TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification

Zhuchen Shao*,[1], Hao Bian*,[1], Yang Chen*,[1], Yifeng Wang[2], Jian Zhang[3], Xiangyang Ji[4]
Yongbing Zhang†,[2]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Harbin Institute of Technology (Shenzhen)
[3]School of Electronic and Computer Engineering, Peking University
[4]Department of Automation, Tsinghua University

CONTENTS

Whole Slide Image (WSI)

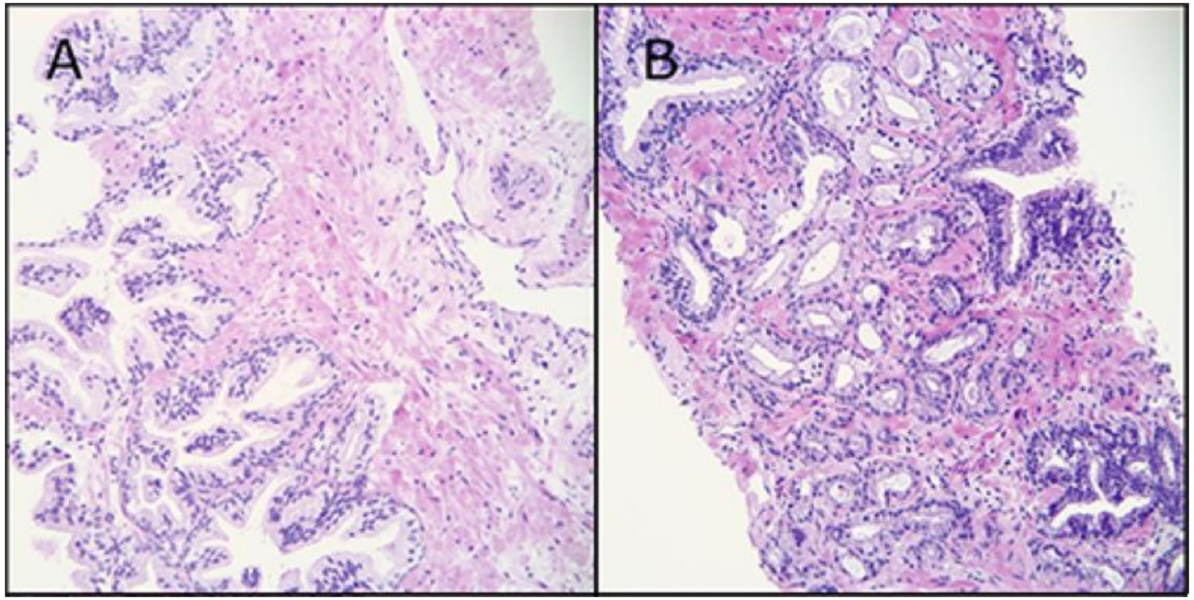

Fig. 1 Whole slide image

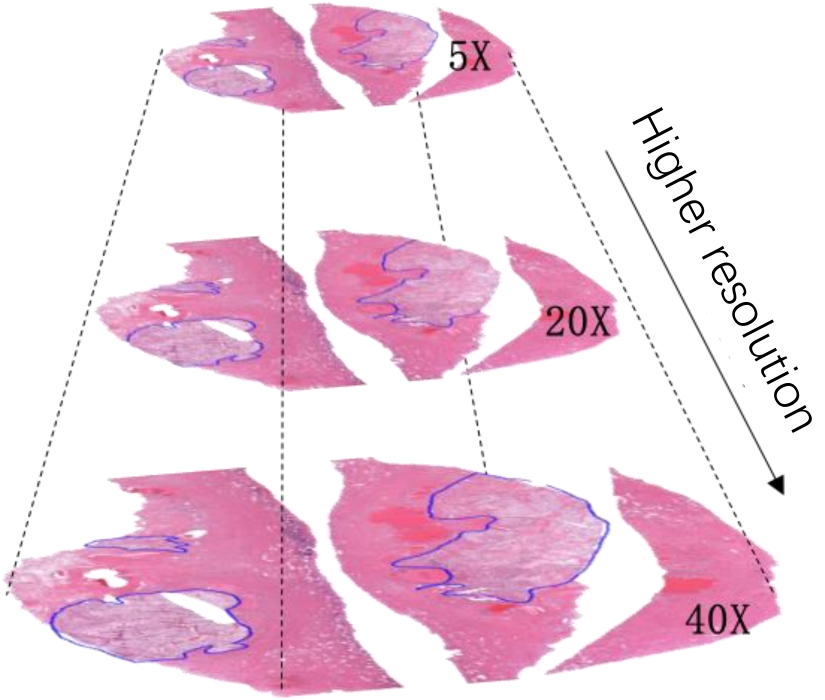

Fig. 2 WSI storage

- Huge size (~50000×50000 pixels at 20×)
- Lack of pixel-level annotations

- Multiple resolution images
- Tissue to cell level information

Multiple Instance Learning (MIL)

Description of MIL problem

- MIL is a weak supervision problem
- Each bag contains an unequal number of instances
- Bag level label is known, instance label is unknown

Traditional assumption in MIL problem

- All the instances in each bag are independent and identically distributed (i.i.d.)

Correlated Multiple Instance Learning

- **Difference:** Consider the correlation and spatial information between different instances in a bag.

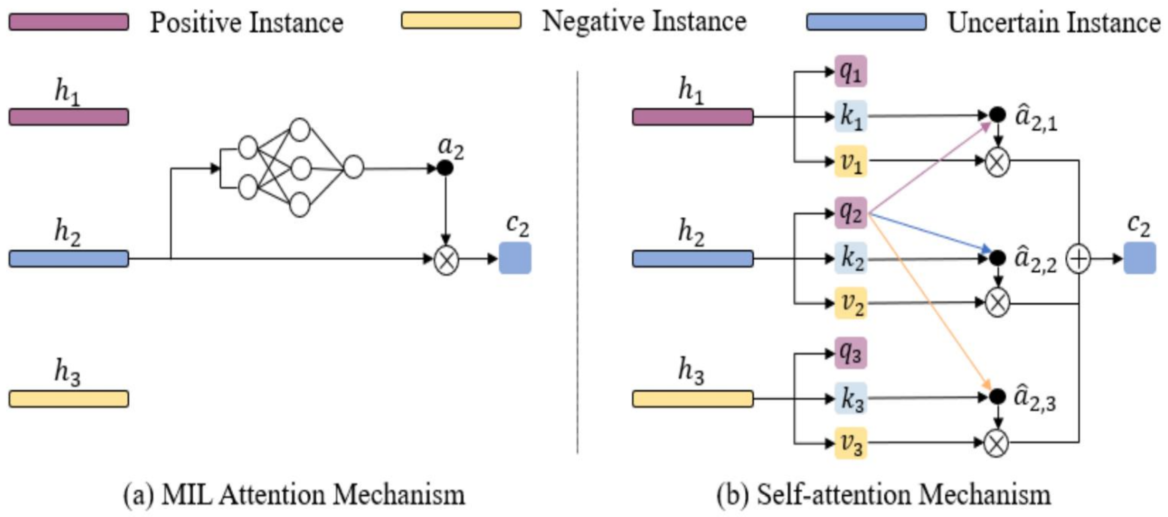

Fig. 3 Difference between MIL attention mechanism and self-attention mechanism

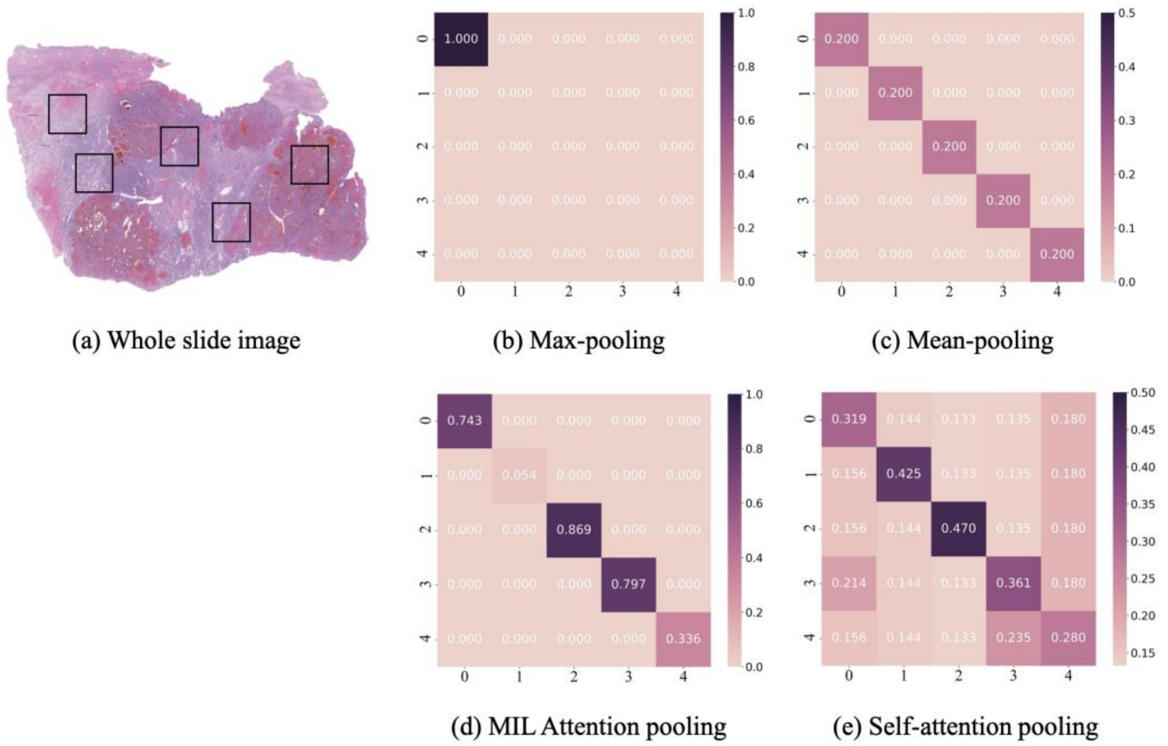

Fig.4 The difference between different Pooling Matrix

Method

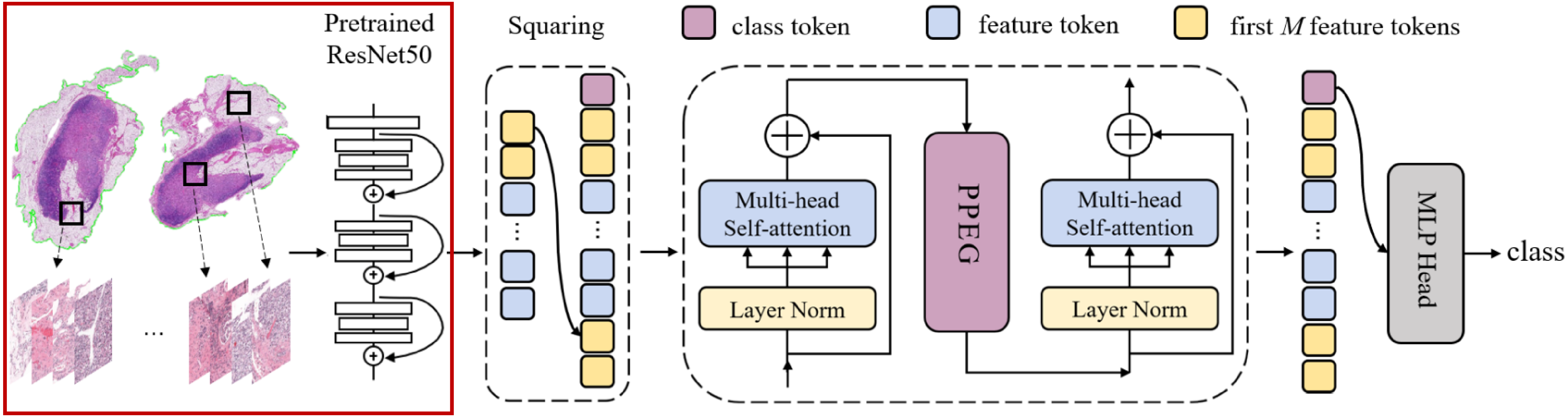

Fig. 5 Overview of our TransMIL

- **Preprocessing** :
  - 1) Each WSI is cropped into a series of 256×256 non-overlapping patches, where background region (saturation<15) is discarded.
  - 2) The feature of each patch is embedded in a 1024-dimensional vector by a ResNet50 model pre-trained on ImageNet.

Method

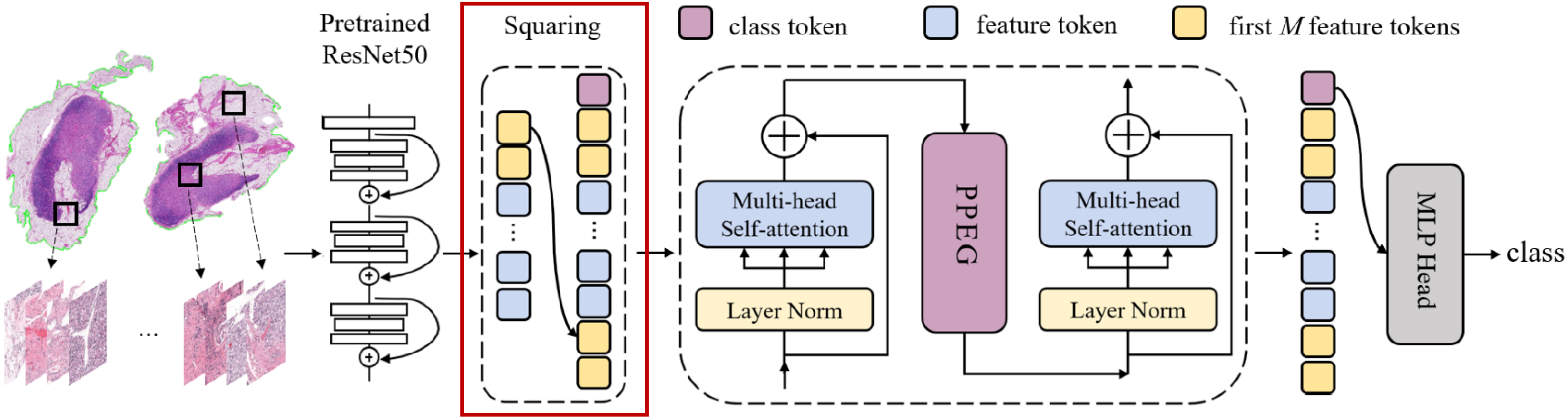

Fig. 5 Overview of our TransMIL

- **Squaring of sequence**:
  - Square the length of the sequence, and add the class token, then reduce the dimension of each feature embedding from 1024 to 512.

Method

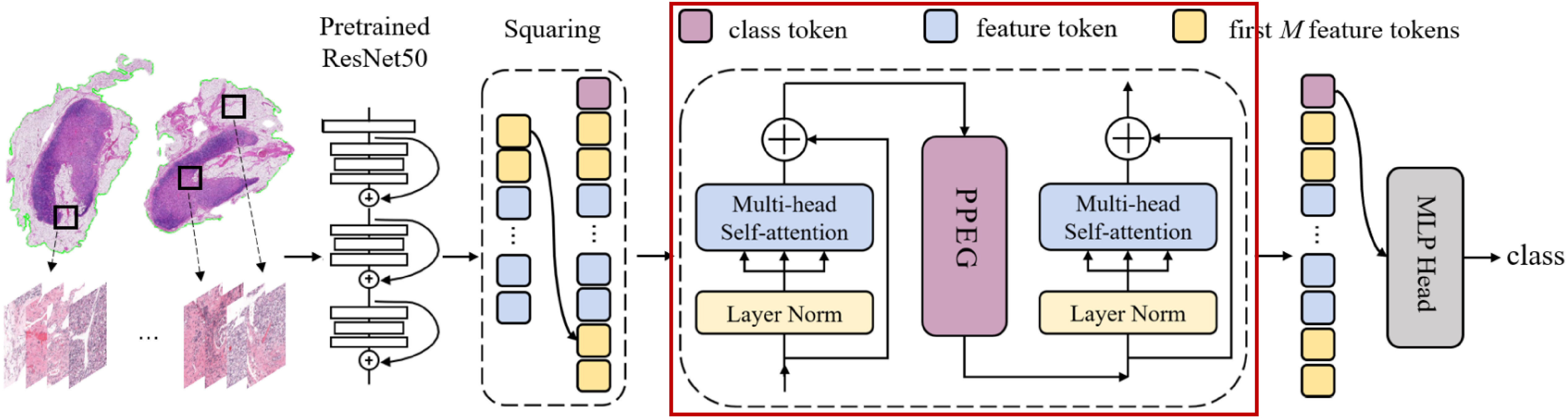

Fig. 5 Overview of our TransMIL

- 1） Correlation modelling of the sequence $\mathbf{H}_S^{\ell} \leftarrow \mathrm{MSA}(\mathbf{H}_S)$
- 2） Conditional position encoding and local information fusion $\mathbf{H}_S^{P} \leftarrow \mathrm{PPEG}(\mathbf{H}_S^{\ell})$
- 3） Deep feature aggregation $\mathbf{H}_S^{\ell+1} \leftarrow \mathrm{MSA}(\mathbf{H}_S^{P})$

Position encoding with PPEG

- **Background** : Zero padding can provide an absolute position information to convolution[1]

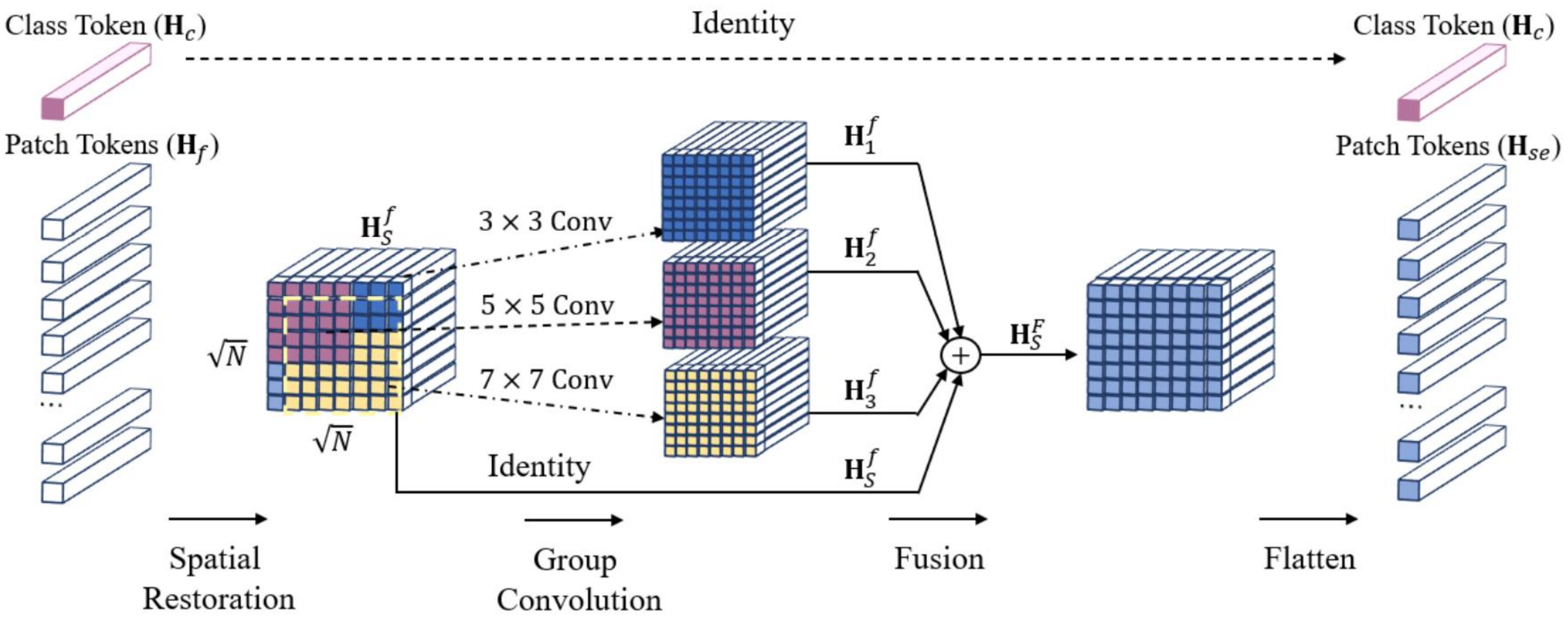

Fig.6 Pyramid Position Encoding Generator (PPEG)

[1] How much position information do convolutional neural networks encode? In International Conference on Learning Representations, 2020.

Method

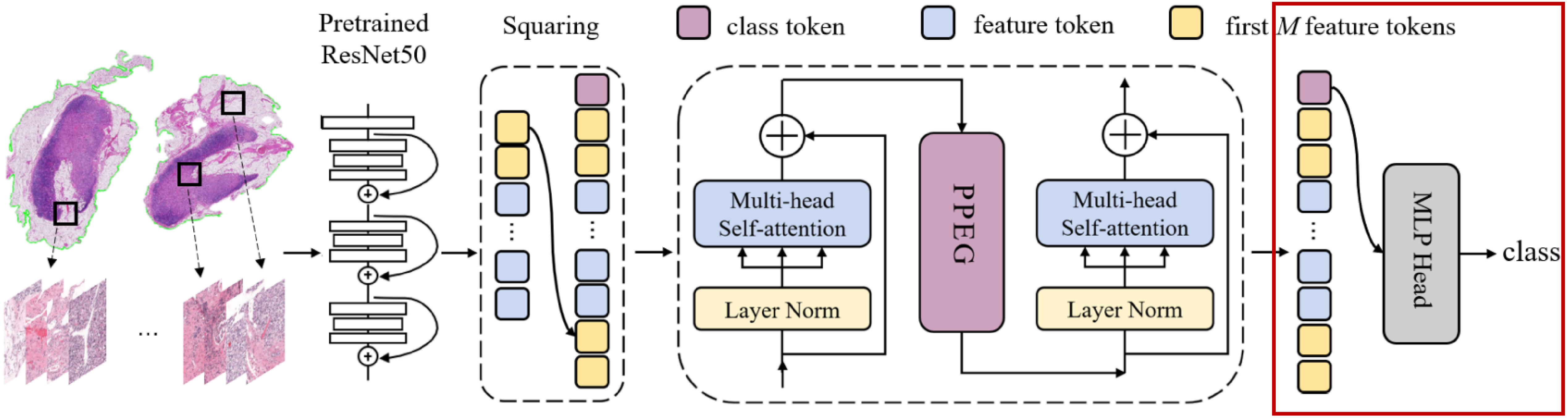

Fig. 5 Overview of our TransMIL

- **Mapping of $\mathbb{T} \to \mathcal{Y}$**
  - Use the class token to get the slide-level label of WSI
  - $\hat{Y} \leftarrow \mathrm{MLP}\left(\mathrm{LN}\left((\mathbf{H}_S^{\ell+1})^{(0)}\right)\right)$

Experiments and Results

Table 1: Results on CAMELYON16, TCGA-NSCLC and TCGA-RCC.

	CAMELYON16		TCGA-NSCLC		TCGA-RCC	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Mean-pooling	0.6389	0.4647	0.7282	0.8401	0.9054	0.9786
Max-pooling	0.8062	0.8569	0.8593	0.9463	0.9378	0.9879
ABMIL[1]	0.8682	0.8760	0.7719	0.8656	0.8934	0.9702
MIL-RNN[2]	0.8450	0.8880	0.8619	0.9107	\	\
DSMIL[3]	0.7985	0.8179	0.8058	0.8925	0.9294	0.9841
CLAM-SB[4]	0.8760	0.8809	0.8180	0.8818	0.8816	0.9723
CLAM-MB[4]	0.8372	0.8679	0.8422	0.9377	0.8966	0.9799
TransMIL	**0.8837**	**0.9309**	**0.8835**	**0.9603**	**0.9466**	**0.9882**

[1] Attention-based deep multiple instance learning. In International Conference on Machine Learning, 2018.
[2] Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine, 2019.
[3] Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.
[4] Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering, 2021.

THANK YOU