Kernel Attention Transformer for Histopathology Whole Slide Image Analysis and Assistant Cancer Diagnosis

Yushan Zheng, *Member, IEEE*, Jun Li, Jun Shi, *Member, IEEE*, Fengying Xie, Jianguo Huai, Ming Cao, and Zhiguo Jiang

***Abstract*— Transformer has been widely used in histopathology whole slide image analysis. However, the design of token-wise self-attention and positional embedding strategy in the common Transformer limits its effectiveness and efficiency when applied to gigapixel histopathology images. In this paper, we propose a novel kernel attention Transformer (KAT) for histopathology WSI analysis and assistant cancer diagnosis. The information transmission in KAT is achieved by cross-attention between the patch features and a set of kernels related to the spatial relationship of the patches on the whole slide images. Compared to the common Transformer structure, KAT can extract the hierarchical context information of the local regions of the WSI and provide diversified diagnosis information. Meanwhile, the kernel-based cross-attention paradigm significantly reduces the computational amount. The proposed method was evaluated on three large-scale datasets and was compared with 8 state-of-the-art methods. The experimental results have demonstrated the proposed KAT is effective and efficient in the task of histopathology WSI analysis and is superior to the state-of-the-art methods.**

***Index Terms*— WSI, transformer, cross-attention, gastric cancer, endometrial cancer.**

I. Introduction

HISTOPATHOLOGY whole slide image (WSI) analysis based on image processing and deep learning has proven effective in building computer-aided applications for cancer screening [1], [2], [3], tumor grading [4], [5], [6], prognosis analysis [7], [8], [9], gene mutant prediction [10], [11], etc.

Recently, Vision Transformer (ViT) [12], [13], the extensively studied model in natural scene image recognition and object detection, was introduced to this problem [14], [15], [16]. The WSI analysis can be achieved by taking the WSI local features as token inputs. Theoretically, the self-attention mechanism of Transformer enables it to detect the useful spatial and spectral relations of the local features on the images. The recent studies [17], [18], [19] have proven that Transformer-based models can further improve the WSI classification accuracy when compared to the previous methods. Nevertheless, the calculation flowchart of self-attention, the main operation in Transformer, occurs notable problems when applied to the WSIs containing gigapixels.

The positional embedding in ViT is designed for the natural sense image dataset, e.g., ImageNet [20], where all the images are in the same size, e.g. 224×224. The image patches are arranged in a consistent sequence for the ViT tokens. It ensures a token indicates consistent positional information throughout the training and inference stages. However, the size and shape of histopathology WSIs are not fixed, and the tissue region varies a lot in different WSI. It makes the patches from different WSIs but for a certain token are positional inconsistent or even positional conflicting, especially under the setting that only the foreground features are extracted to fill the tokens. This makes ViT ambiguous to describe the structural information for WSIs and thereby affects its performance for fine-grained WSI classification tasks that rely on tissue distribution. Secondly, the self-attention operation permits the equivalent conjunction of tokens in every Transformer block. However, the features for WSI diagnosis are expected to be extracted hierarchically from the local level to the global level. Moreover, the computational complexity of the self-attention operation is $\mathcal{O}(n^2)$ assuming that n denotes the number of tokens. The inference of the Transformer becomes rather inefficient when facing a WSI that generates thousands of tokens. These problems have limited the performance and efficiency of the Transformer-based method for WSI analysis.

In this paper, we propose a novel model named kernel attention Transformer (KAT) for whole slide image analysis

Manuscript received 18 January 2023; revised 21 March 2023; accepted 29 March 2023. Date of publication 5 April 2023; date of current version 31 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62171007, Grant 61901018, and Grant 61906058; in part by the Anhui Provincial Natural Science Foundation under Grant 1908085MF210; in part by the Project 111 under Grant B13003; and in part by the Fundamental Research Funds for the Central Universities of China under Grant JZ2022HGTB0285. *(Corresponding authors: Zhiguo Jiang; Jun Shi.)*

Yushan Zheng is with the Beijing Advanced Innovation Center for Biomedical Engineering, School of Engineering Medicine, Beihang University, Beijing 100191, China (e-mail: yszheng@buaa.edu.cn).

Jun Li, Fengying Xie, and Zhiguo Jiang are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 102206, China, and also with the Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing 100191, China (e-mail: junl20@buaa.edu.cn; xfy_73@buaa.edu.cn; jiangzg@buaa.edu.cn).

Jun Shi is with the School of Software, Hefei University of Technology, Hefei 230601, China (e-mail: juns@hfut.edu.cn).

Jianguo Huai and Ming Cao are with the Department of Pathology, The First People's Hospital of Wuhu, Wuhu 241000, China (e-mail: 478205286@qq.com; 854901420@qq.com).

Digital Object Identifier 10.1109/TMI.2023.3264781

and a flowchart for assistant cancer diagnosis. The framework is illustrated in Fig. 1. Compared to the common Transformer structure, the proposed KAT can describe hierarchical context information of the local regions of the WSI and therefore is more effective in histopathology WSI description and analysis. Meanwhile, the kernel-based cross-attention paradigm maintains a near-linear computational complexity to the size of the WSI. The proposed method was evaluated on a gastric dataset with 2040 WSIs, an endometrial dataset with 2560 WSIs and a lung dataset with 3064 WSIs, and was compared with 8 state-of-the-art methods [12], [14], [18], [21], [22], [23]. The experimental results have demonstrated the proposed KAT is effective and efficient in the task of histopathology WSI classification and is superior to the state-of-the-art methods.

The contribution of the paper can be summarized in three aspects.

(1) A novel model named kernel attention Transformer (KAT) is proposed. Compared to ViT [12], the token-wise self-attention is replaced by cross-attention between the tokens and a set of kernels to achieve information transmission. The experiments have demonstrated that the kernel-based cross-attention contributes to a competitive performance for WSI classification and meanwhile significantly reduces the computational complexity of ViT in both the training and the inference stages.

(2) An anchor-based token masking approach is designed for KAT to describe the structural information of histopathology whole slide image. Specifically, a set of anchors are defined based on the spatial positions of the patches and then bound to the kernels in KAT. Then, hierarchical anchor-based soft masks are created to guide the cross-attention to perceive multi-scale features of WSI. It helps the KAT to learn hierarchical representations from the local to the global scale of the WSI, thus delivering better WSI classification performance.

(3) A kernel contrastive representation learning (KCL) strategy is proposed to improve the effectiveness and interpretability of KAT for WSI analysis. Based on KCL, the kernel tokens can extract more discriminative regional representations, which are promising to build informative assistant WSI diagnosis systems.

The remainder of this paper is organized as follows. Section II reviews the related works. Section III introduces the methodology of the proposed method. The experiment and discussion are presented in Section IV and Section V, respectively.

II. Related Works

The gigapixel property of histopathology WSI makes it a challenging task to build end-to-end classification models under the limitation of the current computer hardware [24]. Therefore, the present methods generally separate the task into several stages.

One of the popular methods for WSI analysis is based on patch classification [25], [26], [27]. For example, Wang et al. [27] trained a patch-level CNN based on the annotation of the WSI to select the discriminative patches and then built a recalibrated multi-instance deep learning (RMDL) model based on these patches to achieve the WSI classification. These methods generally rely on the fine-grained annotations of pathologists. It limits the application to the clinical situation where the WSIs lack annotations.

To achieve the annotation-free WSI classification, Divide et al. [28] proposed dividing the WSI into patches and compressing the WSIs into feature cubes by training a patch-level convolution neural network (CNN) [29] and then using a second CNN to predict the label of the WSI from these feature cubes. Recently, Xiang et al. [30] proposed a dual stream convolution neural network that took both the WSI thumbnail images and transformed patch embeddings as input. The multi-level modeling strategy delivers greater performance than the previous method. These methods don't rely on annotations and have proven effective for WSI classification. However, the restriction of the input size of the second CNN makes the padded feature cubes usually involve a large area of background data, which leads to a substantial waste of computation in both the training and application stage.

More typically, multiple instance learning (MIL) is introduced into this domain and has become of the most popular techniques for whole slide image classification [21], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]. Ilse et al. [32] proposed an attention-based MIL model, where the CNN for patch feature extraction and the attention module for feature aggregation can be trained end-to-end. Then, Campanella et al. [34] proposed training a patch feature extractor based on WSI labels, i.e., the pseudo levels, within the tok-K confident patches from the WSIs and then trained a recurrent neural network (RNN) [41], [42] to achieve the WSI classification based on the tok-K patches. The performance of these methods depends on the CNN trained within the dataset. And, the top-K strategy is hard to be applied to the fine-grained WSI subtyping tasks. More recently, the embedding-based MIL methods are widely studied [21], [37], [40]. Typically, CLAM [21] utilized CNN pre-trained on the ImageNet dataset [20] as the patch feature extractor and achieved multi-type WSI classification by the designed multi-class attention module. The major weakness of the MIL-based methods is that the local regions are generally regarded as individual instances. The structural and spatial allocation information of these instances is rarely considered, which causes significant precision degradation when facing the WSI classification tasks relying on the tissue structural patterns.

To describe the relationship of the sub-regions, graph-based modeling methods are proposed for WSI analysis [18], [23], [36], [43], [44], [45]. In [43] and [44], graphs are constructed for WSIs based on the patch similarities in the feature spaces. The studies [18], [23] built graphs based on the adjacency of the patches on the WSI, for which the spatial relationship of the patches is properly described. Based on these graphs, the WSI analysis can be realized by graph neural networks (GCNs).

Most recently, Transformer structures [12], [22], [46] were introduced for WSI analysis [14], [19], [47], [48], [49], [50]. Transformer enables more extensive communication of the patches compared to RNN, MIL, and GCN-based models. This advantage decides it is more promising for fine-grained WSI analysis. However, the self-attention operation makes

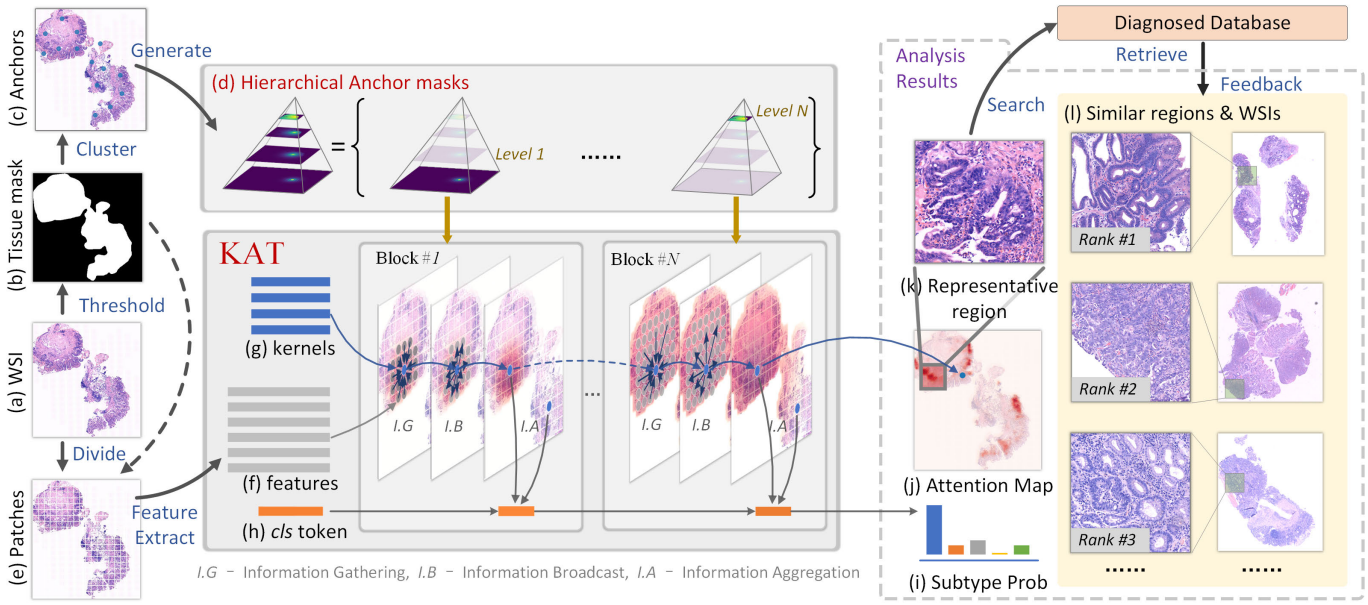


Fig. 1. The proposed histopathology WSI analysis framework, where (a) is the original WSI, (b-d) illustrates the process of anchor definition and anchor-related mask generation, (e-f) is the feature extraction flowchart, (g) represents a set of trainable embedding that serves as the kernels, (h) is a trainable vector used for information aggregation and WSI classification, and (i-l) shows the outputs of the framework that consists of the sub-typing probability, the hot map for the region of interest indication, and the most attentive region with its similar regions/cases from the database. The kernel attention Transformer (KAT) is the core of the framework, which is detailed in section III and Fig. 2.

Transformer expensive in both computational amount and memory cost, which is a problem that the present methods are trying to solve. For example, Huang et al. [47] utilized random sampling to reduce the number of patches fed into the Transformer. Li et al. [19] applied the deformable Transformer [51] to reduce the computational complexity of the self-attention operation. Typically, TransMIL [14] equipped the linear-approximated self-attention model, Nyströmformer [22], to maintain a low memory cost. Furthermore, TransMIL enhanced the spatial description ability with a pyramid position encoding generator (PPEG) module. But, its design still assumes that tissue is a fixed-size square region, for which the positional inconsistent and conflicting problems were not yet properly tackled.

This paper tries to solve the inherent problem of ViT in the aspect of structural information preservation and computational efficiency when applied to WSI analysis. A part of this work has been presented in the conference paper [52].

III. Method

A. Overview

The proposed histopathology WSI analysis framework is shown in Fig. 1. The foreground regions of a WSI are first divided into patches. Meanwhile, a set of anchors and anchor-related soft masks are generated based on the shape and size of the tissue, as shown in Fig. 1c-d. Then, the patch features along with a set of trainable kernels that bound with the anchors are fed into the proposed KAT model for analysis. The information flowing in KAT is achieved by cross-attention operation between the kernels and patches, which includes the procedures of information gathering (I.B), information broadcast (I.B), and information aggregation (I.A). The KAT model outputs the class prediction and the attentive regions for the WSI, as shown in Fig. 1i-j. Furthermore, similar WSI regions from the diagnosed database can be retrieved for diagnosis reference, as shown in Fig. 1l.

B. Pre-Processing and Data Preparation

1) Local Feature Extraction: The units of WSI analysis in our framework are image patches that are obtained following the window sliding paradigm, as shown in Fig. 1e. The patch features are extracted using a convolution neural network (CNN). The blank regions of the WSI are less informative for diagnosis and therefore were removed beforehand by a threshold on the intensity [53]. The features within the foreground of the tissue mask were rearranged to be a feature matrix (as shown in Fig. 1f) which is formulated as $\mathbf{F} \in \mathbb{R}^{n_p \times d_f}$, where n_p denotes the number of the foreground patches on the WSI, d_f denotes the dimension of the feature, and $\mathbf{f}_i$ is the i-th row of the $\mathbf{F}$ that represents the feature of the i-th patch.

2) Anchor Definition: In our method, the structure information of the WSI is handled by a set of anchors. The anchors are designed to be uniformly located in the foreground of the WSI, i.e. the tissue region. This is achieved by clustering the spatial coordinates of the foreground patches. Specifically, we define $p(\mathbf{f}_i) = (m_i, n_i)^{\mathrm{T}}$ to represent the patch-wise position of $\mathbf{f}_i$ on the WSI, and collect all the feature positions within the WSI as set $\{p(\mathbf{f}_i)\}_{i=1}^{n_p}$. These positions are clustered into K centers based on the K-means algorithm. Then, the nearest position to each center is recognized as an anchor position, which is represented as $\mathbf{c}_k = (m_k, n_k)^{\mathrm{T}}, k = 1, 2 \ldots, K$. To make the proposed method scalable to the size of WSI, we assign an adaptive number of anchors for each WSI by defining $K = [n_p/\bar{n}_k]$, where $\bar{n}_k$ denotes the average number

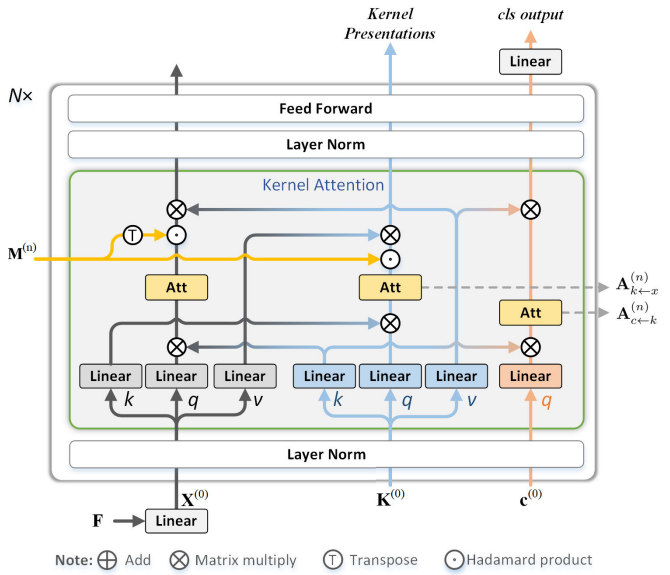


Fig. 2. The structure of kernel attention network (KAT), where Att denotes a scaling operation followed by a softmax operation, F, $K^{(0)}$, and $c^{(0)}$ denotes the input of the feature tokens, kernel tokens, and classification token, respectively, $M^{(n)}$ is the anchor mask for the n-th stage, and $A^{(n)}_{k\leftarrow x}$ and $A^{(n)}_{c\leftarrow k}$ are the attention weights used for generating visualization outputs.

of patches processed per kernel. $\bar{n}_k$ is a hyper-parameter of KAT, which will be discussed in the experimental sections.

3) Soft-Mask Generation: Afterward, we propose using the spatial distances between the anchors and the patches to describe the structure of the tissue. Specifically, a soft mask is generated for each anchor by calculating the weight between the k-th anchor and the i-th patch with the equation

$$m_{ki}(\delta) = \exp(-\| p(\mathbf{f}_i) - \mathbf{c}_k \|_2^2 / 2\delta^2), \tag{1}$$

where δ controls the scale of the Gaussian-like mask. By computing the weights between all the features and anchors, we obtain the mask matrix $\mathbf{M}(\delta) \in (0, 1)^{K\times n_p}$, where each row defines the soft mask for an anchor, and each column records the weights of a patch to the anchors. Moreover, we define multi-scale masks by adjusting δ. Finally, the hierarchical masks are represented by the collection

$$\mathbb{M} = \{\mathbf{M}(\delta_n)\}_{n=1}^{N}, \tag{2}$$

with N representing the number of scales. Here, we empirically set $\delta_1 = \sqrt{\bar{n}_k}$ to make the intersection of the ($-\delta_1$, $+\delta_1$) area of all the anchors cover all the patches, and set $\delta_n = \sqrt{\bar{n}_k \cdot 2^{n-1}}$ to make the receptive field of the anchor gradually expand with the deepening of the network. The visualization of the anchors and soft masks can be found in Fig. 8.

C. Kernel Attention Transformer (KAT)

1) Network Structure: The input of KAT for a WSI includes the feature matrix $\mathbf{F}$ and the anchor masks $\mathbb{M}$. As shown in Fig. 2, the outline structure of KAT follows ViT [12], which is constructed by stacking N repeated blocks. Each block is composed of modules in the sequence of layer normalization (LN), kernel attention (KA), layer normalization, and feed-forward (FF) module. The feed-forward module is a two-layer full-connected neural network that has the same structure as that in ViT. The basic inputs of the n-th KA module include the patch token representations $\mathbf{X}^{(n)} \in \mathbb{R}^{n_p\times d_e}$ and the classification token $\mathbf{c}^{(n)} \in \mathbb{R}^{d_e}$ after layer normalization. Besides, A set of trainable tokens $\mathbf{K}^{(n)} \in \mathbb{R}^{K\times d_e}$ (as shown in Fig. 1h) named *kernels* are defined and bound with the anchors and their soft-masks $\mathbf{M}^{(n)}$. The inference process of a block can be briefly formulated as

$$\begin{aligned}
&\mathbf{X}'^{(n+1)}, \mathbf{K}'^{(n+1)}, \mathbf{c}'^{(n+1)} \\
&\quad = KA\left(LN(\mathbf{X}^{(n)}, \mathbf{K}^{(n)}, \mathbf{c}^{(n)})\right), \\
&\mathbf{X}^{(n+1)}, \mathbf{K}^{(n+1)}, \mathbf{c}^{(n+1)} \\
&\quad = FF\left(LN(\mathbf{X}'^{(n+1)}, \mathbf{K}'^{(n+1)}, \mathbf{c}'^{(n+1)})\right)
\end{aligned} \tag{3}$$

Additionally, the KA module and FF module are with residual connections. The KA module is the major difference between KAT and the common Transformer, which is detailed in this section.

2) Kernel Attention (KA) Module: Instead of doing self-attention among the patch tokens, we propose performing cross-attention between the kernels and the patch tokens. Namely, the information transmission in the KA module is achieved by a bi-direction message passing flow. One direction is the *information gathering (I.G)* flow, which is defined by the equation

$$\begin{aligned}
\mathbf{A}^{(n)}_{k\leftarrow x} &= \sigma\left(\bar{\mathbf{K}}^{(n)}\mathbf{W}^{(n)}_q \cdot (\bar{\mathbf{X}}^{(n)}\mathbf{W}^{(n)}_k)^{\mathrm{T}}/\tau\right), \\
\mathbf{K}'^{(n+1)} &= (\mathbf{A}^{(n)}_{k\leftarrow x} \odot \mathbf{M}^{(n)}) \cdot \bar{\mathbf{X}}^{(n)}\mathbf{W}^{(n)}_v,
\end{aligned} \tag{4}$$

where $\mathbf{A}^{(n)}_{k\leftarrow x} \in \mathbb{R}^{K\times n_p}$ is the weighting matrix that indicates the attention weights of each kernel to all the patches, $\bar{\mathbf{K}}^{(n)}$ and $\bar{\mathbf{X}}^{(n)}$ are the normalized tensors of $\mathbf{K}^{(n)}$ and $\mathbf{X}^{(n)}$ after the LN layer, $\mathbf{W}^{(n)}_k$, $\mathbf{W}^{(n)}_q$, $\mathbf{W}^{(n)}_v \in \mathbb{R}^{d_e\times d_h}$ denote the trainable weights for the linear projections from the embedding dimension d_e to the head dimension d_h for *Key*, *Query*, and *Value* entries, respectively, σ denotes the row-wise softmax function, $\tau = \sqrt{d_h}$ is the scaling factor, and $\mathbf{K}'^{(n+1)}$ is the output of *I.G* flow. Particularly, $\mathbf{M}^{(n)} = \mathbf{M}(\delta_n) \in \mathbb{M}$ denotes the anchor-related masks defined by Eq. (2). Based on the Hadamard product of $\mathbf{M}^{(n)}$ and $\mathbf{A}^{(n)}_{k\leftarrow x}$, the attention of each kernel is restricted to its nearby region.

Another direction is the *information broadcast (I.B)* flow. The calculation of *I.B* is symmetric with *I.G* that is defined by the equation

$$\begin{aligned}
\mathbf{A}^{(n)}_{x\leftarrow k} &= \sigma\left(\bar{\mathbf{X}}^{(n)}\mathbf{W}^{(n)}_q \cdot (\bar{\mathbf{K}}^{(n)}\mathbf{W}^{(n)}_k)^{\mathrm{T}}/\tau\right), \\
\mathbf{X}'^{(n+1)} &= (\mathbf{A}^{(n)}_{x\leftarrow k} \odot \mathbf{M}^{\mathrm{T}(n)}) \cdot \bar{\mathbf{K}}^{(n)}\mathbf{W}^{(n)}_v,
\end{aligned} \tag{5}$$

where $\mathbf{A}^{(n)}_{x\leftarrow k} \in \mathbb{R}^{n_p\times K}$ is a weighting matrix that indicates the weights of each patch receiving the information of all the kernels and $\mathbf{X}'^{(n+1)}$ is the output of *I.B* flow.

Through *I.G* flow, the local information described by the patch representations is reported to their nearby kernels for information gathering. Then, through the *I.B* flow, the regional information summarized by the kernels is broadcast back to

the patches. Based on the bi-directional message passing flow, communication among the patch representations of the WSI can be accomplished. The soft masks defined by $\mathbf{M}^{(n)}$ for each stage of the KA module expand as n enlarges, and thereby the commutation range between the kernels and the patches gradually increases along with the inference process. It allows the kernels to get hierarchical structural information from the micro to macro view of tumor tissue. This will benefit the fine-grained recognition of the tissues, especially in the sub-typing tasks depending on the spatial distribution of the tissue.

In each KA module, a classification token is used for summing up the information from all the kernels, which is named as *information aggregation (I.A)* flow and formulated as

$$\begin{aligned}\mathbf{A}_{c\leftarrow k}^{(n)} &= \sigma\left(\bar{\mathbf{c}}^{(n)}\mathbf{W}_q^{(n)}\cdot(\bar{\mathbf{K}}^{(n)}\mathbf{W}_k^{(n)})^{\mathrm{T}}/\tau\right),\\ \mathbf{c}'^{(n+1)} &= \mathbf{A}_{c\leftarrow k}^{(n)}\cdot\bar{\mathbf{K}}^{(n)}\mathbf{W}_v^{(n)},\end{aligned} \tag{6}$$

where $\mathbf{c}'^{(n+1)}$ is the output of the *I.A* flow.

Here, we provide a computational complexity analysis for the KA module. According to Eq. (4), the computation for $\mathbf{A}_{k\leftarrow x}^{(n)}$ costs $\mathcal{O}(Kd_ed_h + n_pd_ed_h + Kn_pd_h)$ and the computation for $\mathbf{K}'^{(n+1)}$ costs $\mathcal{O}(Kn_p + n_pd_ed_h + Kn_pd_h)$. Then, the total computation for Eq. (4) costs $\mathcal{O}(Kd_ed_h + n_pd_ed_h + Kn_pd_h + Kn_p)$. Because d_e and d_h are constant dimension parameters for the network, the computation complexity evolves to $\mathcal{O}(K + n_p + Kn_p) = \mathcal{O}(Kn_p)$. Similarly, we can also derive that the computational complexity for Eq. (5) is $\mathcal{O}(Kn_p)$ and is $\mathcal{O}(K)$ for Eq. (6). Therefore, the computational complexity for KA is $\mathcal{O}(Kn_p + K) = \mathcal{O}(Kn_p)$. If we set the number of kernels K as constant, the computational complexity for KA is $\mathcal{O}(n_p)$, i.e., scales linearly w.r.t. the number of patches for a WSI.

Finally, we extended it to the multi-head KA module following the paradigm of Transformer [12] to further improve the performance of the KA module.

D. Objective and Optimization

1) WSI Classification Objective: A fully connected layer is built on the output of the *cls* token for WSI classification, which is formulated as $\mathbf{z}_c = \mathcal{F}_{fc}(\mathbf{c}^{(N)})$, where $\mathbf{z}_c \in \mathbb{R}^{n_c}$ is the output of the classification token with n_c denoting the number of neurons of the layer, i.e. the type number of the WSIs. The objective function for WSI classification is the cross-entropy between $\mathbf{z}_c$ and the label of the WSI, which is formulated as

$$L_{ce} = -\mathbf{y}^{\mathrm{T}}\log(\sigma(\mathbf{z}_c)), \tag{7}$$

where $\mathbf{y}$ denotes the one-hot label of the WSI.

2) Kernel Contrastive Representation Learning (KCL): We propose building a contrastive representation learning module for the the kernels, to enhance the discrimination of the kernel representations. As shown in Fig. 3, two augmentations of the feature matrix $\mathbf{F}$ for a WSI are generated based on patch-wise Monte Carlo dropout with a probability of p_{drop} and are denoted as $\mathbf{F}'_1$ and $\mathbf{F}'_2$. Correspondingly, a contrastive representation learning structure is built to the kernel outputs

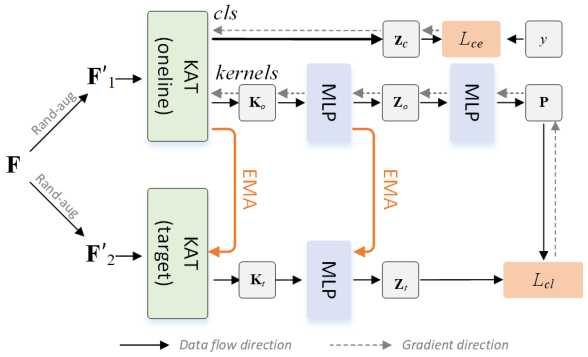


Fig. 3. Data flow for training KAT, based on the composite loss function for WSI classification and contrastive kernel representation learning. For the details please refer to section III-D.

of KAT referring to BYOL [54]. $\mathbf{F}'_1$ and $\mathbf{F}'_2$ are fed into the KAT of the online branch and the target branch, respectively, based on which the kernel representations of $\mathbf{K}_o^{(n)}$ and $\mathbf{K}_t^{(n)}$ are obtained. The $\mathbf{K}_o^{(n)}$ for the online branch is sequentially fed into a projector $\mathcal{F}_o^{proj}$ and a predictor $\mathcal{F}_o^{pred}$ that are composed of multilayer perceptrons (MLPs) following [54]. This process is formulated as

$$\begin{aligned}\mathbf{Z}_o^{(n)} &= \mathcal{F}_o^{proj}(\mathbf{K}_o^{(n)}),\\ \mathbf{P}^{(n)} &= \mathcal{F}_o^{pred}(\mathbf{Z}_o^{(n)}),\end{aligned} \tag{8}$$

and $\mathbf{K}_t^{(n)}$ for the target branch is fed into a projector $\mathcal{F}_t^{proj}$ that has the same structure with $\mathcal{F}_o^{proj}$ to obtain

$$\mathbf{Z}_t^{(n)} = \mathcal{F}_t^{proj}(\mathbf{K}_t^{(n)}), \tag{9}$$

where $\mathbf{P}^{(n)}, \mathbf{Z}_t^{(n)} \in \mathbb{R}^{K\times d_p}$ with d_p denote the dimension of the projector and predictor output, respectively. Note that the number of kernels is determined based on the original WSI patches and kept fixed during the Monte Carlo dropout. Therefore, the anchors related to the rows in $\mathbf{P}^{(n)}$ and $\mathbf{Z}_t^{(n)}$ maintain aligned. Based on this, we define the contrastive objective function as

$$L_{cl} = \sum_{n=1}^{N}\|\bar{\mathbf{P}}^{(n)} - \bar{\mathbf{Z}}_t^{(n)}\|_2^2, \tag{10}$$

where $\bar{\mathbf{P}}^{(n)}$ and $\bar{\mathbf{Z}}_t^{(n)}$ represent the row-wise-l2-normalized $\mathbf{P}^{(n)}$ and $\mathbf{Z}_t^{(n)}$, respectively.

3) Optimization: Finally, the entire KAT is trained end-to-end based on the composite objective function

$$L = L_{ce} + \lambda L_{cl} \tag{11}$$

with λ controls the composite weight. The parameters of the modules in the online target are optimized by the gradient descent algorithm, and the modules in the target branch are updated by the exponential moving average (EMA) mechanism [54]. AdamW optimizer is employed to handle the mini-batch-based back-propagation procedure. The inputs of the kernel tokens and the classification token, i.e. $\mathbf{K}^{(0)}$ and $\mathbf{c}^{(0)}$, are randomly initialed and kept trainable in the training stage. To ensure all the kernels have consistent action for the same allocation of nearby features, we make all the kernels share the same set of trainable parameters.

E. WSI Analysis With KAT

As shown in Fig. 1i, the basic output of the KAT is $\sigma(\mathbf{z}_c)$, i.e., the probability of the WSI for each type. Moreover, we can find which kernel contributes the most to the decision by checking $\mathbf{A}^{(n)}_{c\leftarrow k}$, and correspondingly resolve the attention map of the kernels by projecting the values $\mathbf{A}^{(n)}_{k\leftarrow x}$ on the WSI. Afterward, the representative region, i.e. the most attentive regions on the attention map can be highlighted and fed back to the doctors.

IV. Experiment and Result

A. Experimental Settings

The proposed method was evaluated on three large-scale WSI datasets, which are briefly introduced below.

- *Gastric-2K* contains 2040 WSIs of gastric biopsy histopathology collected from 2040 patients by No.1 people's hospital of Wuhu, China. These WSIs are categorized into 6 subtypes of gastric pathology, including Low-grade intraepithelial neoplasia (LGIN), High-grade intraepithelial neoplasia (HGIN), and Adenocarcinoma (A.), Mucinous adenocarcinoma (MA), Signet-ring cell carcinoma (SRCC), and non-tumor tissue (Normal).
- *Endometrial-2K* contains 2650 WSIs of endometrium histopathology collected from 2650 patients by Tianjin Fifth Central Hospital of China. These WSIs are categorized into 5 subtypes of endometrial pathology, including Well/Moderately/Low-differentiated endometrioid adenocarcinoma (WDEA/MDEA/LDEA), Serous endometrial intraepithelial carcinoma (SEIC), and cancer-free tissue (Normal).
- *TCGA-Lung-3K* contains 3064 WSIs of lung histopathology collected from the cancer genome atlas (TCGA) program of NCI. These WSIs are categorized into 3 subtypes including lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and cancer-free tissue (Normal). This dataset is publicly available.[1]

In each dataset, the WSIs were randomly separated into train and test subsets by the proportion of 7:3. The distribution of WSIs in the three datasets is given in Table I. Five-fold cross-validation was conducted within the training set where the validation part in each fold was used to perform early stops and hyper-parameter selection. Then, the mean and standard deviation of the results on the test set obtained by these early stopped models were used for ablation study and quantitative comparison with other methods.

The structure of CNN was Resnet-50 [55] because of its comprehensive performance. An increasing number of methods [16], [37], [47], [56], [57], [58] introduced the contrastive representation learning to train the CNN and have achieved a significantly better WSI classification performance than the previous methods. Following these methods, we trained the Resnet-50 within the training data using the contrastive representation learning framework proposed in BYOL [54].

[1]The list of these WSIs is available at https://github.com/Zhengyushan/kat/tree/main/dataset/tcga_lung

TABLE I
The WSI Distribution of the Experimental Datasets

Gastric-2K	Normal	LGIN	HGIN	A.	MA	SRCC
Train	563	354	120	259	74	88
Test	227	148	58	117	14	18
Endometrial-2K	Normal	WDEA	MDEA	LDEA	SEIC	
Train	413	557	590	187	111	
Test	162	261	234	91	44	
TCGA-Lung-3K	Normal	LUAD	LUSC			
Train	395	863	886			
Test	158	394	368			

The window sliding and CNN feature extraction were performed on the WSIs under 20× lenses (the resolution is 0.48μm/pixel). The window size, as well as the image patch size, was set 224×224 and the feature dimension $d_f = 2048$ to fit the Resnet50 structure.

The average accuracy, macro-F1 score, weighted-F1 score, the macro area under the receiver operating characteristic curve (macro-AUC), and weighted-AUC are calculated for evaluating the sub-typing tasks.[2]

The proposed method was implemented in Python with torch, and run on a computer cluster with 10 Xeon 2.66GHz CPUs and 10 GPUs of Nvidia Geforce 2080Ti. For more details please refer to the source code at https://github.com/zhengyushan/kat.

B. Hyper-Parameter Verification

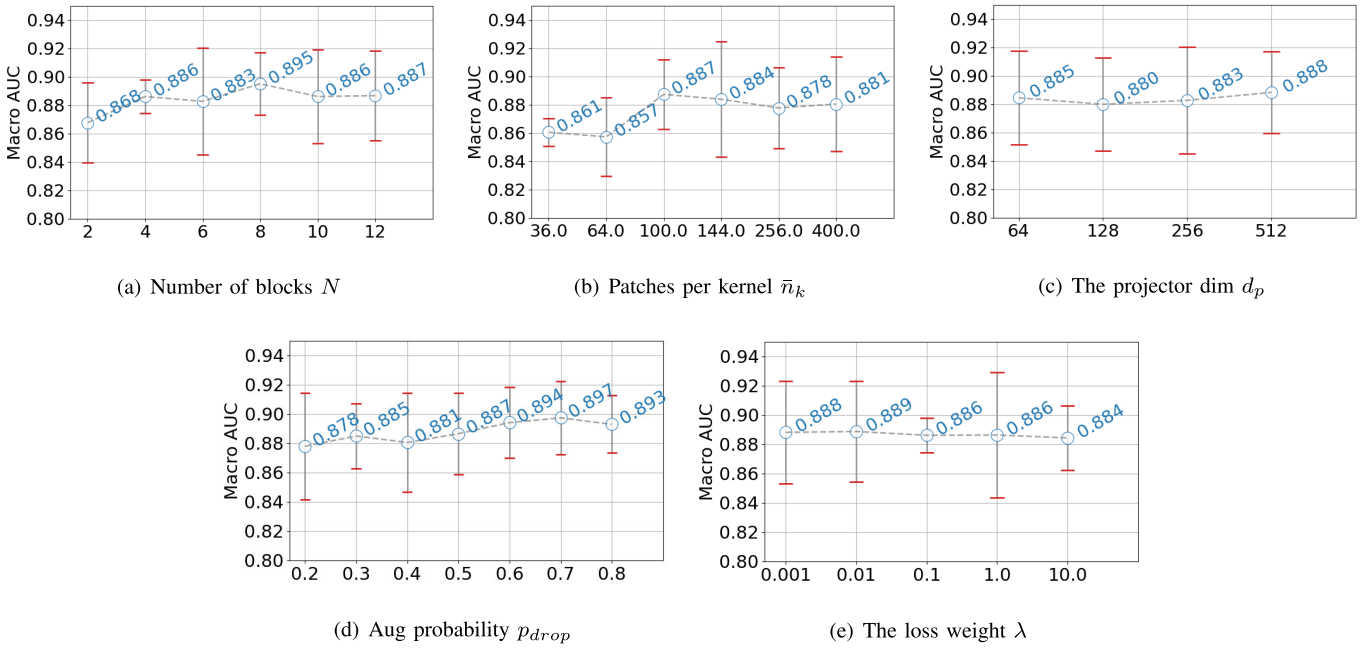
We first conducted experiments to verify the design of the proposed KAT model. There are five hyper-parameters, $(N, \bar{n}_k, d_p, p_{drop}, \lambda)$ that decide the capacity of KAT. We discuss the effect of these hyper-parameters on the *Gastric-2K* dataset. The mean and standard deviation of macro-AUCs for the validation data in the cross-validation as functions of these hyper-parameters are plotted in Fig. 4. Note that the other hyper-parameters were set fixed when one hyper-parameter was being tuned.

1) The Number of Stacked Blocks: N is positively correlated with the computational amount of KAT and also controls the scale upper bound of the anchor-related soft masks. N was tuned in the range of [2] and [12] with a step of 2. The curve in Fig. 4a shows that the classification performance generally increases as N enlarges and then keeps stable when $N > 4$. Therefore, we set $N = 8$ for its best classification performance and relatively low computational amount.

2) Average Patch Number Per Kernel: $\bar{n}_k$ decides the basic range of an anchor gathering and broadcasting information. A larger $\bar{n}_k$ helps the model quickly get contextual information on the WSI but also weakens its perception of micro tissue structures. Fig. 4b indicates that $\bar{n}_k = 100$ is a compromise for that it achieves the best classification performance. Hence, We set $\bar{n}_k = 100$ in the following experiments.

3) The Output Dimension of Projector: d_p determines the capacity of the contrastive representation learning. In this experiment, we tuned $d_p \in \{64, 128, 256, 512\}$. Correspondingly, the dimension of the hidden layer in the projector and

[2]These metrics were calculated using sklearn library.



(a) Number of blocks N

(b) Patches per kernel $\bar{n}_k$

(c) The projector dim d_p

(d) Aug probability p_{drop}

(e) The loss weight λ

Fig. 4. Performance curves of KAT on the validation data in the five-fold cross-validation as functions of the hyper-parameters, where the red bar indicates the standard deviation of the macro AUCs.

TABLE II
MEAN ± STANDARD DEVIATION OF THE EVALUATION RESULTS ON THE TEST SUBSET FOR THE ABLATION STUDY

Method	Macro-AUC	Weighted-AUC	Accuracy	Macro-F1	Weighted-F1
KAT w/o kernels (ViT)	0.773 ± 0.019	0.940 ± 0.009	0.726 ± 0.056	0.415 ± 0.056	0.732 ± 0.047
KAT w/o softmask	0.816 ± 0.013	0.953 ± 0.003	0.783 ± 0.015	0.488 ± 0.025	0.791 ± 0.014
KAT w/ $\delta_n = \sqrt{\bar{n}_k \cdot 2^0}$	0.843 ± 0.007	0.964 ± 0.004	0.802 ± 0.021	0.506 ± 0.010	0.812 ± 0.017
KAT w/ $\delta_n = \sqrt{\bar{n}_k \cdot 2^{N-1}}$	0.824 ± 0.033	0.963 ± 0.003	0.788 ± 0.042	0.473 ± 0.015	0.788 ± 0.038
KAT w/ $\delta_n = \sqrt{\bar{n}_k \cdot 2^{N-n-1}}$	0.826 ± 0.020	0.965 ± 0.004	0.809 ± 0.017	0.502 ± 0.029	0.818 ± 0.012
KAT w/o KCL	0.847 ± 0.012	0.966 ± 0.002	0.812 ± 0.016	0.513 ± 0.021	0.823 ± 0.013
KAT	0.857 ± 0.007	0.966 ± 0.001	0.823 ± 0.011	0.525 ± 0.015	0.832 ± 0.011

predictor was set $4 \times d_p$ to preserve the inverted-bottle-neck design in BYOL. As shown in Fig. 4c, KAT performs stable. It indicates KAT is not sensitive to d_p. Here, we set $d_p = 64$ to keep KAT light-weighed in the training parse.

4) The Augmentation Probability: p_{drop} controls the average number of the tokens discarded in the augmentation. It affects the diversity of the kernel representations on the online and target branches. Fig. 4d indicates that KAT performed stable to different p_{drop}. We set $p_{drop} = 0.7$ in the following experiments for it achieved the best performance.

5) The Weight in the Loss Function: λ controls the weight in the composition of the WSI classification objective and KCL (Eq. (11)). As shown in Fig. 4e, λ is also an insensitive hyper-parameter for KAT. Here, we empirically set $\lambda = 0.1$. For more discussion about λ please refer to section V.

C. Ablation Study

The results of the ablation study are summarized in Table II. KAT w/o kernels denote KAT without the kernels and masks, where the communication of the patch tokens is achieved by self-attention which is the same as ViT. We observed a significant decrease when comparing KAT w/o kernels with KAT, where the macro-AUC dropped from 0.857 to 0.773 and the macro-F1 dropped from 0.525 to 0.415. It demonstrates the design of the anchor-related kernel attention module in KAT effectively covers the incompatibility of ViT in WSI analysis.

More specifically, we studied the impact of the scale of the masks by changing δ_n in Eq. (2). Specifically, we assigned equivalent small and large masks by respectively setting $\delta_n = \sqrt{\bar{n}_k \cdot 2^0}$ and $\delta_n = \sqrt{\bar{n}_k \cdot 2^{N-1}}$, and also inverted the scale of the masks by setting $\delta_n = \sqrt{\bar{n}_k \cdot 2^{N-n-1}}$. The results are also given in Table II. The macro AUC drops from 0.857 to 0.824 ($\delta_n = \sqrt{\bar{n}_k \cdot 2^{N-1}}$) and 0.826 ($\delta_n = \sqrt{\bar{n}_k \cdot 2^{N-n-1}}$) when large masks are used in the early stages of KAT. Meanwhile, there is also a significant decrease in the metrics when only small masks ($\delta_n = \sqrt{\bar{n}_k \cdot 2^0}$) are used. These results have certified that the design of hierarchical masks from small to large scale is necessary for the KAT model. It helps KAT learn structural patterns from the micro to the macro of the WSI.

TABLE III
COMPARISON OF THE STATE-OF-THE-ART METHODS ON THE *Gastric-2K*, *Endometrial-2K*, AND *TCGA-Lung-3K* DATASETS, WHERE AUC, ACC, AND F1 DENOTE THE MACRO-AVERAGE AREA UNDER ROC CURVE, ACCURACY, AND F1 SCORE, RESPECTIVELY

Method	Gatric-2K (*cls=6*)			Endometrial-2K (*cls=5*)			TCGA-Lung-3K (*cls=3*)		
	AUC (%)	ACC (%)	F1 (%)	AUC (%)	ACC (%)	F1 (%)	AUC (%)	ACC (%)	F1 (%)
RNN-MIL [34]	75.3 ± 3.4	72.4 ± 2.3	41.5 ± 3.1	79.2 ± 1.8	50.4 ± 2.8	48.7 ± 2.5	82.7 ± 2.6	56.0 ± 5.7	54.0 ± 7.8
DSMIL [37]	81.9 ± 0.6	76.0 ± 1.3	43.6 ± 2.3	83.8 ± 1.2	54.3 ± 1.9	52.3 ± 2.1	92.3 ± 0.1	73.9 ± 0.7	75.5 ± 1.0
CLAM [21]	79.2 ± 2.1	77.0 ± 4.2	44.7 ± 3.4	83.5 ± 0.5	55.1 ± 0.4	53.0 ± 0.9	94.6 ± 0.2	83.3 ± 0.6	84.5 ± 0.4
ViT [12]	77.3 ± 1.9	72.6 ± 5.6	41.5 ± 5.6	82.8 ± 1.1	54.4 ± 3.9	50.8 ± 3.5	92.9 ± 0.3	77.8 ± 0.5	79.1 ± 0.8
Nyströmformer [22]	81.6 ± 2.2	77.0 ± 2.5	46.3 ± 3.8	81.5 ± 0.4	56.1 ± 0.9	53.4 ± 0.3	95.5 ± 0.4	83.8 ± 1.3	85.2 ± 1.1
TransMIL [14]	78.1 ± 1.7	81.1 ± 1.4	48.1 ± 2.7	85.0 ± 0.7	55.1 ± 2.3	51.4 ± 1.7	95.6 ± 0.5	82.6 ± 2.0	84.0 ± 2.0
Patch-GCN [18]	81.5 ± 1.9	79.3 ± 3.5	46.5 ± 3.9	84.2 ± 0.7	54.4 ± 4.5	52.2 ± 3.6	96.0 ± 0.2	83.9 ± 0.6	84.8 ± 0.7
LAGE-Net [23]	81.4 ± 0.9	76.9 ± 3.8	44.2 ± 3.5	84.4 ± 0.3	54.4 ± 4.5	52.2 ± 3.6	95.6 ± 0.6	84.2 ± 1.0	85.5 ± 1.1
KAT w/o KCL	84.7 ± 1.2	81.2 ± 1.6	51.3 ± 2.1	86.4 ± 0.8	58.8 ± 1.4	56.1 ± 1.3	96.5 ± 0.1	84.9 ± 0.6	86.4 ± 0.6
KAT	**85.7** ± 0.7	**82.3** ± 1.1	**52.5** ± 1.5	**86.9** ± 0.2	**59.7** ± 0.8	**57.5** ± 0.8	**97.1** ± 0.2	**85.9** ± 0.7	**87.4** ± 0.8

Then, we removed the contrast loss by setting $\lambda = 0$. Table II shows a significant performance gap between KAT and KAT w/o KCL. This result has certified the necessity of the proposed KCL mechanism.

D. Comparison for WSI Classification

We compared our method with 8 methods, including RNN-MIL [34], DSMIL [37], CLAM [21], ViT [12], Nyströmformer [22], TransMIL [14], Patch-GCN [18], and LAGE-Net [23]. For all the Transformer-based models, we uniformly set the embedding dimension $d_e = 256$ and the head dimension $d_h = 64$. The mean and standard deviation results on the test subset through the five-fold cross-validation are presented in Tables III and the subtype ROC curves are drawn in Figs. 5, 6, and 7.

Overall, the proposed method achieved the best performance with a macro-AUC of 0.857 on the *Gastric-2K* dataset, 0.869 on the *Endometrial-2K* dataset, and 0.971 on the *TCGA-Lung-3K* dataset for WSI sub-typing tasks.

RNN-MIL, DSMIL, and CLAM are typical MIL-based methods. The structural relationship of the patches is not considered in these methods. The abandonment of the structural modeling caused generally gaps to the structure-aware methods, e.g., PatchGCN, LAGE-Net, and KAT.

ViT is our baseline method, which describes the patch allocation by trainable positional embeddings. Nyströmformer is a variant of Transformer where a linear approximation solution is proposed to substitute the original self-attention computation. The results show that the macro AUCs of the two methods are 4.3% to 8.4% inferior to the proposed KAT model due to the problem of positional inconsistency and conflict in tokens.

TransMIL achieved better AUCs than Nyströmformer on the endometrium and lung histopathology datasets. Nevertheless, the spatial relationship described by PPEG is also inconsistent with different WSIs. This inconsistency introduced additional noise, especially on the gastric dataset where the tissue area varies a lot to different WSI. It made TransMIL perform even worse than the basic Nyströmformer regarding the AUC on the gastric dataset.

In comparison, the proposed KAT builds uniformly distributed anchors that are adaptive to the shape and size of the tissue region and generates hierarchically region masks for the anchors to describe the local to the global relationship of the patches. The spatial relationship described by KAT is more complete and consistent compared to the previous methods and thereby achieves relatively better performance.

The subtype ROC curves are presented in Figs. 5, 6, and 7 indicate that the proposed KCL strategy significantly improves the performance of KAT for recognizing the lesions with fewer training samples, e.g. gastric MA, SRCC, and endometrial SEIC.

E. Comparison of Computational Complexity

We calculated the model parameters, floating point operations (FLOPs) for model inference, and the maximum GPU memory cost for train/test stages to show the computation and memory efficiency of the compared methods. The results are presented in Table IV, where only the models for WSI prediction are counted and the CNN models for feature extraction are not included. For all the Transformer-based methods, the results are obtained based on 6 stacked blocks with 8 heads.

RNN-MIL has a constant low cost in computation and memory for all the WSIs since it only takes the top-K activated patches as the input. DSMIL and CLAM are typical instance-based MIL models, where no interactive operations for patch pairs are involved. Table IV shows that the computational amount and the memory usage for the two methods are linearly increased as n_p enlarges. It also demonstrates that both the FLOPs and memory cost for ViT are quadratic related to n_p. Especially, for large WSIs that contain 16,000 foreground patches, the memory cost of ViT is over 24 GB, which is already out-of-memory for an Nvidia RTX 3090 GPU. In comparison, the proposed KAT has a linear increase in computational amount and memory cost to n_p owing to the linear property of the KA module. Specifically, the inference FLOPs and memory cost for a WSI within 16,000 patches is 95.1 G and 540 MB, respectively, which is comparable with the models based on linear-approximated Transformer, i.e., TransMIL and Nyströmformer. This property makes KAT has significant advantages to ViT in processing large WSIs.

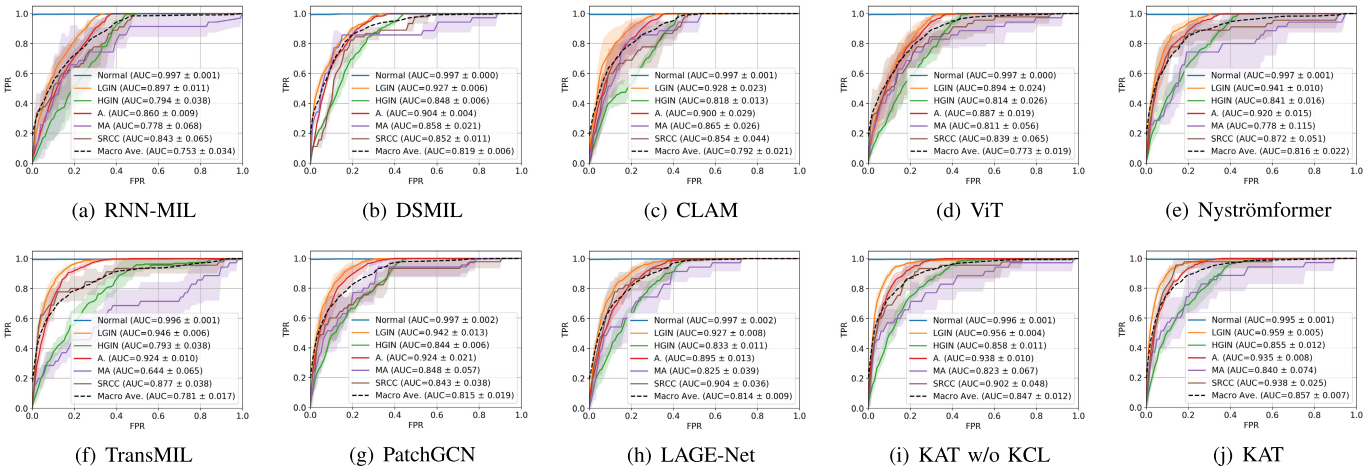


Fig. 5. Comparison of sub-typing ROC curves of state-of-the-art methods on the Gastric-2K dataset, where the solid line curves are the mean values of the cross-validation and shaded area under the curve represents the standard deviation.

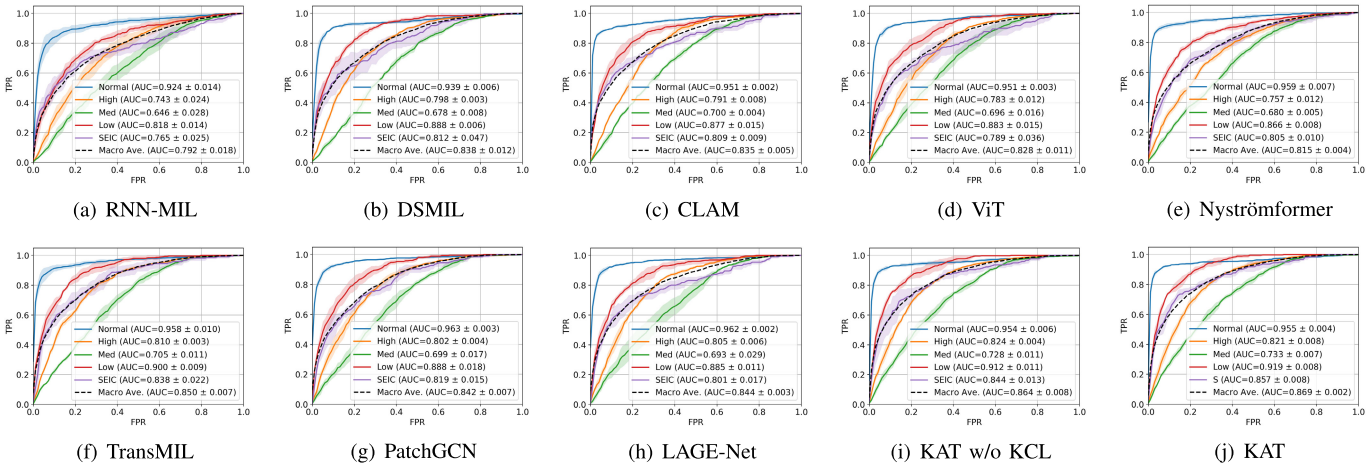


Fig. 6. Comparison of sub-typing ROC curves of state-of-the-art methods on the Endometrial-2K dataset.

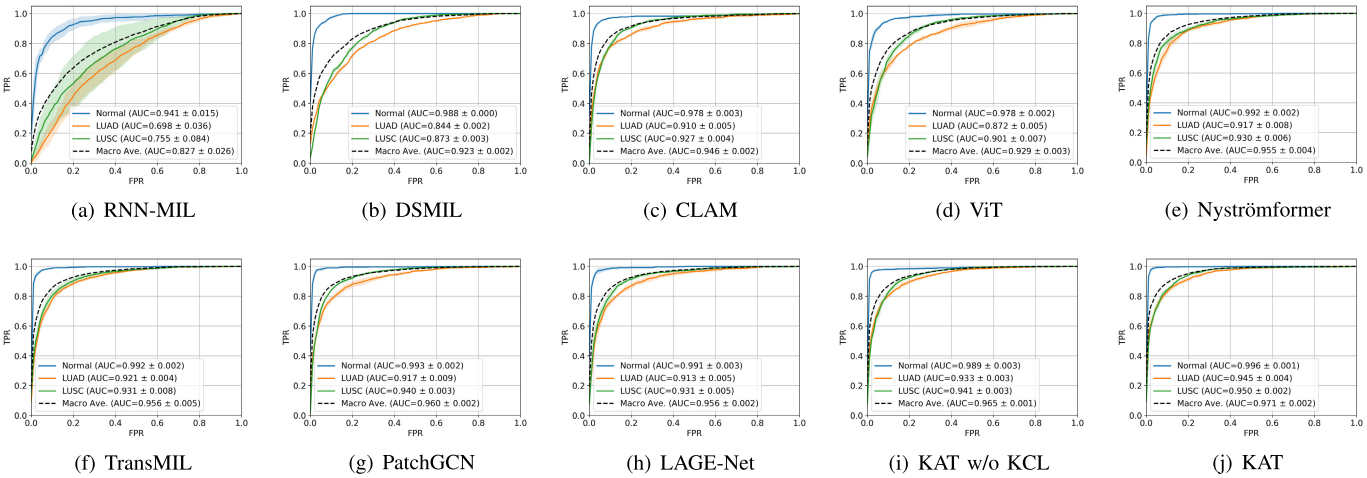


Fig. 7. Comparison of sub-typing ROC curves of state-of-the-art methods on the TCGA-LUNG-3K datasets.

F. Visualization

The interpretability is equally important as the accuracy for assistant cancer diagnosis based on WSI analysis. It can help pathologists determine whether to adopt the suggestion of AI. The design of the kernel attention and kernel representation learning of KAT make it more eligible to generate visual outputs. In this section, we present the interpretability of KAT in aspects of representative regions indication and content-based similar region retrieval.

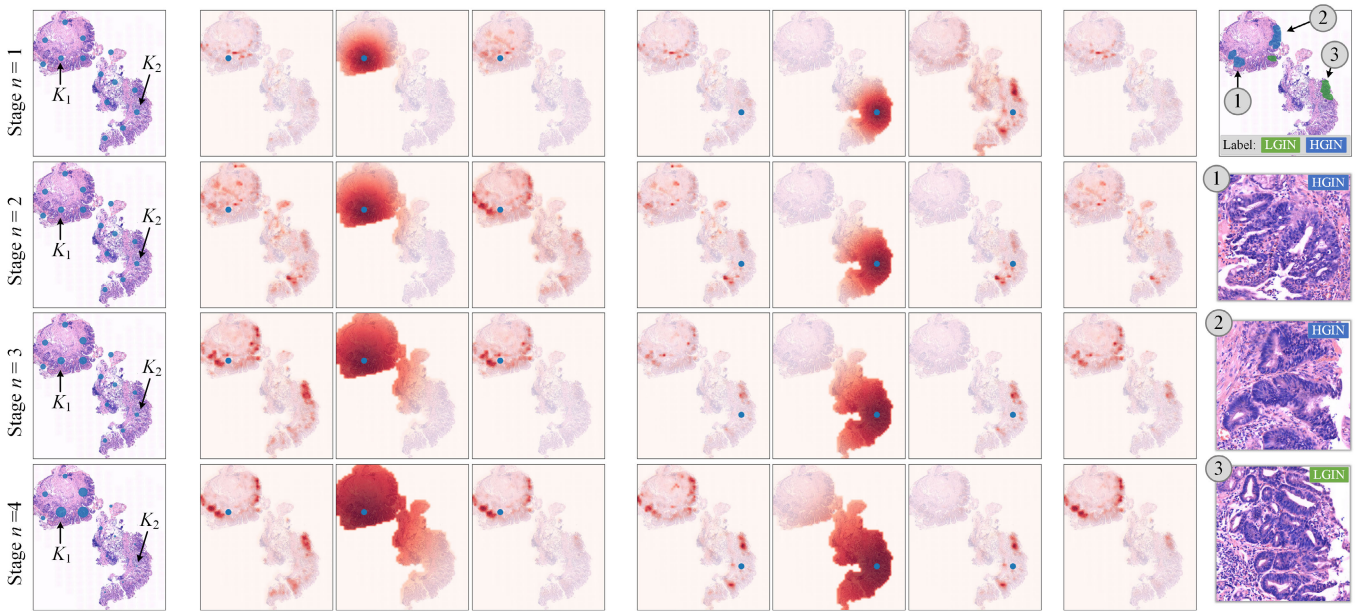
TABLE IV

COMPARISON OF MODEL PARAMETERS (PARAMS), FLOATING POINT OPERATIONS (FLOPS), AND TRAIN/TEST GPU MEMORY COST (MEM.) FOR WSIS WITH THE DIFFERENT NUMBER OF FOREGROUND PATCHES (n_p), WHERE THE GPU MEMORY COST IS TESTED WITH ONE SAMPLE BY SETTING THE BATCH SIZE TO 1

Method	Params (MB)	$n_p = 500$ FLOPs ($\times 10^9$)	$n_p = 500$ Mem. (MB)	$n_p = 2000$ FLOPs ($\times 10^9$)	$n_p = 2000$ Mem. (MB)	$n_p = 4000$ FLOPs ($\times 10^9$)	$n_p = 4000$ Mem. (MB)	$n_p = 8000$ FLOPs ($\times 10^9$)	$n_p = 8000$ Mem. (MB)	$n_p = 16000$ FLOPs ($\times 10^9$)	$n_p = 16000$ Mem. (MB)
RNN-MIL [34]	1.25	0.09*	44* / 27*	-	- / -	-	- / -	-	- / -	-	- / -
DSMIL [37]	0.34	0.14	11 / 5	0.56	24 / 19	1.1	47 / 37	2.3	90 / 73	4.5	183 / 150
CLAM [21]	1.38	0.77	66 / 19	2.8	78 / 44	5.3	138 / 77	10.9	265 / 142	21.8	514 / 277
ViT [12]	2.51	5.8	119 / 49	56.7	1192 / 367	205.2	4442 / 1336	777.5	17230 / 5152	3023.7	OOM
Nyströmformer [22]	1.51	3.1	169 / 29	12.2	351 / 79	24.6	634 / 155	50.6	1162 / 296	111.7	2260 / 588
TransMIL [14]	1.52	3.1	170 / 29	12.3	358 / 80	24.7	646 / 155	50.8	1186 / 296	112.0	2309 / 588
PatchGCN [18]	1.44	2.4	30 / 14	7.8	103 / 43	18.5	200 / 78	51.2	396 / 151	162.4	786 / 299
LAGE-Net [23]	3.13	5.9	116 / 42	61.0	1204 / 363	221.4	4473 / 1336	840.5	17297 / 5160	3271.8	OOM
KAT (w/o KCL)	1.21	3.0	58 / 19	11.8	154 / 49	23.7	333 / 96	47.4	792 / 210	95.1	2154 / 540
KAT (w/ KCL)	2.51**		62 / 19		171 / 49		360 / 96		846 / 210		2297 / 540

* RNN-MIL takes the top-K activated patches as the input. The FLOPs and memory cost is constant for different WSI.

** The complete model for self-supervised training contains the parameters for both the online branch and the target branch. The parameter number for inference is 1.21 MB, the same as KAT (w/o KCL).



(a) Kernel Weight (b) K_1's raw attention map (c) K_1's soft-mask (d) K_1's masked attention map (e) K_2's raw attention map (f) K_2's soft-mask (g) K_2's masked attention map (h) Weighted attention map (i) Ground Truth

Fig. 8. Visualization based on the kernel attention mechanism in KAT with $N = 4$, where (a) displays the attention weights of the kernels with blue dots, (b-d) respectively visualize the raw attention maps, the soft-masks, and the masked attention map for the kernel K_1, (e-g) are the maps for kernel K_2, (h) is the integrated attention map, and (i) provides the ground truth (the top-right) and representative tumor regions (indexed by digits). For more details please refer to section IV-F.

1) Representative Regions Indication: The $\mathbf{A}^{(n)}_{c\leftarrow k} \in \mathbb{R}^K$ obtained in Eq. (6) indicates the weights of the kernels, i.e. the contribution of the anchor-related regions, for the WSI classification. The visualization of $\mathbf{A}^{(n)}_{c\leftarrow k}, n = 1, 2, 3, 4$ for a typical gastric WSI is presented in Fig. 8a, where the blue dots indicate the position of the anchor and the size of the dots is positively correlated to the attention weight. As shown in the ground truth (Fig. 8i), the WSI contains two pathological types of tissue, i.e. HGIN and LGIN, and therefore the WSI is categorized as HGIN according to the diagnosis priority. It is obvious that KAT has gradually increased the attention in the anchors located near HGIN tissues, and simultaneously suppressed the attention in the normal and LGIN regions. This behavior made KAT correctly predict the subtype of the WSI.

Furthermore, the attention map for a certain anchor can be obtained by projecting the values of $\mathbf{A}^{(n)}_{k\leftarrow x}$ and $\mathbf{M}^{(n)}$ (see Eq. (4)) on the WSI. Fig. 8(b-d) visualizes a typical anchor K_1, where Fig. 8b shows the raw attention maps resolved from the K_1-th row of $\mathbf{A}^{(n)}_{k\leftarrow x}$, Fig. 8c displays K_1's soft-masks resolved from $\mathbf{M}^{(n)}$, and Fig. 8d provides K_1's masked attention maps that are resolved from the masked attention matrix $\tilde{\mathbf{A}}^{(n)}_{k\leftarrow x} = \mathbf{A}^{(n)}_{k\leftarrow x} \odot \mathbf{M}^{(n)}$. Similarly, Fig. 8(e-g) provides the visualization for another typical anchor K_2. It shows that both K_1 and K_2 originally highlight the regions within LGIN and HGIN tissues, which are highly correlated with the

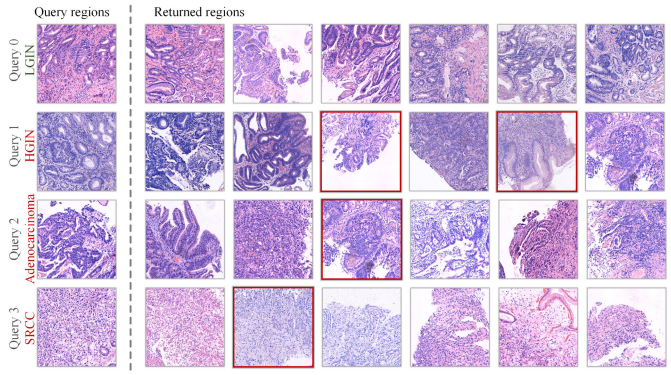


Fig. 9. Display of similar regions retrieval, where the first column provides the regions for 4 query items in different lesion types, the top-returned regions from the database for each query item are ranked on the right, and the irrelevant return regions (have different labels with the query item) are framed in red.

TABLE V
RESULTS FOR THE ABLATION STUDY, WHERE THE AVERAGE PRECISION OF THE TOP-K RETURNED ITEMS (P@K), THE MEAN AVERAGE PRECISION (MAP), MEAN RECIPROCAL RANK (MRR) OF THE TOP-1 CORRECT ITEM, AND THE MEAN RECALL OF THE TOP-K ITEMS (R@K) ARE CALCULATED AS METRICS

Stage	P@5	P@20	mAP	MRR	R@5
1	0.722	0.720	0.718	0.796	0.897
2	0.783	0.769	0.753	0.839	0.905
3	0.808	0.809	0.808	0.847	0.887
4	0.826	0.814	0.811	0.859	0.892

pathologist's annotation (see Fig. 8i). Then, the two kernels changed to observe their nearby lesions after being masked. In this way, K_1 turned to be an HGIN-representative anchor and K_2 to be an LGIN anchor. It is the reason that the classification token gradually enhanced the attention to K_1 and meanwhile weakened the attention to K_2 in decision making.

Finally, we can generate an integrated attention map by resolving the matrix $\mathbf{A}^{(n)}_{c\leftarrow x} = \mathbf{A}^{(n)\mathrm{T}}_{c\leftarrow k}\tilde{\mathbf{A}}^{(n)}_{k\leftarrow x}$ as visualized in Fig. 8h. It is apparent that $\mathbf{A}^{(n)}_{c\leftarrow x}$ explicitly highlights the diagnostic relevant regions of the WSI. Therefore, it is valuable in assisting pathologists in diagnosis.

2) Performance for Similar Anchor Retrieval: One motivation of the proposed kernel contrastive learning is to improve the discrimination of the kernel representation. As visualized in the previous section, KAT can highlight the diagnostically relevant regions of the WSI. These properties determine that KAT is adequate to build a content-based histopathology image retrieval system, which is also a promising application for interpretable assistant diagnosis [23], [59], [60].

In this experiment, we extracted the most attentive representation from $\mathbf{K}^{(n)}$ in the n-th scale and represent it as $\hat{\mathbf{k}}^{(n)}$. For each scale, we built the retrieval database by collecting $\hat{\mathbf{k}}^{(n)}$ from the training set. Then, we regarded the $\hat{\mathbf{k}}^{(n)}$ for each WSI in the test set as the query item. Then, the top-K similar items for each query item were retrieved from the database. The retrieval performance was evaluated following [23]. The results are presented in Table V. It shows that the accuracy of retrieval gradually increases as the scale n enlarges and the mAP for retrieval reached 0.811 based on $\hat{\mathbf{k}}^{(4)}$, i.e. the kernel representation from the fourth block of KAT. It is a considerable accuracy for histopathology image retrieval systems. Meanwhile, P@5 and R@5 are 0.826 and 0.892, respectively, for $n = 4$. This means that doctors can efficiently receive the diagnostically relevant cases by only checking the top-5 returned results. Fig. 9 displays 4 instances of retrieval with $\hat{\mathbf{k}}^{(4)}$, where a region in size of 2048 × 2048 centered on the most attentive area of $\hat{\mathbf{k}}^{(4)}$ under 20× lens is extracted to represent the WSI. It shows that the returned regions are contently similar to the query ones and meanwhile the majority WSIs share the correct labels. The visualization is consistent with the metrics presented in Table V. These results demonstrate the effectiveness of the KCL strategy in the aspect of discriminative kernel representation learning. It also shows the capacity of KAT in diagnostically relevant regions mining and retrieval for WSIs, which is practical for building intelligent and informative cancer diagnosis systems.

TABLE VI
INFERENCE TIME (IN SECOND) OF THE PROPOSED FRAMEWORK AS A FUNCTION OF THE NUMBER OF PATCHES IN THE WSI. THE MEAN AND STANDARD DEVIATION TIME ON THE GASTRIC-2K DATASET FOR PREPROCESSING AND SIMILAR REGION RETRIEVAL ARE REPORTED FOR THEIR TIME COST ARE NOT DIRECTLY CORRELATED WITH THE NUMBER OF FOREGROUND PATCHES

#Patches (n_p)	500	2000	4000	8000	16000
Preprocessing	0.024 ± 0.013		(mean ± std on the dataset)		
Data loading	2.239	8.959	17.919	35.838	71.676
CNN inference	0.956	1.670	2.568	4.396	8.009
Anchor clustering	0.291	0.372	0.502	0.650	0.829
KAT inference	0.048	0.048	0.050	0.052	0.052
Retrieval	0.012 ± 0.001		(mean ± std on the dataset)		
Total	3.6	11.1	21.1	40.9	80.6

G. Running Time

In the end, we evaluated the running time of the proposed framework under our experiment environment. The results are shown in Table VI. The major running time was from the patch data loading. Typically, it costs 35.838 s to load 8,000 patches from the disk to GPU. The time for extracting features from the 8,000 patches (with a batch size of 512) is 4.396 s and for anchor clustering is 0.650 s. The inference time for KAT is around 50 ms. This time is insensitive to the number of patches owning to the parallel computation capacity of the GPU. The preprocessing stage, including the foreground segmentation and foreground patch detection, spent 24 ms, and the average retrieval time is 12 ms, which is negligible for the entire running time.

V. DISCUSSION

Notably, the kernels in KAT are bound with anchors that indicate explicit locations on the WSI. Therefore, we can regard the representation of a kernel as the feature of the region where the corresponding anchor mask is located. It has a higher level of semantics than patch-level representations.

Moreover, the proposed hierarchical masking strategy makes the kernel presentation able to describe the tissue on different scales. These properties of the kernels allow KAT to produce more interpretable outputs for assistant diagnosis. That is one of the motivations we built contrastive constraints on the kernel representations rather than patch representations.

As one role of the validation data in our framework was to perform early stop, we report the best AUCs achieved on the validation data in Fig. 4e. It is one reason that the results for different λ settings appear to be very close. Actually, we found λ has a distinct effect on the training procedure of the model. Fig. 10 plots the average macro AUCs of the five trials as a function of the training epoch, where the best value for each curve is marked by a dot. When $\lambda = 0.001$ and $\lambda = 0$, i.e., without contrastive loss, the best AUCs appeared in the early stage of training. But, the AUCs decreased as training continued. It means the model was already overfitted to the validation data. In contrast, when set $\lambda = 0.01$ and $\lambda = 0.1$, the AUCs stably increased along with the training progress and reached the best value at the end of training. It indicates a good generalization performance of the model to unseen data, e.g. the test data. That is the reason why the model with $\lambda = 0.1$ achieved noticeable improvement on the test set compared to the model without contrastive loss, as shown in Table II. Based on these results, we concluded that the contrastive loss with a weighting of $\lambda = 0.01$ and $\lambda = 0.1$ are appropriate to our method.

SimCLR [61], MoCov2 [62], and BYOL [54] are popular self-supervised representation learning frameworks applied in histopathology WSI analysis. SimCLR commonly requires a large batch size of up to 4096 for full performance [61]. MoCov2 needs a memory bank in size of 65K and multiple GPUs for full performance [62]. In contrast, BYOL does not rely on a memory bank and meanwhile maintains satisfactory performance with a small batch size of 256 [54], which is computationally efficient. Moreover, it achieves better performance than SimCLR and MoCov2 for multiple downstream tasks [54]. These properties made us choose BYOL as the representation learner for the CNN and KAT.

In our previous work [52], we utilized the EfficientNet-B0 [63] as the feature extractor for the reason that it achieved an overall best performance in the previous studies based on the EfficientNet-B0 pre-trained on ImageNet dataset [64], [65]. But, through an extended study, we found that ResNet-50 achieved a stable improvement from 0.71% to 1.83% of macro AUC compared to EfficientNet-B0 when trained under the contrastive representation learning paradigm. We suppose the main reason is that contrastive representation learning requires a considerable number of parameters to achieve good performance. Therefore, we changed the feature extraction network from EfficientNet-B0 to ResNet-50 in this work.

The ground truths of the subtypes of the endometrial dataset and gastric dataset are certified by senior pathologists. The subtype of HGIN in the gastric dataset is a borderline tumor between the subtype of LGIN and adenocarcinoma, which is relatively hard to distinguish in clinical diagnosis. In the experiments, we also observed an apparent prediction confusion by the compared models for the subtypes LGIN, HGIN, and adenocarcinoma, which is consistent with the clinical diagnosis. Similarly, MDEA is an intermediate histopathology type between WDEA and LDEA, and thereby there is an apparent prediction confusion of these subtypes. That is the main reason all the compared models obtained a relatively low accuracies and F1-scores, around 0.5 to 0.6, on the two datasets.

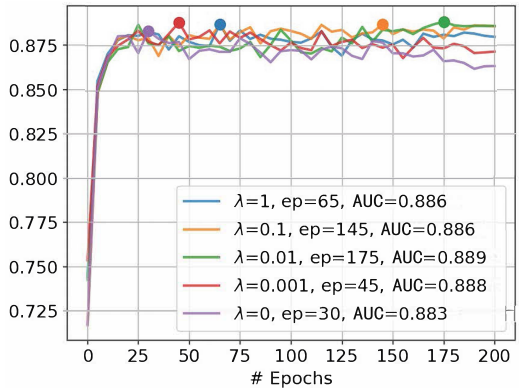


Fig. 10. Average macro AUCs of the five trials on the validation data as a function of the training epoch, where the best AUC values and the corresponding epochs are marked dots.

The proposed method utilized the dot-product operation, following Transformer, to calculate the cross-attention matrix between the anchor representations and patch representations. Theoretically, the dot-product operation can also be replaced by other operations, e,g. radial basis function (RBF) and Laplace kernel function as suggested in [66].

There are also related works that build trainable vectors and perform cross-attention operations for fine-grained image understanding. These trainable vectors are named as *Learned Part Dictionary* in [67]. And in more recent works [68], [69], these vectors are also named as *Kernels*. The outline structure of the proposed model fully follows the Transformer, including the usage of the LayerNorm module, the structure of the MLP, the usage of residual connections, the multi-head design, and the stacking strategy of the KA blocks. It is the reason that we named the proposed model kernel attention Transformer.

The current model defines uniformly distributed masks to guide anchors in learning specific regions of tissue. It is a generic modeling strategy that does not take into account differences between cancer types. This is a drawback of the work that can be further improved. One of the future works will focus on building cancer-type sensitive anchor masks based on trainable relative distance embeddings, to improve the description ability for tissue structure. Another future work will focus on aligning the kernels with the diagnosis reports of the WSIs and building a case-level vision-language model based on KAT for more intelligent computer assistant cancer diagnosis.

REFERENCES

[1] P. Bándi et al., "From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 550–560, Feb. 2019.

[2] Z. Song et al., "Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning," *Nature Commun.*, vol. 11, no. 1, p. 4294, 2020.

[3] H. Yu et al., "Large-scale gastric cancer screening and localization using multi-task deep neural network," *Neurocomputing*, vol. 448, pp. 290–300, Aug. 2021.

[4] W. Bulten et al., "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge," *Nature Med.*, vol. 28, no. 1, pp. 154–163, 2022.

[5] A. Raju, J. Yao, M. M. Haq, J. Jonnagaddala, and J. Huang, "Graph attention multi-instance learning for accurate colorectal cancer staging," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 529–539.

[6] J. Xu et al., "Computerized spermatogenesis staging (CSS) of mouse testis sections via quantitative histomorphological analysis," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101835.

[7] Y. Fu et al., "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis," *Nature Cancer*, vol. 1, no. 8, pp. 800–810, 2020.

[8] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101789.

[9] 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.

[10] N. Coudray et al., "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nature Med.*, vol. 24, no. 10, pp. 1559–1567, 2018.

[11] R. Yamashita et al., "Deep learning model for the prediction of microsatellite instability in colorectal cancer: A diagnostic study," *Lancet Oncol.*, vol. 22, no. 1, pp. 132–141, 2021.

[12] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–22.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 213–229.

[14] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, and X. Ji, "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2136–2147.

[15] Q. Da et al., "DigestPath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102485.

[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.*, Jun. 2022, pp. 16144–16155.

[17] Z. Gao et al., "Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 299–308.

[18] R. J. Chen et al., "Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 339–349.

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

[20] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[21] M. Y. Lu, D. F. 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.

[22] Y. Xiong et al., "Nyströmformer: A Nyström-based algorithm for approximating self-attention," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 16, p. 14138.

[23] Y. Zheng et al., "Encoding histopathology whole slide images with location-aware graphs for diagnostically relevant regions retrieval," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102308.

[24] J. van der Laak, G. Litjens, and F. Ciompi, "Deep learning in histopathology: The path to the clinic," *Nature Med.*, vol. 27, no. 5, pp. 775–784, May 2021.

[25] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2424–2433.

[26] 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, no. 22, pp. 2199–2210, Dec. 2017.

[27] S. Wang et al., "RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101549.

[28] 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.

[29] J. Donahue, P. Krähenbuhl, and T. Darrell, "Adversarial feature learning," 2016, *arXiv:1605.09782*.

[30] T. Xiang et al., "DSNet: A dual-stream framework for weakly-supervised gigapixel pathology image analysis," *IEEE Trans. Med. Imag.*, vol. 41, no. 1, pp. 2180–2190, Aug. 2022.

[31] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 316–325, Jan. 2017.

[32] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.

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

[34] 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, Aug. 2020.

[35] 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.*, Jun. 2020, pp. 3852–3861.

[36] Y. Zhao et al., "Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4837–4846.

[37] 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.*, Jun. 2021, pp. 14318–14328.

[38] L. Qu, X. Luo, S. Liu, M. Wang, and Z. Song, "DGMIL: Distribution guided multiple instance learning for whole slide image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 24–34.

[39] 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.*, Jun. 2022, pp. 18802–18812.

[40] J. Yang et al., "ReMix: A general and efficient framework for multiple instance learning based whole slide image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 35–45.

[41] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 273–278.

[42] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[43] 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. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 174–182.

[44] Y. Guan et al., "Node-aligned graph convolutional network for whole-slide image representation and classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 18813–18823.

[45] W. Lu, M. Toss, M. Dawood, E. Rakha, N. Rajpoot, and F. Minhas, "SlideGraph^{+}: Whole slide image level graphs to predict HER2 status in breast cancer," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102486.

[46] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 10012–10022.

[47] 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. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 561–570.

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

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

[50] 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.*, Oct. 2021, pp. 4015–4025.

[51] 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. ICLR*, 2021, pp. 1–16.

[52] Y. Zheng, J. Li, J. Shi, F. Xie, and Z. Jiang, "Kernel attention transformer (KAT) for histopathology whole slide image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2022, pp. 283–292.

[53] Y. Zheng et al., "Histopathological whole slide image analysis using context-based CBIR," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1641–1652, Jul. 2018.

[54] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

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

[56] X. Wang et al., "Transformer-based unsupervised contrastive learning for histopathological image classification," *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 102559.

[57] J. Li, Y. Zheng, K. Wu, J. Shi, F. Xie, and Z. Jiang, "Lesion-aware contrastive representation learning for histopathology whole slide images analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 273–282.

[58] J. Li et al., "Weakly supervised histopathological image representation learning based on contrastive dynamic clustering," in *Proc. SPIE*, vol. 12039, pp. 14–19, Apr. 2022.

[59] P. Chen, X. Shi, Y. Liang, Y. Li, L. Yang, and P. D. Gader, "Interactive thyroid whole slide image diagnostic system using deep representation," *Comput. Methods Programs Biomed.*, vol. 195, Oct. 2020, Art. no. 105630.

[60] S. Kalra et al., "Yottixel—An image search engine for large archives of histopathology whole slide images," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101757.

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

[62] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[63] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[64] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, and J. Shi, "Tracing diagnosis paths on histopathology WSIs for diagnostically relevant case recommendation, in *Medical Image Computing and Computer Assisted Intervention—MICCAI* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2020, pp. 459–469.

[65] Y. Zheng et al., "Diagnostic regions attention network (DRA-Net) for histopathology WSI recommendation and retrieval," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 1090–1103, Mar. 2021.

[66] D. Rymarczyk, A. Borowa, J. Tabor, and B. Zielinski, "Kernel self-attention for weakly-supervised image classification using deep multiple instance learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 1721–1730.

[67] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8662–8672.

[68] W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-Net: Towards unified image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 10326–10338.

[69] X. Li et al., "Video K-Net: A simple, strong, and unified baseline for video segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 18847–18857.