transformer for Whole Slide Image Classification

Saisai Ding , Juncheng Li , Jun Wang , *Member, IEEE*, Shihui Ying , *Member, IEEE*, and Jun Shi 

**Abstract**—The multi-scale information among the whole slide images (WSIs) is essential for cancer diagnosis. Although the existing multi-scale vision Transformer has shown its effectiveness for learning multi-scale image representation, it still cannot work well on the gigapixel WSIs due to their extremely large image sizes. To this end, we propose a novel Multi-scale Efficient Graph-Transformer (MEGT) framework for WSI classification. The key idea of MEGT is to adopt two independent efficient Graph-based Transformer (EGT) branches to process the low-resolution and high-resolution patch embeddings (i.e., tokens in a Transformer) of WSIs, respectively, and then fuse these tokens via a multi-scale feature fusion module (MFFM). Specifically, we design an EGT to efficiently learn the local-global information of patch tokens, which integrates the graph representation into Transformer to capture spatial-related information of WSIs. Meanwhile, we propose a novel MFFM to alleviate the semantic gap among different resolution patches during feature fusion, which creates a non-patch token for each branch as an agent to exchange information with another branch by cross-attention mechanism. In addition, to expedite network training, a new token pruning module is developed in EGT to reduce the redundant tokens. Extensive experiments on both TCGA-RCC and CAMELYON16 datasets demonstrate the effectiveness of the proposed MEGT.

**Index Terms**—Cancer diagnosis, cross-attention, graph-Transformer, multi-scale feature fusion, Whole slide images.

## I. INTRODUCTION

HISTOPATHOLOGICAL images are considered as the “gold standard” for cancer diagnosis, and the analysis

of whole slide images (WSIs) is a critical approach for cancer diagnosis, prognosis, and survival prediction [1], [2], [3]. Nowadays, deep learning (DL) has been successfully applied to the field of computational pathology to develop the computer-aided diagnosis (CAD) system, which can help pathologists improve diagnosis accuracy together with good consistency and reproducibility [4], [5].

Due to the huge size of WSIs and lack of pixel-level annotation, the weakly-supervised learning framework is generally adopted for analysis of gigapixel WSIs, among which multiple instance learning (MIL) is a typical method [6], [7]. Under the MIL framework, each WSI is regarded as a bag with numerous cropped patches as instances. The features of instances are then extracted and aggregated to produce a bag-level representation for the followed tasks. Existing MIL-based methods have shown effectiveness in WSI analysis [8], [9], [10], [11], [12]. However, these works generally focus on the single resolution of WSIs, and ignore the important multi-resolution information.

Inspired by the diagnosis process of pathologists, several works have extended the previous single-resolution based MIL frameworks to the multi-resolution oriented approaches [13], [14], [15], [16], and then achieve superior performance. However, different resolutions of WSIs present quite different diagnostic information ranging from tissue-scale to cellular-scale, resulting in a semantic gap among different resolution patches [15], [16]. Existing multi-resolution works do not pay enough attention to this intrinsic semantic gap, and mainly perform simple concatenation or direct message passing to fuse features extracted from patches with different resolutions. To this end, instead of directly using these patch-level features, we introduce a non-patch agent for each resolution to exchange information.

Transformer has been widely used in various vision tasks [17], [18], [19], which can capture the correlation between different tokens in a sequence to learn long-range information. Compared with CNN, Transformer introduces a learnable class token to aggregate information from all patch tokens. Recently, several multi-scale vision Transformers (MVTs) have adopted class tokens to effectively learn and fuse multi-scale features [20], [21], [22]. Since class tokens uniformly convert different scale features into discriminative representations for classification, they can effectively alleviate the semantical gap in different scales. However, the existing MVTs are mainly developed for natural images with small sizes. While WSIs are the high-resolution scans of tissue sections, whose full spatial sizes can be over  $100000 \times 100000$  pixels at  $40\times$  resolution. Therefore, it is a

Manuscript received 25 May 2023; revised 22 August 2023; accepted 12 September 2023. Date of publication 19 September 2023; date of current version 6 December 2023. This work was supported in part by the National Key R&D Program of China under Grant 2021YFA1003004, in part by the National Natural Science Foundation of China under Grants 62271298 and 81871428, and in part by the 111 Project under Grant D20031. (Corresponding author: Jun Shi.)

Saisai Ding, Juncheng Li, Jun Wang, and Jun Shi are with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: dingsaisai0114@shu.edu.cn; junchengli@shu.edu.cn; wangjun\_shu@shu.edu.cn; junshi@shu.edu.cn).

Shihui Ying is with the Department of Mathematics, School of Science, Shanghai University, Shanghai 200444, China (e-mail: shyg@shu.edu.cn).

Digital Object Identifier 10.1109/JBHI.2023.3317067

time-consuming and tedious task to apply existing MVTs to WSIs directly.

Recently, many efficient vision Transformer (EVT) models have been proposed to reduce computational complexity recently [23], [24]. Among them, token pruning is one of the most direct and effective ways to remove redundant tokens, which contain semantically meaningless or distracting image backgrounds. The current strategies for token pruning mainly focus on retaining attentive and removing inattentive tokens. These studies have shown that not all patch tokens in the Transformer model contribute to the final prediction [23], [24]. Since there are abundance of similar and potentially redundant patches in WSIs, the token pruning strategy can be used to reduce redundant instances in bags, thus applying existing MVT to MIL-based WSI analysis.

On the other hand, the spatial information in WSIs has great diagnostic significance for WSI analysis. For example, since the tumor cells and tissues of different cancer subtypes often have great similarities, the spatial relationship between the tumor and its surrounding tissue patches requires be fully considered for more accurate cancer diagnosis. Recently, several graph-based MIL methods [25], [26], [27] regard the WSI as a graph-based data structure, where the nodes correspond to patches and the edges are constructed between adjacent patches. The constructed graph can be fed into the graph convolutional network (GCN) to learn the spatial relationships among different regions in a WSI. Therefore, integrating the GCN into Transformer to capture spatial-related information of WSIs for more efficient representation learning is feasible.

In this work, we propose a Multi-scale Efficient Graph-Transformer (MEGT) framework to effectively fuse the multi-scale information of WSIs for more accurate diagnosis. Specifically, MEGT adopts a dual-branch efficient Graph-Transformer (EGT) to process the low-resolution and high-resolution patch embeddings (i.e., tokens in the Transformer), respectively. Meanwhile, a multi-scale feature fusion module (MFFM) is developed to fuse these tokens. The proposed EGT integrates the graph representation into the Transformer to capture the spatial information of WSIs. Moreover, to accelerate EGT computation, a novel token pruning module (TPM) is developed to reduce the redundant tokens. MFFM introduces a non-patch token for each branch as an agent to exchange information with another branch by a specially designed cross-attention. The experimental results on two public WSIs datasets indicate the effectiveness of the proposed MEGT.

The main contributions of this work are as follows:

- 1) We propose EGT, a new graph-Transformer structure to learn both the spatial information of WSIs and the correlations between image patches by innovatively embedding GCN into Transformer. Thus, the EGT can work as an effective backbone to learn superior feature representations.
- 2) A simple yet effective TPM is developed in the EGT to reduce the redundant tokens without introducing additional parameters. It can significantly expedite EGT computation and preserve the most important tokens for model training.

- 3) A new MFFM is developed to fuse multi-resolution features, which uses the non-patch token as an agent for each resolution to exchange information via a cross-attention mechanism. Thus, it can effectively alleviate the semantic gap in different resolution patches.

## II. RELATED WORK

### A. Multiple Instance Learning for WSI

MIL has been successfully applied to WSI analysis, such as cancer diagnosis and survival prediction [28], [29]. According to the network structures, existing MIL methods for WSI analysis generally can be divided into two categories [30]: structure-specific and structure-free models.

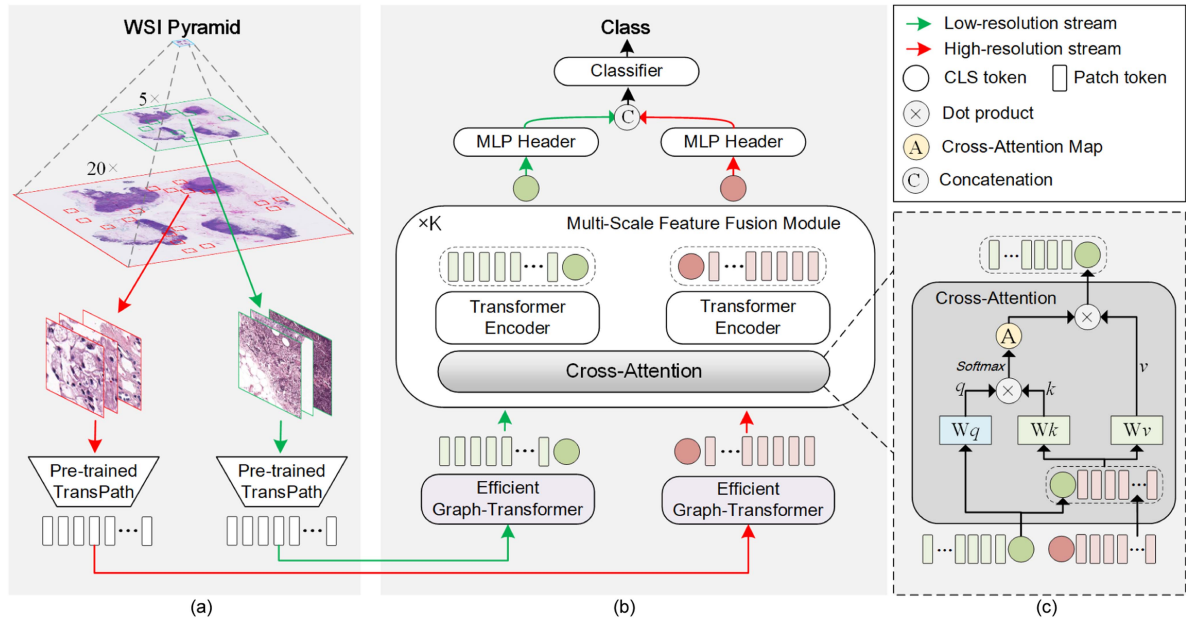
The structure-specific model usually adopts the GCN to learn the spatial information from WSI. For example, DeepGraphSurv applied the GCN to WSIs for survival prediction, in which the extracted patches were adopted as nodes for graph construction [25]; Li et al. [31] proposed a context-aware GCN for survival analysis by using the true spatial coordinates of WSIs to connect edges between adjacent patches. These structure-specific models can represent the contextual information between patches for learning the spatial representation of WSIs.

The structure-free model is generally developed based on the attention-MIL. It uses attention scores to weight the instance embeddings to learn the slide-level representation. For example, Li et al. [15] proposed a dual-stream MIL (DSMIL) for WSI classification, in which the attention scores for each instance were computed with the critical instance; Shao et al. [19] designed a Transformer-based MIL (TransMIL) for WSIs classification, it utilized self-attention to capture long-range dependencies between patches; Huang et al. [17] combined self-supervised learning and Transformer to generate superior feature representation for survival analysis with WSIs. The structure-free model can capture the correlation among different patches to improve the performance of the WSI-based CADs.

Different from the previous MIL-based methods, we aim to explore a novel Graph-Transformer structure to combine the strengths of the above two models, which integrates GCN into Transformer to learn both the spatial information of WSIs and the long-range dependencies between image patches.

### B. Multi-Scale WSI Analysis

In recent years, the multi-scale oriented WSI analysis has attracted more attention. Compared with the single-scale method, the multi-scale approach can learn more semantic information for classification tasks [7], [13], [14], [15], [30]. For example, Campanella et al. [7] trained different MIL branches for different resolutions and then used a max-pooling operation to fuse these multi-resolution embedding for learning the WSI-level representation; Hashimoto et al. [14] mixed patches from different resolutions into a bag and then fed it to the MIL model; Li et al. [15] adopted a pyramidal concatenation strategy to spatially align patches from different resolutions for WSIs classification; Liu et al. [30] proposed a square pooling layer to align patches from two resolutions, which spatially pooled patches



**Fig. 1.** Overview of MEGT for WSI classification. (a) WSI processing and feature extraction. A WSI pyramid is divided into non-overlapping patches at low and high resolutions, and then their features are extracted by a pre-trained TransPath model, respectively. (b) Flowchart of MEGT. The multi-resolution patch embeddings are fed into the proposed MEGT framework, equipped with the efficient Graph-Transformer and multi-scale feature fusion module, to learn slide-level representation for WSI classification. (c) Cross-attention operation for the low-resolution branch. The CLS token of the low-resolution branch is used as a query token to interact with the patch tokens from the high-resolution branch through cross-attention. The high-resolution branch follows the same procedure, but swaps CLS and patch tokens from another branch.

from high-resolution under the guidance of low-resolution; Hou et al. [13] also proposed a heterogeneous graph neural network for tumor typing and staging, in which the heterogeneous graph was constructed based on the feature and spatial-scaling relationship of the multi-resolution patches. These works demonstrate that the multi-scale features can learn more effective slide-level representation to improve the performance of WSI analysis.

However, existing methods usually directly combine features from different scales without considering the semantic differences between different resolution patches. Thus, we propose to introduce a non-patch agent for each resolution to exchange information.

### C. Multi-Scale Vision Transformer

Inspired by the feature pyramid of images in CNNs, the MVTs have been designed to learn multi-scale visual representations from images [21], [22], [32], [33]. For example, Zhang et al. [21] proposed a nested hierarchical Transformer, which stacked Transformer layers on non-overlapping image blocks independently, and then nested them hierarchically; Chen et al. [22] developed a cross-attention multi-scale Transformer for image classification, which used the class token as an agent to exchange information among different branch; Liu et al. [32] proposed a general Transformer backbone that provided hierarchical feature representations for computer vision; Fan et al. [33] designed a multi-scale vision Transformer for video and image recognition, which hierarchically expanded the feature complexity while reducing visual resolution; These works indicate that the MVTs can learn more effective feature representation for different computer vision tasks.

However, the existing MVTs algorithms are mainly developed for natural images with small sizes. Since WSIs have an extremely large size, these algorithms cannot be directly applied to gigapixel WSIs. Therefore, we investigate how to learn multi-scale feature representations in Transformer models for WSI classification.

## III. METHODOLOGY

In this work, a novel MEGT framework is proposed for WSIs classification, which can effectively exploit the multi-scale feature information of WSIs. Given a WSI pyramid, our framework aggregates patch features of different resolutions to implement the slide-level prediction. As shown in Fig. 1, a WSI pyramid is first cropped into non-overlapping patches at low and high resolutions, and their features are extracted by a pre-trained TransPath [34], respectively. Then, the multi-resolution patch embeddings will be fed into the proposed MEGT framework, which consists of two EGT branches and a MFFM, to extract discriminative representation and fully mine multi-scale information. Finally, a WSI-level classifier is employed to generate the slide-level prediction based on the learned representation. In the following subsections, we will introduce the EGT, MFFM, and the learning strategy of the whole framework in detail.

### A. Efficient Graph-Transformer

As shown in Fig. 2, the EGT contains two Transformer encoders, a TPM, and a Graph-Transformer layer (GTL). The first Transformer encoder adopts class token to learn the global information of patch tokens and provides attention scores for token pruning. Then, the token pruning module selects Top- $k$



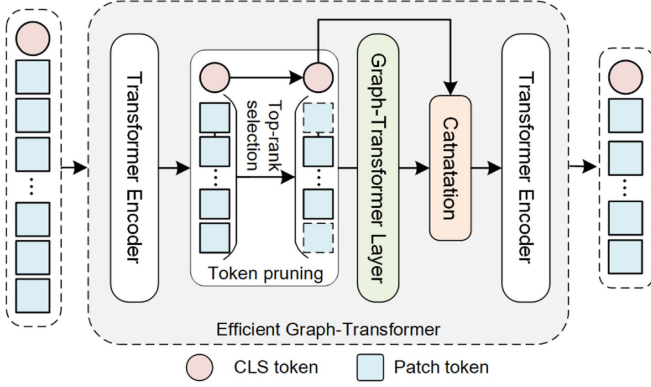


Fig. 2. Structure of Efficient Graph-Transformer (EGT), which is composed of two Transformer encoder layers, a token pruning module, and a Graph-Transformer layer.

most important tokens according to the attention scores to reduce the number of tokens. Subsequently, the Graph-Transformer layer uses these selected tokens to simultaneously learn the local and global information of the WSIs. Finally, the class token learns all the information again through the second Transformer encoder for the subsequent MFFM.

1) **Transformer Encoder:** A Transformer encoder [35] is employed to learn potential long-term dependencies between instances. It contains multiple Transformer layers, each of which has a multi-head self-attention (MSA) and a multi-layer perceptron (MLP). MSA uses the self-attention mechanism to calculate the correlation matrix between instances, and the complexity of memory and time are both  $O(n^2)$ . In WSI processing, a WSI may be divided into tens of thousands of patches. To address the issue of long instance sequence, we employ the Nystrom-attention (NA) method [36] here, which utilizes the Nystrom method to approximate the standard self-attention. The NA method can be defined as:

$$Q = XW_q, K = XW_k$$

$$NA = \text{softmax} \left( \frac{Q\tilde{K}^T}{\sqrt{d}} \right) \left( \text{softmax} \left( \frac{\tilde{Q}\tilde{K}^T}{\sqrt{d}} \right) \right)^\dagger$$

$$\times \text{softmax} \left( \frac{\tilde{Q}K^T}{\sqrt{d}} \right) \quad (1)$$

where  $W_q$  and  $W_k$  are learnable parameters,  $Q$  and  $K \in \mathbb{R}^{n \times d}$ ,  $d$  is the patch embedding feature dimension,  $\tilde{Q}$  and  $\tilde{K} \in \mathbb{R}^{m \times d}$  are the  $m$  selected landmarks from the  $Q$  and  $K$ ,  $A^\dagger$  is the Moore-Penrose inverse of  $A$ . When  $m$  is much less than  $n$ , the computational complexity is reduced from  $O(n^2)$  to  $O(n)$ . The output of the  $i$ -th Transformer layer can be defined as:

$$T'_i = MSA(LN(T_{i-1})) + T_{i-1}$$

$$T_i = MLP(LN(T'_i)) + T'_i \quad (2)$$

where LN is Layer normalization.

2) **Token Pruning Module:** Although the above NA solves the long-sequence problem in Transformer, the computational

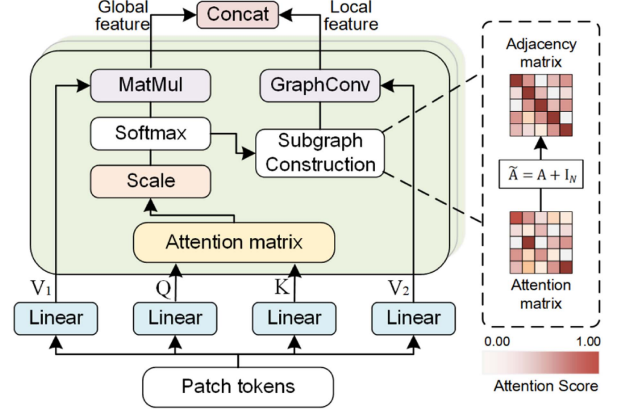


Fig. 3. Structure of Graph-Transformer layer.

complexity is still heavy, when all tokens are used to construct a graph in GCN. Therefore, we perform token pruning to reduce redundant tokens.

Let  $n$  denote the number of patch tokens, and an extra class token is added for classification in the Transformer. The interactions between class and other tokens are performed via the attention mechanism in NA, and an attention map  $A \in \mathbb{R}^{(n+1) \times (n+1)}$  is obtained, in which the first row of  $A$  represents the attention score  $a = A[0, 1:] \in \mathbb{R}^{1 \times n}$  from class to all patch tokens. Thus, the attention scores are used to determine the importance of each patch token.

In the multi-head self-attention layer, there are multiple class attention vectors  $a_h$ , where  $h = [1, \dots, H]$ , and  $H$  is the total number of attention heads. We compute the average value of all heads by:

$$\bar{a} = \sum_{h=1}^H a_h / H \quad (3)$$

After that, we select the tokens corresponding to the  $k$  largest (top- $k$ ) elements in  $\bar{a}$ , and further fuse the other tokens into a new token using the attention scores in  $\bar{a}$ . Therefore, the token fusion can be written as:

$$x_{fusion} = \sum_{i=1}^{i=n-k} x_i \bar{a}_i \quad (4)$$

where  $x_i$  is the  $i$ -th patch token, and  $\bar{a}_i$  represents the  $i$ -th attention score in  $\bar{a}$ .

3) **Graph-Transformer Layer:** As shown in Fig. 3, GTL has two branches, one branch performs the self-attention operation of Transformer to learn the correlation between image patches, and the other branch performs graph convolution to learn the spatial information of WSIs. Furthermore, instead of using the  $K$ -NN algorithm to construct the adjacency matrix based on a pair of node features, we adopt the attention matrix in self-attention to adaptively generate the adjacency matrix, thereby further speeding up the training of the network.

After token pruning, a total  $(k+1)$  patch tokens are used in the GTL. Note that we do not use class token, since it may affect the learning of GCN. Given an input patch embedding  $X_{patch} \in \mathbb{R}^{(k+1) \times d}$ , the matrices of query  $Q$ , key  $K$  and values  $V_1, V_2$

are first calculated through four different linear projections as follows:

$$\begin{aligned} Q &= \text{LinearProj}_1 (X_{patch}) = X_{patch} W_Q, \\ K &= \text{LinearProj}_2 (X_{patch}) = X_{patch} W_K, \\ V_1 &= \text{LinearProj}_3 (X_{patch}) = X_{patch} W_{V_1}, \\ V_2 &= \text{LinearProj}_4 (X_{patch}) = X_{patch} W_{V_2}. \end{aligned} \quad (5)$$

where  $W_Q$ ,  $W_K$ ,  $W_{V_1}$ , and  $W_{V_2} \in \mathbb{R}^{d \times d_m}$  are the corresponding weight matrices of linear projections.

In addition, we also adopt the multi-head structure to expand the GTL. As shown in Fig. 3, this structural design can project the inputs into different subspaces to learn different features, thereby improving the performance of the model. Specifically, the input features are evenly split into  $h$  parts, and the attention matrix can be calculated as follows:

$$A_i = \text{Score} (Q_i, K_i) = \frac{Q_i K_i^T}{\sqrt{d_m/h}} \quad (6)$$

where  $A_i \in \mathbb{R}^{(k+1) \times (k+1)}$ ,  $i = [1, \dots, h]$ ,  $Q_i \in \mathbb{R}^{(k+1) \times \frac{d_m}{h}}$ ,  $K_i \in \mathbb{R}^{(k+1) \times \frac{d_m}{h}}$ , and  $1/\sqrt{d_m/h}$  is a scaling factor.

For the Transformer branch, the outputs of the multi-head structure are first concatenated together and then fed into linear projections to obtain the complete outputs:

$$\begin{aligned} V'_{1i} &= \text{softmax} (A_i) V_{1i} \\ V'_1 &= \text{Concat} (V'_{11}, \dots, V'_{1h}) W_{o1}. \end{aligned} \quad (7)$$

where  $V'_{1i} \in \mathbb{R}^{(k+1) \times \frac{d_m}{h}}$ ,  $V'_1 \in \mathbb{R}^{(k+1) \times d}$ , and  $W_{o1} \in \mathbb{R}^{d_m \times d}$  the weight matrices of linear projection.

For the GCN branch, the adjacency matrix  $\tilde{A}_i$  is first transformed by the normalized  $A_i$  with self-connections, and then the  $\tilde{A}_i \in \mathbb{R}^{(k+1) \times (k+1)}$  and node embedding  $V_{2i} \in \mathbb{R}^{(k+1) \times \frac{d_m}{h}}$  are fed into the GCN [37] to learn graph representations. The forward propagation of GCN can be written as follows:

$$V'_{2i} = \sigma \left( \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} V_{2i} W_l \right) \quad (8)$$

where  $V'_{2i} \in \mathbb{R}^{(k+1) \times \frac{d_m}{h}}$ ,  $\tilde{D}$  is the degree matrix of  $\tilde{A}$ , and  $W_l$  is trainable weight matrix. After that, the multiple subgraph representations are concatenated together and then fed into the linear projections to obtain the complete outputs:

$$V'_2 = \text{Concat} (V'_{21}, \dots, V'_{2h}) W_{o2} \quad (9)$$

where  $V'_2 \in \mathbb{R}^{(k+1) \times d}$  and  $W_{o2} \in \mathbb{R}^{d_m \times d}$  are the weight matrices of linear projection.

The adjacency matrix can effectively represent the spatial distribution and adjacent relationship between nodes, and then the spatial-related information is learned through GCN layer. Thus, the generated graph representation contains the local information and short-range structure, which are ignored in the original Transformer.

Finally, the Transformer branch output  $V'_1$  and GCN output  $V'_2$  are concatenated together and then fed into linear projections to fuse the local and global information:

$$X'_{patch} = \text{Concat} (V'_1, V'_2) W_{o3} \quad (10)$$

where  $X'_{patch} \in \mathbb{R}^{(k+1) \times d}$  and  $W_{o3} \in \mathbb{R}^{2d \times d}$  are the weight matrices of linear projection.

## B. Multi-Scale Feature Fusion Module

As mentioned before, the patches with different resolutions present different diagnostic information ranging from cellular-scale (e.g., nucleus and micro-environment) to tissue-scale (e.g., vessels and glands), which has an intrinsic semantical gap. However, existing multi-scale methods have not paid enough attention to this issue. To this end, inspired by several previous multi-scale Transformers [20], [22], we propose a novel MFFM, which alleviates the semantical gap by using the class token as an agent to exchange information between two resolutions.

As shown in Fig. 1(b), MEGT contains  $K$  MFFM, each of which consists of two Transformer encoders and an efficiently cross-attention, and the value of  $K$  is set to 2 in this work. Since the class token uniformly converts different resolution features into discriminative representation for classification, we use class tokens to effectively fuse features of different scales via a Cross-Attention (CA). Fig. 1(c) shows the basic idea of CA, which uses the class token of a branch to exchange information with patch tokens of the other branch. Since the class token already learns the global information of all patch tokens in the corresponding branch, interacting with patch tokens of another branch can help to learn more information at different scales. After the CA, the class token passes the learned information to its patch tokens at the later Transformer encoder, thereby enriching the representation of patch tokens.

Let  $X^i = [x_{CLS}^i | X_{patch}^i]$  be the token sequence at branch  $i \in [\text{low-resolution}, \text{high-resolution}]$ , where  $x_{CLS}^i$  and  $X_{patch}^i$  represent the class and patch token of branch  $i$ , respectively. For the low-resolution branch, we first collect  $X_{patch}^h$  from high-branch, and then concatenate it with  $x_{CLS}^l$  to obtain the  $X' = [x_{CLS}^l | X_{patch}^h]$ . Then,  $x_{CLS}^l$  is used as the query and  $X'$  works as the key and value to perform the CA operation, which can be expressed as:

$$\begin{aligned} Q &= x_{CLS}^l W_q, K = X' W_k, V = X' W_v \\ CA &= \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V \end{aligned} \quad (11)$$

where the  $x_{CLS}^l \in \mathbb{R}^{1 \times d}$ , so the computational complexity of CA is  $O(n+1)$  instead of  $O(n^2)$  for self-attention,  $n$  is the number of patch tokens in the high-branch. The output  $O^l$  with the CA operation can be expressed as:

$$\begin{aligned} y_{CLS}^l &= x_{CLS}^l + CA(x_{CLS}^l, X') \\ O^l &= [y_{CLS}^l | X_{patch}^l] \end{aligned} \quad (12)$$

TABLE I

ARCHITECTURE OF OUR PROPOSED MEGT.  $M$  AND  $N$  ARE THE NUMBERS OF PATCH TOKENS IN THE  $5\times$  AND  $20\times$  BRANCHES, RESPECTIVELY.  $M'$  AND  $N'$  ARE THE NUMBERS OF REMAINED PATCH TOKENS AFTER TPM IN THE  $5\times$  BRANCH AND  $20\times$  BRANCHES, RESPECTIVELY.  $C$  IS THE NUMBER OF CLASSES

Module	Layer	Input size		Output size	
		$5\times$ branch	$20\times$ branch	$5\times$ branch	$20\times$ branch
EGT	FC_1	$1 \times M \times 768$	$1 \times N \times 768$	$1 \times M \times 256$	$1 \times N \times 256$
	Transformer_1	$1 \times (M+1) \times 256$	$1 \times (N+1) \times 256$	$1 \times (M+1) \times 256$	$1 \times (N+1) \times 256$
	Token pruning	$1 \times M \times 256$	$1 \times N \times 256$	$1 \times M' \times 256$	$1 \times N' \times 256$
	Graph-Transformer	$1 \times M' \times 256$	$1 \times N' \times 256$	$1 \times M' \times 256$	$1 \times N' \times 256$
	Transformer_2	$1 \times (M'+1) \times 256$	$1 \times (N'+1) \times 256$	$1 \times (M'+1) \times 256$	$1 \times (N'+1) \times 256$
MFFM	Cross-attention_1	$1 \times 1 \times 256,$ $1 \times (N'+1) \times 256$	$1 \times 1 \times 256,$ $1 \times (M'+1) \times 256$	$1 \times 1 \times 256$	$1 \times 1 \times 256$
	Transformer_3	$1 \times (M'+1) \times 256$	$1 \times (N'+1) \times 256$	$1 \times (M'+1) \times 256$	$1 \times (N'+1) \times 256$
	Cross-attention_2	$1 \times 1 \times 256,$ $1 \times (N'+1) \times 256$	$1 \times 1 \times 256,$ $1 \times (M'+1) \times 256$	$1 \times 1 \times 256$	$1 \times 1 \times 256$
	Transformer_4	$1 \times (M'+1) \times 256$	$1 \times (N'+1) \times 256$	$1 \times (M'+1) \times 256$	$1 \times (N'+1) \times 256$
Classifier	FC_2	$1 \times 1 \times 256$	$1 \times 1 \times 256$	$1 \times 1 \times 256$	$1 \times 1 \times 256$
	FC_3 (fusion)	$1 \times 1 \times 512$		$1 \times 1 \times 256$	
	FC_4	$1 \times 1 \times 256$		$1 \times C$	

The high-resolution branch follows the same procedure, but  $\mathbf{x}_{CLS}^h$  is used as the query and  $\mathbf{X}' = [\mathbf{x}_{CLS}^h \| \mathbf{X}_{patch}^l]$  is used as the key and value. The output  $\mathbf{O}^h$  with the CA operation can be expressed as:

$$\begin{aligned} \mathbf{y}_{CLS}^h &= \mathbf{x}_{CLS}^h + CA(\mathbf{x}_{CLS}^h, \mathbf{X}') \\ \mathbf{O}^h &= [\mathbf{y}_{CLS}^h \| \mathbf{X}_{patch}^h] \end{aligned} \quad (13)$$

note that the weights of the CAs in the two branches are not shared, because each branch needs to exchange information separately with the other.

### C. Network Architecture and Training Strategy

EGT and MFFM are the basic components of the proposed MEGT. As shown in Fig. 1(b), MEGT contains two separate branches, and EGT is used in each branch to provide superior feature representation for the MFFM. Then, MFFM employs the class tokens to fuse multi-scale features multiple times via a cross-attention layer. Finally, the dual-resolution class tokens are concatenated to produce slide-level representation  $\mathbf{z}$  for WSI classification, which can be defined by:

$$\begin{cases} \mathbf{z} = \text{Concat}(\mathbf{x}_{CLS}^l, \mathbf{x}_{CLS}^h) \\ \hat{Y} = \text{softmax}(MLP(\mathbf{z})) \end{cases} \quad (14)$$

For the model training, the loss function  $\mathcal{L}$  is defined by the cross entropy between the bag class labels  $Y$  and bag class predictions  $\hat{Y}$  which can be expressed as:

$$\mathcal{L} = -\frac{1}{M} \sum_{i=1}^M \sum_{j=1}^C Y_{ij} \log(\hat{Y}_{ij}) \quad (15)$$

where  $M$  is the number of patients,  $C$  is the number of classes. The gradient descent algorithm is adopted for optimizing the whole model. The architecture of the proposed framework is shown in Table I.

## IV. EXPERIMENTS AND RESULTS

### A. Datasets

The proposed MEGT was evaluated on two commonly used WSI datasets, namely the Cancer Genome Atlas Renal Cell Carcinoma (TCGA-RCC) dataset (<https://portal.gdc.cancer.gov>) and CAMELYON16 dataset [38].

TCGA-RCC is a WSI dataset for Renal Cell Carcinoma classification, and it contains three categories: Kidney Chromophobe Renal Cell Carcinoma (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), and Kidney Renal Papillary Cell Carcinoma (KIRP). A total of 914 slides are collected from 876 cases, including 111 KICH slides from 99 cases, 519 KIRC slides from 513 cases, and 284 KIRP slides from 264 cases. After WSI pre-processing, the mean numbers of patches extracted on each slide at  $5\times$  and  $20\times$  magnification were 4263 and 14627, respectively.

CAMELYON16 is a public dataset for metastasis detection in breast cancer, including 270 training and 129 test slides. After WSI pre-processing, there were about mean 918 and 3506 patches selected from each slide at  $5\times$  and  $20\times$  magnifications, respectively.

### B. Experiment Setup

In our experiments, for the TCGA-RCC dataset, we conducted a 5-fold cross-validation on the 936 slides to evaluate these algorithms. The results on this dataset were given in the format of mean  $\pm$  SD (standard deviation). For CAMELYON16 dataset, all algorithms were trained on the 270 official training slides and tested on the 129 official test slides. Thus, the results calculated on the individual testing set did not have the mean and SD values. The widely used accuracy, F1-score (F1), and area under the curve (AUC) were used as evaluation indices to compare the classification performance of different algorithms. Since the classification algorithms consistently achieved better results

TABLE II

COMPARISON RESULTS ON THE CAMELYON16 AND TCGA DATASETS. WE USE THE SAME PATCH FEATURE EXTRACTOR FOR ALL METHODS (UNIT: %)

Method	Resolution	TCGA-RCC			CAMELYON16			FLOPs	Params	Inference (bags/s)
		ACC	F1	AUC	ACC	F1	AUC			
Patch-GCN [31]	20×	93.28±1.10	91.90±1.29	94.57±1.33	93.02	90.53	93.80	573M	1.35M	67
TransMIL [19]	20×	93.94±1.09	93.43±0.75	94.89±1.67	94.57	92.93	95.53	614M	2.66M	77
GTP [39]	20×	94.27±1.12	93.44±1.58	95.01±1.49	95.35	93.75	95.79	778M	1.84M	9
<b>EGT (Ours)</b>	20×	<b>95.37±0.64</b>	<b>94.73±0.55</b>	<b>95.91±1.19</b>	<b>96.12</b>	<b>95.20</b>	<b>96.34</b>	215M	1.09M	33
Patch-GCN (GBU)	20×+5×	90.21±1.45	88.97±1.24	91.24±1.67	89.92	84.71	90.43	615M	2.75M	59
TransMIL (GBU)	20×+5×	94.63±1.34	93.89±1.37	95.43±1.97	95.35	94.84	96.01	2.43G	5.87M	65
GTP (GBU)	20×+5×	91.78±1.62	90.59±1.57	92.19±1.04	93.02	90.72	94.53	957M	3.74M	4
TransMIL (CA)	20×+5×	95.24±1.27	94.76±1.53	96.19±1.41	96.12	94.65	96.65	2.85G	6.76M	31
GTP (CA)	20×+5×	94.89±1.37	93.79±1.55	95.34±1.28	95.35	94.00	94.71	1.38G	5.53M	2
H <sup>2</sup> -MIL [13]	20×+5×	95.44±0.96	94.89±1.34	96.27±1.41	96.12	94.74	96.70	1.29G	2.11M	3
<b>MEGT (Ours)</b>	20×+5×	<b>96.91±1.24</b>	<b>96.26±1.19</b>	<b>97.89±1.58</b>	<b>96.89</b>	<b>95.74</b>	<b>97.30</b>	423M	4.28M	18

The best classification results are highlighted in bold.

on 20× images than those on 5× ones, we only reported the experimental results on a single 20× scale for both datasets.

### C. Implementation Details

In WSI pre-processing, each WSI was divided into non-overlapping  $299 \times 299$  patches in both the magnifications of 20× and 5×, and a threshold was set to filter out background patches. After patching, we used the pre-trained TransPath [34] model, a pre-training vision Transformer for histopathology images, to extract a feature vector with a dimensional of 768. Thereafter, for the proposed MEGT, the Adam optimizer was used in the training stage with a learning rate of 1e-4 with a weight decay of 1e-5, to update the models. The size of the mini-batch was set as 1 (bag). The MEGT and other models were trained for 150 epochs with a cross-entropy loss function, and they would early stop if the loss would not decrease in the past 30 epochs. All models were implemented by Python 3.7 with PyTorch toolkit 1.10.0 on a platform equipped with one NVIDIA GeForce RTX 3090 GPU.

### D. Comparison Experiment

Since the proposed MEGT leverages graph-based learning, Transformer architectures, and multi-resolution learning, we compared it with the following SOTA MIL algorithms:

- 1) Patch-based GCN (Patch-GCN) [31]: It is a graph-based MIL algorithm that uses the true spatial coordinates of WSIs to connect edges between adjacent patches.
- 2) Transformer-based MIL (TransMIL) [19]: It is a Transformer-based MIL algorithm that approximates patch self-attention with Nyström method for WSI classification.
- 3) Graph-Transformer for Processing Pathology Images (GTP) [39]: It is a novel graph-Transformer MIL algorithm that integrates a graph-based representation of WSIs and a vision Transformer.
- 4) Hierarchical Representation with Heterogeneous MIL (H<sup>2</sup>-MIL) [13]: It is a novel multi-resolution MIL algorithm that constructs a heterogeneous graph with different

resolutions to learn the hierarchical representation of WSIs.

We also compare the proposed MEGT with the following baselines for multi-resolution learning:

- 5) Gated Bimodal Unit (GBU): We combine Patch-GCN [31], TransMIL [19], and GTP [39] of different resolutions using the GBU operator [40], denoted as Patch-GCN (GBU), TransMIL (GBU), and GTP (GBU).
- 6) Cross-Attention: As TransMIL [19] and GTP [39] also introduce class tokens for classification, the Cross-Attention (CA) in MFFM can be used to exchange information between different resolutions, denoted as TransMIL (CA), and GTP (CA).

Table II shows the classification results of different algorithms on the TCGA-RCC and CAMELYON16 datasets. It can be found that the proposed MEGT outperforms all the compared algorithms on both datasets. In particular, on the TCGA-RCC dataset, MEGT achieves the best mean accuracy of  $96.91 \pm 1.24\%$ , F1-score of  $96.26 \pm 1.19\%$ , and AUC of  $97.89 \pm 1.58\%$ . Compared to other algorithms, it improves at least 1.47%, 2.02%, and 1.32% on the corresponding indices, respectively. Similarly, MEGT outperforms all the compared algorithms with the best accuracy of 96.89%, F1-score of 95.74%, and AUC of 97.30% on the CAMELYON16 dataset, improves by 0.77%, 1.00%, and 0.60% on corresponding indices, respectively.

1) *Single-Scale vs. Multi-Scale Learning*: The multi-scale MIL algorithms achieve better results than the single-scale MIL ones, which indicates that the multi-scale information of WSIs is important for cancer diagnosis. As a SOTA multi-resolution MIL algorithm, H<sup>2</sup>-MIL achieves the second-best performance due to a newly proposed GCN algorithm in it, which can learn hierarchical representation from a heterogeneous graph. Nevertheless, our MEGT still outperforms H<sup>2</sup>-MIL.

2) *Graph-Based vs. Transformer-based vs. Graph-Transformer*: The Graph-Transformer MIL algorithms significantly outperform graph-based and Transformer-based ones on both datasets. This observation supports our design choice of using the graph-Transformer structure to learn both the spatial information of WSIs and the correlation between patches.



TABLE III  
ABLATION STUDY RESULTS FOR EVALUATING THE TPM AND GTL IN EGT ON THE CAMELYON16 AND TCGA DATASETS (UNIT: %)

Method	TPM	GTL	TCGA-RCC			CAMELYON16			Aver. Sec. (Epoch)		Params
			ACC	F1	AUC	ACC	F1	AUC	TCGA	Came16	
EGT-m	×	×	93.83±1.11	92.65±1.09	94.65±1.17	93.80	93.17	94.72	42s	27s	495K
EGT-TPM	√	×	94.41±1.32	92.88±1.28	95.38±1.34	94.57	93.78	95.55	29s	18s	495K
EGT-GTL	×	√	94.58±1.03	93.24±1.11	95.49±1.24	95.35	94.40	95.69	65s	39s	1.09M
EGT-KNN	√	√	95.21±0.98	94.53±1.24	95.61±1.58	95.35	94.31	95.77	141s	101s	1.09M
EGT	√	√	<b>95.37±0.64</b>	<b>94.73±0.55</b>	<b>95.91±1.19</b>	<b>96.12</b>	<b>95.20</b>	<b>96.34</b>	38s	26s	1.09M

The best classification results are highlighted in bold.

3) *GBU vs. Cross-Attention Baselines*: The Cross-Attention baselines perform better than the GBU ones on both datasets. We attribute this observation to the use of class tokens, which can alleviate the semantic gap across different scales.

4) *Computational Complexity*: For the single-resolution experiment at  $20\times$  resolution, the proposed EGT outperforms all the other single-resolution MIL algorithms on all indices with fewer model parameters and FLOPs. Due to the integration of GCN and Transformer architectures, EGT has a longer inference time compared to Patch-GCN and TransMIL but is still significantly shorter than GTP. For the multi-resolution experiment, our MEGT model parameters are larger than those of  $H^2$ -MIL, but its FLOPs are much smaller due to the significant reduction in the number of patches by our TPM. In addition, the inference time of MEGT is also significantly shorter than  $H^2$ -MIL.

## E. Ablation Study

To further evaluate the effectiveness of MEGT, we conducted a series of ablation experiments to delineate the contributions of two major components in the proposed MEGT: the EGT and the MFFM.

1) *Effects of Efficient Graph-Transformer*: To evaluate the effectiveness of the proposed token pruning module (TPM) and Graph-Transformer Layer (GTL) in EGT, we compared the proposed EGT with its three variants:

- 1) EGT-m: This variant removed both the TPM and GTL.
- 2) EGT-TPM: This variant only maintained the TPM, but removed the GTL.
- 3) EGT-GTL: This variant only maintained the GTL, but removed the TPM.
- 4) EGT-KNN: This variant used the same structure as the EGT, but used the KNN algorithm to perform the sub-graph construction.

Table III shows the results of different variants on both the TCGA-RCC and CAMELYON16 datasets. It can be observed that EGT-TPM and EGT-GTL outperform EGT-m, suggesting the effectiveness of TPM and GTL in EGT. Moreover, TPM can effectively reduce the number of redundant tokens without additional parameters, and GTL can learn more important local-global features with lower computational complexity. Therefore, by integrating TPM and GTL, EGT can efficiently learn both the spatial-related information of WSI and the correlation between

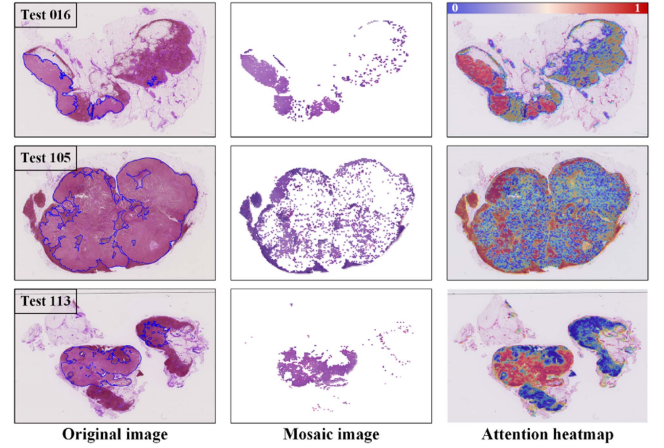


Fig. 4. Visualizations of pruning results on CAMELYON16 test set. Each representative slide is annotated by pathologists, who roughly highlighted the tumor tissue regions (left). A mosaic map is generated by overlaying remained patches in a grid-wise fashion (second column). An attention heatmap corresponding to each slide is generated by computing the attention scores of self-attention maps (right).

patch tokens for improved performance. Although the EGT-KNN achieves similar performance to our EGT, it needs more time to construct the adjacency matrix, resulting in a significant decrease in the training speed of the network.

To verify the effectiveness of TPM, we used the attention scores of the Transformer to visualize the remained patches after applying TPM. As shown in Fig. 4, the CAMELYON16 testing set with pixel-level annotations is used to evaluate the ability of our TPM to locate important patches. It can be observed that the remaining patches tend to appear within the annotated ROIs on all WSIs, indicating that our TPM can effectively remove redundant tokens and keep the most discriminative patches to expedite network training.

In addition, to further evaluate the effectiveness of the GTL, we visualized the input and output features of GTL on the randomly selected testing images from the CAMELYON16 dataset using the t-SEN [41]. As shown in Fig. 5, the images in left and right columns show the input and output features, respectively. It can be observed that the selected two features exhibit significant distinction. Compared to the input features obtained from Transformer, the output features of GTL have more local correlations between patches instead of being uniformly distributed in the feature space. The results demonstrate



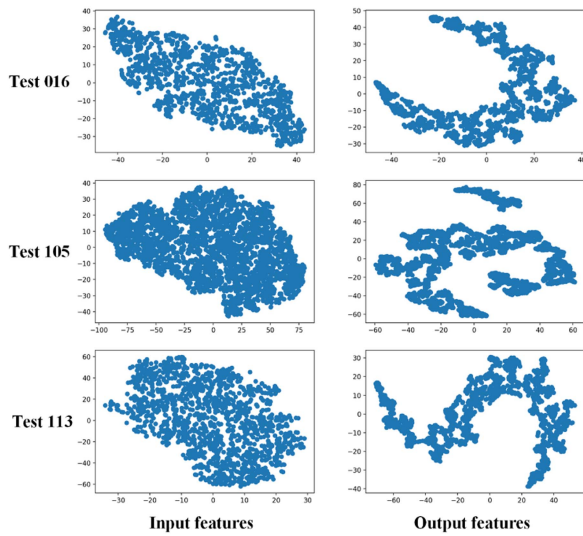


Fig. 5. Visualizations of the input and output features of GTL on the CAMELYON16 test set by using the t-SEN.

that the GTL can effectively capture the spatial information in WSIs to improve the feature representation of Transformer.

**2) Effects of Multi-Scale Feature Fusion Module:** We compared our cross-attention fusion strategy to several other feasible multi-scale fusion methods, including:

- 1) Concatenation: It removed the cross-attention in the MFFM, and then only concatenated the class tokens of the two branches after MFFM.
- 2) All-attention: It concatenated all tokens from different branches together and then fused them via the MSA.
- 3) Class-Token: It simply averaged the class tokens on the two branches, so that the information was passed back to patch tokens through the later transformer encoder.

Fig. 6 shows the classification results of four different token fusion strategies on the TCGA-RCC and CAMELYON16 datasets. The specially designed cross-attention outperforms all the other strategies on all indices, which indicates it can effectively fuse multi-scale information of WSIs with superior performance. It is worth noting that although the All-attention mechanism uses additional self-attention to learn information between two branches, it fails to achieve better performance compared to the simple Class-Token. The experimental results demonstrate that class tokens can avoid the semantic gap in different resolution patches, resulting in the performance improvement of the MFFM.

In addition, we further visualized the cross-attention maps to evaluate the effectiveness of MFFM on CAMELYON16 testing set. For all the cross-attention maps in MFFM, the attention weights were normalized between 0 to 1 (i.e., blue to red) in each cross-attention map. For better visualization, we dropped 80% of the smaller attention value. As shown in Fig. 7, there are a total of four cross-attention maps in two MFFMs, where the red regions in the attention heatmaps represent the highest contribution instances for classification in each bag. It can be found that the attention heatmaps in the first MFFM gradually extend from the global low-related information to the local

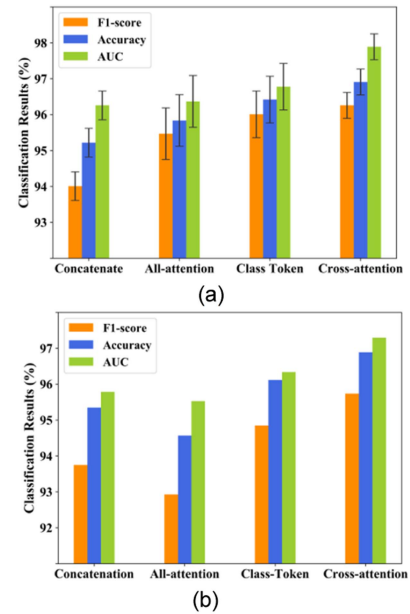


Fig. 6. Classification results for evaluating different fusion strategies in MFFM on (a) TCGA-RCC dataset and (b) CAMELYON16 dataset.

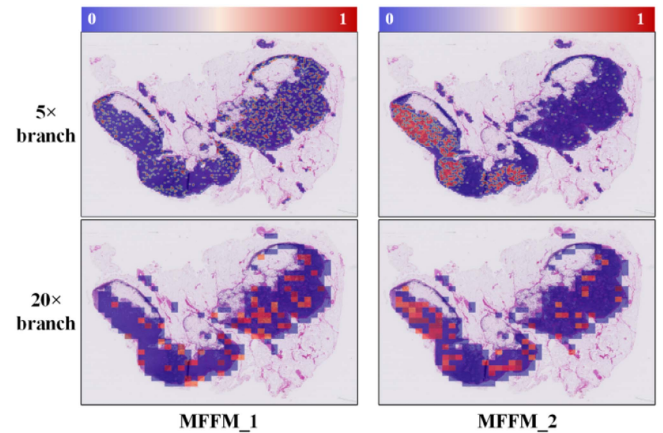


Fig. 7. Visualization of cross-attention maps in the MFFMs on CAMELYON16 testing set. MFFM\_1 and MFFM\_2 represent the first and second MFFMs.

high-related information in the second MFFM on both branches. These results demonstrate that our framework effectively fuses multi-scale information through cross-attention to implement a more accurate WSI classification.

## F. Parameter Sensitivity Analysis

A parameter sensitivity analysis was also conducted for the proposed MEGT. Two architecture parameters in MEGT will affect the classification performance, i.e., the number of Transformer layers  $L$  in MFFM and the number of MFFM  $K$ . We then changed both parameters to investigate their impact on representing different combinations of architecture parameters  $L$  and  $K$ .

Table IV shows the classification results of different numbers of Transformer layers on the low-branch and high-branch. It can be found that both models A and C significantly increase

TABLE IV

CLASSIFICATION RESULTS WITH DIFFERENT ARCHITECTURE PARAMETERS ON CAMELYON16 AND TCGA-RCC DATASETS

Model	$L$			Came16 Acc. (%)	TCGA Acc. (%)	Params. (M)
	$l$	$h$	$K$			
MEGT	1	2	2	96.89	96.91±1.24	4.28
A	<b>2</b>	2	2	96.12	96.21±1.13	4.81
B	1	<b>1</b>	2	95.35	95.98±1.38	3.76
C	1	<b>4</b>	2	96.12	96.67±1.26	5.34
D	1	2	<b>1</b>	94.57	95.77±1.85	3.23
E	1	2	<b>4</b>	95.35	96.79±1.47	6.39

THE BLACK COLOR INDICATES CHANGES FROM MEGT.

parameters but without any improvement in accuracy compared to the original MEGT, because more Transformer layers lead to a larger number of parameters, which may suffer from the overfitting problem. It is worth noting that the performance of MEGT will be decreased by reducing the depth of the high-branch in model B, which indicates that the high-branch plays the main role in learning the features of WSIs, while the low-branch only provides additional information.

The number of MFFM  $K$  is an important parameter in our MEGT, which controls the fusion frequency of the two branches. Table IV shows the classification results with different numbers of MFFMs on the TCGA-RCC dataset and CAMELYON16 dataset. With MEGT as the baseline, the accuracies of the models D and E on both datasets are much degraded by using only one MFFM, because the class token cannot pass the learned information to its patch tokens. In addition, too much branch fusion does not improve performance, but introduces more parameters. This is because the cross-attention is a linear operation without any nonlinearity function, which results in overfitting of the model due to over-parameterization. Considering the performance and parameters of the model, we finally select 1, 2 and 2 as the values of Transformer layers  $L_{low}$ ,  $L_{high}$  and MFFM  $K$ .

## V. DISCUSSION

In this work, a novel MEGT network is proposed for cancer diagnosis on gigapixel WSIs. MEGT designs an effective MFFM to aggregate different resolution patches to produce stronger WSI-level representation, which is easy to implement without complicated patch processing. Experimental results on both TCGA-RCC and CAMELYON16 datasets indicate the effectiveness of our proposed MEGT.

Previous works on WSIs generally focused on single-resolution methods, which fail to capture multi-scale information of WSIs. Inspired by the diagnosis process of pathologists, some researchers have extended previous MIL algorithms to learn multi-scale representations from the WSI pyramid. From the aspect of multi-scale feature fusion, existing schemes are restricted to simple concatenation or multi-scale feature pyramid construction, and they do not pay enough attention to the intrinsic semantic gap among the patches with different resolutions. To this end, our proposed MEGT introduces a class token for each resolution as an agent to fuse multi-scale information of WSIs.

Since the class token uniformly converts different resolution features into slide-level representations for classification, the semantic gap in different resolution patches is alleviated. In addition, our framework avoids the complicated patch processing, such as building feature pyramids and heterogeneous graphs in the WSI. Therefore, the proposed MEGT is a simple yet effective framework for learning the multi-scale information of WSIs for a CAD model.

The spatial information of WSIs is also essential for cancer diagnosis. Different from previous graph-based MIL or Transformer-based MIL, the proposed EGT aims to learn the spatial-related features from graph data to enhance the performance of the Transformer. Here, each node in the graph data corresponds to a patch in the original WSI, and the edges are computed by the embedded features from these patches. Thus, the patch-based graph actually represents the spatial relationships among different regions in a WSI. In addition, EGT utilizes the attention scores of Transformer encoder for token pruning, which greatly reduces the computational complexity of graph construction and graph convolution. Therefore, EGT can efficiently learn the spatial information and the correlation between patches simultaneously to produce a superior feature representation.

Although the proposed MEGT achieves superior performance over the compared SOTA algorithms, it still has some room for improvement. For example, we will focus on the data-driven pretext task design in self-supervised learning to learn more effective multi-scale feature representation, which can alleviate the issue of a small sample size in histopathological images. On the other hand, MEGT is currently only suitable for dual resolution on WSIs due to the multi-scale feature fusion strategy of MFFM. Future studies can explore other efficient strategies, such as feature pyramids and hierarchical networks, to combine more resolution for feature fusion.

## VI. CONCLUSION

In this work, we proposed Multi-scale Efficient Graph-Transformer (MEGT), a dual-branch Transformer for aggregating image patches of different resolutions, to promote the accuracy of cancer diagnosis on WSIs. Particularly, an effective MFFM was developed to learn the multi-scale features and reduce the semantic gap in different resolution patches. Meanwhile, the EGT was specifically designed to improve the ability of the branches in MEGT for learning spatial information of WSIs. Experimental results on two public WSI datasets demonstrated the effectiveness of the proposed MEGT framework. It suggests that MEGT has the potential for WSI-based CAD in clinical practice.

## REFERENCES

- [1] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, "Artificial intelligence in digital pathology — New tools for diagnosis and precision oncology," *Nature Rev. Clin. Oncol.*, vol. 16, pp. 703–715, 2019.
- [2] M. Zarella et al., "A practical guide to whole slide imaging: A white paper from the Digital Pathology Association," *Arch. Pathol. Lab. Med.*, vol. 143, no. 2, pp. 222–234, 2019.

- [3] J. Shi, X. Zheng, J. Wu, B. Gong, Q. Zhang, and S. Ying, "Quaternion Grassmann average network for learning representation of histopathological image," *Pattern Recognit.*, vol. 89, pp. 67–76, 2019.
- [4] X. Zhu, J. Yao, F. Zhu, and J. Huang, "WSISA: Making survival prediction from whole slide histopathological images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6855–6863.
- [5] D. Di, S. Li, J. Zhang, and Y. Gao, "Ranking-based survival prediction on histopathological whole-slide images," in *Proc. Int. Conf. Med. Image Comput. Assist. Interv.*, 2020, pp. 428–438.
- [6] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.
- [7] G. Campanella et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Med.*, vol. 25, no. 8, pp. 1301–1309, 2019.
- [8] W. Shao, T. Wang, Z. Huang, Z. Han, J. Zhang, and K. Huang, "Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3739–3747, Dec. 2021.
- [9] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, "Neural image compression for gigapixel histopathology image analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 567–578, Feb. 2021.
- [10] H. Zhang et al., "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18780–18790.
- [11] J. Yao, X. Zhu, J. Jonnagaddala, N. J. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Med. Image Anal.*, vol. 65, 2020, Art. no. 101789.
- [12] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, 2021.
- [13] W. Hou et al., "H<sup>2</sup>-MIL: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 933–941.
- [14] N. Hashimoto et al., "Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3851–3860.
- [15] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14313–14323.
- [16] R. J. Chen et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16123–16134.
- [17] Z. Huang, H. Chai, R. Wang, H. Wang, Y. Yang, and H. Wu, "Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 561–570.
- [18] H. Li et al., "DT-MIL: Deformable transformer for multi-instance learning on histopathological image," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 206–216.
- [19] Z. Shao et al., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 2136–2147.
- [20] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Mach. Learn. Representations*, 2021, pp. 1–11.
- [21] Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. Ö. Arik, and T. Pfister, "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3417–3425.
- [22] R. Chen, Q. Fan, and R. Panda, "CrossVit: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 357–366.
- [23] Y. Liang, C. Ge, Z. Tong, Y. Song, J. Wang, and P. Xie, "Not all patches are what you need: Expediting vision transformers via token reorganizations," in *Proc. Int. Conf. Mach. Learn. Representations*, 2022, pp. 1–13.
- [24] S. Long, Z. Zhao, J. Pi, S. Wang, and J. Wang, "Beyond attentive tokens: Incorporating token importance and diversity for efficient vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10334–10343.
- [25] R. Li, J. Yao, X. Zhu, Y. Li, and J. Huang, "Graph CNN for survival analysis on whole slide pathological images," in *Proc. Int. Conf. Med. Image Comput. Assist. Interv.*, 2018, pp. 174–182.
- [26] P. Pati et al., "Hierarchical graph representations in digital pathology," *Med. Image Anal.*, vol. 75, 2021, Art. no. 102264.
- [27] Y. Zhou, S. Graham, N. A. Koohbanani, M. Shaban, P. Heng, and N. Rajpoot, "CGC-net: Cell graph convolutional network for grading of colorectal cancer histology images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 388–398.
- [28] N. A. Koohbanani, B. Unnikrishnan, S. A. Khurram, P. Krishnaswamy, and N. Rajpoot, "Self-path: Self-supervision for classification of pathology images with limited annotations," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2845–2856, Oct. 2021.
- [29] R. J. Chen et al., "Multimodal Co-attention transformer for survival prediction in gigapixel whole slide images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3995–4005.
- [30] P. Liu, B. Fu, F. Ye, R. Yang, B. Xu, and L. Ji, "DSCA: A dual-stream network with cross-attention on whole-slide image pyramids for cancer prognosis," *Expert Syst. Appl.*, vol. 227, 2023, Art. no. 120280.
- [31] R. J. Chen et al., "Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 339–349.
- [32] Z. Liu et al., "Swin Transformer: Hierarchical vision Transformer using shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
- [33] H. Fan et al., "Multiscale vision Transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6804–6815.
- [34] X. Wang et al., "Transformer-based unsupervised contrastive learning for histopathological image classification," *Med. Image Anal.*, vol. 81, 2022, Art. no. 102559.
- [35] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [36] Y. Xiong et al., "Nyströmformer: A Nyström-based algorithm for approximating self-attention," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 14138–14148.
- [37] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Mach. Learn. Representations*, 2017, pp. 1–10.
- [38] B. E. Bejnordi et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, pp. 2199–2210, 2017.
- [39] Y. Zheng et al., "A graph-transformer for whole slide image classification," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3003–3015, Nov. 2022.
- [40] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated multimodal units for information fusion," in *Proc. Int. Conf. Mach. Learn. Representations*, 2017, pp. 1–17.
- [41] L. van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.