Encoding histopathology whole slide images with location-aware graphs for diagnostically relevant regions retrieval

Yushan Zheng[a], Zhiguo Jiang[b,a,*], Jun Shi[d,*], Fengying Xie[b,a], Haopeng Zhang[b,a], Wei Luo[b,a], Dingyi Hu[b,a], Shujiao Sun[b,a], Zhongmin Jiang[c], Chenghai Xue[e,f]

[a] *Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing 100191, China*
[b] *Image Processing Center, School of Astronautics, Beihang University, Beijing 102206, China*
[c] *Department of Pathology, Tianjin Fifth Central Hospital, Tianjin 300450, China*
[d] *School of Software, Hefei University of Technology, Hefei 230601, China*
[e] *Wankangyuan Tianjin Gene Technology, Inc, Tianjin, 300220, China*
[f] *Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China*

Abstract

Content-based histopathological image retrieval (CBHIR) has become popular in recent years in histopathological image analysis. CBHIR systems provide auxiliary diagnosis information for pathologists by searching for and returning regions that are contently similar to the region of interest (ROI) from a pre-established database. It is challenging and yet significant in clinical applications to retrieve diagnostically relevant regions from a database consisting of histopathological whole slide images (WSIs). In this paper, we propose a novel framework for regions retrieval from WSI database based on location-aware graphs and deep hash techniques. Compared to the present CBHIR framework, both structural information and global location information of ROIs in the WSI are preserved by graph convolution and self-attention operations, which makes the retrieval framework more sensitive to regions that are similar in tissue distribution. Moreover, benefited from the graph structure, the proposed framework has good scalability for both the size and shape variation of ROIs. It allows the pathologist to define query regions using free curves according to the appearance of tissue. Thirdly, the retrieval is achieved based on the hash technique, which ensures the framework is efficient and adequate for practical large-scale WSI database. The proposed method was evaluated on an in-house endometrium dataset with 2650 WSIs and the public ACDC-LungHP dataset. The experimental results have demonstrated that the proposed method achieved a mean average precision above 0.667 on the endometrium dataset and above 0.869 on the ACDC-LungHP dataset in the task of irregular region retrieval, which are superior to the state-of-the-art methods. The average retrieval time from a database containing 1855 WSIs is 0.752 ms. The code is available at https://github.com/zhengyushan/lagenet

Keywords: Histopathological image analysis, CBIR, Computer-aided cancer diagnosis, Graph convolutional network, Self-attention

1. Introduction

With the development of digital pathology and artificial intelligence, computer-aided cancer diagnosis methods based on histopathological image analysis (HIA) Litjens et al. (2017); Gurcan et al. (2009); Hollon et al. (2020) have been widely studied. In recent years, the stud-

[*] Corresponding authors.
e-mail: jiangzg@buaa.edu.cn (Z. Jiang), juns@hfut.edu.cn (J. Shi)

ies focused on the histopathological whole slide image classification Zheng et al. (2017); Xu et al. (2017), segmentation Xu et al. (2014); Bejnordi et al. (2016, 2017); Jia et al. (2017); Falk et al. (2019), object detection Xu et al. (2015); Veta et al. (2019), etc. Generally, these applications can provide the pathologists diagnosis suggestions and even automatically generate reports within quantitative data and diagnostic descriptions. However, these applications can hardly provide the dependence or reason of the decision. The information for diagnoses is limited.

Content-based histopathological image retrieval (CBHIR) is an emerging approach in the domain of HIA Zhang and Metaxas (2016); Li et al. (2018); Zheng et al. (2018a); Kalra et al. (2020). Compared to the typical HIA methods mentioned above, CBHIR methods provide more valuable information, including similar regions from diagnosed cancer cases, the corresponding meta-information, and the diagnosis reports of experts stored along with the cases in the digital pathology platform. It can increase the information and improve the interpretability of the automatic diagnosis, which is of developmental significance to pathologists.

With the rapid expansion of digital WSIs archive, it is recently crucial to develop effective retrieval systems for large-scale WSI database. However, the histopathology WSIs are gigapixel digital images with complex textural information and the query image is a region of interest (ROI) in various size and shape. It makes the CBHIR for the WSI database a challenging task. The existing methods are confronted with many difficulties, which are reflected in three aspects: 1) Due to the constraint of the deep learning model, the regions to establish the database and the query region are limited to rectangle in a fixed size Ma et al. (2017); Shi et al. (2018); Peng et al. (2019). In this case, multiple models need to be established for the retrieval requirement in different sizes. 2) The retrieval for large regions is commonly completed by measuring the distance between two sets of local features Jimenez-del Toro et al. (2017); Zheng et al. (2018b); Chen et al. (2020). The internal adjacency relationship of these local features is not considered, and meanwhile, the computational complexity is expensive. 3) The sub-regions are cropped from the WSI and then regarded as independent items Ma et al. (2018); Zheng et al. (2019) in the database. The global location information of the sub-region in the WSI is discarded. The above problems lead to a series of defects in precision, efficiency, and convenience of the retrieval system when applied to the practical database.

In this paper, we simultaneously address the above three problems and propose a novel CBHIR framework for diagnostically relevant region retrieval from large-scale WSI-database based on graphs and deep hashing method. Unlike the present sub-region retrieval frameworks, we proposed constructing location-aware graphs (LA-Graph) for the sub-regions in the WSI to describe both the structural information within the regions and the global location information of the region in the WSI. Meanwhile, we designed a novel location-aware graph encoding network (LAGE-Net) based on graph convolution and self-attention operations to encode the LA-Graph for retrieval. Moreover, we employed the hashing technique to ensure the efficiency of retrieval. The experiments on two large-scale datasets have demonstrated the effectiveness of our method.

The contribution of this paper to the problem is three-fold:

1) We proposed a novel histopathology image retrieval framework based on location-aware graphs for databases consisting of whole slide images. To our knowledge, we are the first to use the graphs to simultaneously represent the image content, the internal adjacency and the global location information for histopathology ROIs. Specifically, the local features are regarded as the nodes of the graph, the spatial connection relations of the features are described as the edges of the graph and distances of patches to the border of the tissue are represented by distance embedding. The definition of LA-Graph determines the sub-regions for retrieval are size- and shape-scalable. It allows the pathologists to define query regions by free-curves.

2) We designed a LAGE-Net to encode the LA-Graph into binary codes, which are used to index the sub-regions in the WSI and the query ROI. The spatial adjacency information in a LA-Graph is extracted by graph convolution operations and the global location information is modeled along with the local features by self-attention operations. Finally, multiple information is combined and converted into binary codes by a hash module. The local features, spatial information and location information for a sub-region in the WSI are effectively extracted and preserved by the LAGE-Net. The LAGE-Net can be trained

end-to-end from graphs with a variable number of nodes to the binary-like codes and the retrieval can be effectively achieved based on hamming distances between binary codes. It determines the proposed method is applicable to practical large-scale WSI databases.

3) We conducted comprehensive experiments to verify the proposed retrieval framework on an in-house endometrium dataset with 2650 WSIs and the public ACDC-LungHP dataset with 150 WSIs and compared it with 6 state-of-the-art methods. The experimental results have demonstrated that the proposed method achieves the best performance in the task of irregular region retrieval with a mean average precision above 0.667 on endometrium dataset and above 0.869 on the ACDC-LungHP dataset. The average retrieval time from a database within 1855 WSIs is 0.752 ms.

The remainder of this paper is organized as follows. Section 2 reviews the history of histopathological image retrieval. Section 3 introduces the methodology of the proposed method. The experiment and discussion are presented in Section 4. Section 5 summarizes the contributions. A part of this work has been presented on the conference paper Zheng et al. (2019).

2. Related Works

The objects in the studies on CBHIR have been developed through cells/nuclei, image patches and whole slide images with the development of digital pathology. The typical methods related to our work are reviewed in this section.

2.1. Retrieval methods for cells and patches

Early studies focused on the cell retrieval from histological images that were captured under the optical microscopy Comaniciu et al. (1998b,a); Wetzel et al. (1999). With the development of the digitalization of histological sections, CBHIR frameworks for patches retrieval were proposed. Zheng et al. (2004); Zhou and Jiang (2004) and Mehta et al. (2009) employed the classical image features to depict the histopathological images and achieved the patch-level retrieval. Then, the retrieval methodology was studied in various aspects.

A number of works concentrated on extracting high-level features of histopathological images to improve the accuracy of retrieval. Specifically, CBHIR frameworks based on manifold learning Doyle et al. (2007); Sparks and Madabhushi (2011), semantic analysis Caicedo et al. (2008); Caicedo and Izquierdo (2010); Zheng et al. (2014), spectral embedding Sridhar et al. (2011) and fine-designed local descriptors Tizhoosh and Babaie (2018); Erfankhah et al. (2019) have been developed and have proven effective in improving the accuracy of retrieval. Meanwhile, Gu and Jie (2018); Zheng et al. (2018a); Gu and Yang (2019) proposed utilizing the contextual information by combining features from multiple magnifications of histopathological images to enhance the representations of image patches and thus improve the performance of retrieval. As for the online usage of CBHIR, the security of retrieval has also been considered Cheng et al. (2019).

Besides the retrieval accuracy, the efficiency of CBHIR has become increasingly popular in the recent years. To satisfy the application for database consisting of massive histopathological images, hashing techniques were introduced. Typically, Zhang et al. (2015b) ,Zhang et al. (2015a) and Jiang et al. (2016) introduced supervised hashing with kernels (KSH) Liu et al. (2012) into the CBHIR. Shi et al. (2017) utilized a graph hashing model to learn the similarity relationship of histopathological images. With hashing functions, the images are encoded into an array of binary codes. And the similarities among images are measured by Hamming distance, which is able to be calculated very efficiently using bitwise operations by computer. More recently, Shi et al. (2018), Sapkota et al. (2018) and Peng et al. (2019) constructed end-to-end deep learning frameworks based on CNNs to directly encode histopathological images into binary codes. The overall performance of patch-level CBHIR has been further improved.

2.2. Whole slide image database retrieval

The practical digital histopathology scans are generally stored in the format of whole slide images. Therefore, it is crucial to study the approach for retrieving relevant sub-regions from the WSIs for a region the pathologist provided during the diagnosis.

In the previous study, Ma et al. (2017) proposed dividing the WSIs into sub-regions following the sliding window paradigm and encoding the individual regions to establish the retrieval database. It is a convenient strategy to index WSIs for sub-regions retrieval. However, the tissue structure was ignored in the division of WSIs and retrieval instances in the database were limited to rectangle images in fixed sizes. It gaps from the applicable situation where pathologists usually define the ROIs with free-carves in various shapes and sizes.

Then, several retrieval strategies were developed to improve the scalability of the retrieval framework. Zheng et al. (2018a) proposed segmenting a WSI into superpixels and defining the super-pixels as retrieval instances. Further, Ma et al. (2018) proposed merging the superpixels into irregular regions based on selective search Uijlings et al. (2013) to index the WSI. The query ROI in these methods was not restricted in rectangle regions. Chen et al. (2020) proposed to represent the annotation regions by fusing patch-level features and encoding the region representation by supervised hashing for retrieval. However, the representation of an irregular region was obtained by the quantification of local features. The scale information of the region cannot be described. In the methods Jimenez-del Toro et al. (2017) and Zheng et al. (2018b), the composition of the images was considered by measuring the similarity between each pair of local features across two regions. Nevertheless, the adjacency relationship of different types of histological objects was not considered in these methods. Therefore, the structural similarity between tissue regions is difficult to recognize in the retrieval procedure.

To conquest the drawbacks in the present methods, we proposed to establish an end-to-end network based on the LA-Graphs to encode the regions into uniform indexes, where the local features, the adjacency relationships, and the location information are hopefully preserved in the indexes and reflected in the results of retrieval.

3. Method

The overview of the framework is illustrated in Fig. 1. The WSIs are first divided into patches and converted into an image feature tube with a pre-trained convolutional neural network (CNN) He et al. (2016); Huang et al. (2017). Then, sub-region graphs are generated based on the spatial relationships and feature similarities of patches. Moreover, the minimum distances of the patches to the border of the tissue are measured to identify the location of the graph in the WSI. Finally, the location-aware graphs (LA-Graphs) are constructed and fed into the designed LAGE-Net to obtain the binary indexes for retrieval. The method for LA-Graph construction and the LAGE-Net are the main components of the framework, which are elaborated in this section.

3.1. Location-aware graph construction

The flowchart to generate the graphs for a WSI is illustrated in Fig. 2. The patches in a WSI are clustered into sub-regions based on their CNN features. Then, the sub-regions are represented with graphs by regarding the patches as the graph nodes and the spatial adjacency as the graph edges.

Letting I_i represent the i-th patch in a WSI, the process of feature extraction can be described as

$$\mathbf{x}_i = \mathcal{F}_{CNN}(I_i),$$

where $\mathcal{F}_{CNN}$ represents a CNN feature extractor that takes an image patch as the input and outputs d_f-dimensional column vector.

Then, the hierarchical agglomerative clustering (HAC) algorithm Day and Edelsbrunner (1984) is employed to merge the patches in the WSI based on the CNN features. HAC is designed to merge a set of samples to an assigned number of clusters. In each iteration of HAC, the two most similar clusters under specific similarity measurement are merged. Specifically for a WSI, we propose regarding the patch features $\{\mathbf{x}_i\}_{i=1}^{m_s}$ as initial clusters and utilizing error sum of squares (EES) as the similarity measurement between clusters. Besides, an adjacency matrix $\mathbf{A}_s \in \{0,1\}^{m_s \times m_s}$ is generated to indicate the connectivity of m_s patches in the s-th WSI, where $a_{ij} = 1, a_{ij} \in \mathbf{A}_s$ indicates the i-th and the j-th patch are spatially 4-connected and $a_{ij} = 0$ otherwise. To ensure the merged sub-regions are spatially connected, only the pairs associated with $a_{ij} = 1$ are allowed to be merged in the iterations of HAC. Fig. 2b illustrates the merged sub-regions, where a colored area represents a sub-region.

Another important information considered in the graph is the global location of the sub-region in the WSI. The distance of the sub-region, especially the cancerous region, to the border of the tissue implies the information about the size of the tumor, the depth of tumor infiltration, etc., that is important indicators of tumor classification

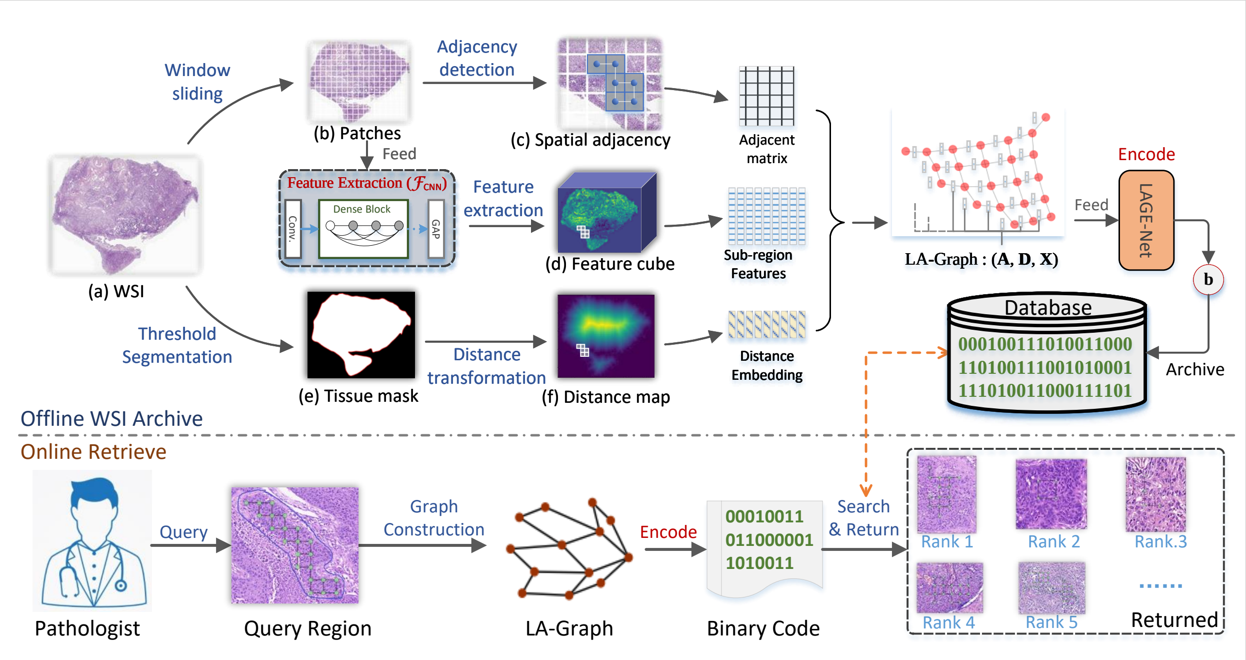


Figure 1: The proposed CBHIR framework. In the offline stage, the WSIs are first divided into patches following the sliding window paradigm and a CNN is trained based on the patch labels to extract image features. Then, a WSI is divided into sub-regions based on the features and the spatial adjacency of the patches. Next, A graph is constructed for each sub-region by regarding the patches within the sub-region as nodes and the spatial adjacency as edges. Additionally, the distances of the patches to the tissue border are considered in the graph. Finally, the LAGE-Net is trained based on the graphs for sub-region encoding and indexing. In the online retrieval stage, the region the pathologist queried is converted into a binary code using the trained models. The most relevant regions are retrieved by measuring the similarities between the query code and those in the database and finally returned to pathologists for diagnosis reference.

and grading. Motivated by this, we propose measuring the minimum distance of each patch to the border of the tissue and adding it to the tissue graph data. It makes the graph involve the tissue depth information, which is expected to benefit the subsequent encoding process. Specifically, we apply distance transformation to the tissue mask segmented from the WSI, as shown in Fig. 1(e-f) and record the minimum distance to the tissue border for each patch center. The border distance of the j-th patch is denoted by φ_j. For a uniform representation, φ_j is scaled to $\varphi_j \in [0, 1]$ by the minimum and maximum values in the dataset.

Finally, we construct the location-aware graph for each subregion, which can be represented as $G = (\mathbf{A}, \mathbf{X}, \phi)$, where $\mathbf{A} \in \mathbb{R}^{m\times m}$ is an adjacent matrix that defines the connectivity in G with m nodes, and $\mathbf{X} = (\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}}, ..., \mathbf{x}_m)^{\mathrm{T}} \in \mathbb{R}^{m\times d_f}$ denotes the node feature matrix assuming each node is represented as a d_f-dimensional vector, and $\phi = (\varphi_1, \varphi_2, ..., \varphi_m)$. For convenience, all the graphs in the s-th WSI are represented by set $\mathcal{G}_s = \{G_i | i = 1, 2, ..., n_s\}$, where n_s denotes the number of graphs in the s-th WSI. The set $\mathcal{G}_s$ covers the entire content of the WSI and thereby can be used to index the WSI.

In summary, the algorithm of the LA-Graph construction is arranged as Algorithm 1.

3.2. LAGE-Net for region encoding

It is challenging to encode the graph node attributes with local adjacency and global location information into a uniform representation. In this paper, we propose a location-aware graph encoding network (LAGE-Net) to achieve this task. The structure of LAGE-Net is presented in Fig. 3. The CNN features for a graph are firstly embedded by a linear transformation, then fed into the

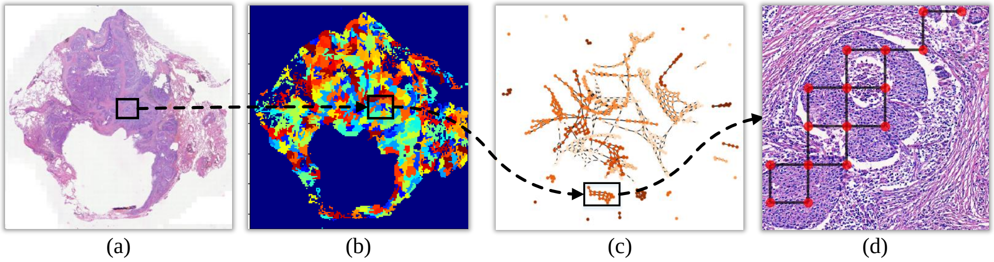


Figure 2: The flowchart of tissue graph generation, where (a) is a digital WSI, (b) illustrates the sub-regions clustered by Algorithm 1, (c) shows the graphs established on the sub-regions, and (d) jointly presents a graph and its corresponding region where the nodes are drawn on the centers of patches.

stacked blocks consisting of the LAGE module and feed-forward linear layers to obtain the graph representation. Finally, the graph representation is transferred into the binary code by a hash module. Meanwhile, layer normalization and residual connection are inserted (as shown in 3a). The main component of LAGE-Net is the LAGE module, which is elaborated as follows.

3.2.1. LAGE module

The proposed LAGE module is composed of graph convolution, self-attention and linear transformation operation. The flowchart of the module is illustrated in Fig. 3b.

1) Internal relationship encoding with graph convolution

The adjacency matrix **A** in a tissue graph describes the internal relationship of the graph nodes. The message-passing based on **A** is essential in the graph representation learning. Therefore, we firstly apply the GCN methodology proposed by Kipf and Welling (2016) to achieving the internal relationship encoding. Generally, a step of graph convolution can be formulated as

$$\mathbf{H}_{gc} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}_e\mathbf{W}_{gc}, \tag{1}$$

where $\mathbf{H}_{gc} \in \mathbb{R}^{m\times d_e}$ denotes the node embeddings after the l-th step of graph convolution, d_e denotes the dimension of the embeddings, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$[1], $\tilde{\mathbf{D}} = diag(\sum_j \tilde{\mathbf{A}}_{1j}, \sum_j \tilde{\mathbf{A}}_{2j}, ..., \sum_j \tilde{\mathbf{A}}_{nj})$, and $\mathbf{W}^{(l)} \in \mathbb{R}^{d_e\times d_e}$ is a trainable weight matrix. For simplification, we define $\bar{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}$ and rewrite the equation 1 as

$$\mathbf{H}_{gc} = \bar{\mathbf{A}}\mathbf{X}_e\mathbf{W}_{gc}, \tag{2}$$

Specifically, $\mathbf{X}_e$ represents the original embeddings of graph nodes, which in our method is defined as the linear transformation of the CNN features followed by layer normalization (LN) operation.

2) Global location encoding with self-attention

In this paper, the global location of the sub-regions in the WSI is proposed to be also important for diagnostically relevant retrieval. Motivated by the usage of the position embedding in the Transformer Vaswani et al. (2017), we determine to build global location embeddings for the graph nodes and merging the global location information into the graph representation. Specifically, we define the distance index $\bar{\varphi}_j = [\varphi_j \times N_{dist}]$ with N_{dist} controls the intervals of the distance indexing and is empirically set as 64 in the experiment. Then, we define

$$\mathbf{D} = (\mathbf{d}^{\mathrm{T}}_{\bar{\varphi}_1}, \mathbf{d}^{\mathrm{T}}_{\bar{\varphi}_2}, ...)^{\mathrm{T}} \in \mathbb{R}^{m\times d_e} \tag{3}$$

to be the global location embeddings that are indexed by $\{\bar{\varphi}_j\}^m_{j=1}$. $\mathbf{d}_{\bar{\varphi}_j}$ is generated following sinusoidal embedding formula in Transformer Vaswani et al. (2017). Then, **D** is merged to the original node embeddings $\mathbf{X}_e$ by the operation

$$\mathbf{X}_{sa} = \mathbf{X}_e + \mathbf{D}, \tag{4}$$

Here, $\mathbf{X}_{sa} \in \mathbb{R}^{m\times d_e}$ involves the information from both the image patterns and global locations of the patches. Next, we apply the self-attention mechanism to build relations between the location-aware representations. This proce-

[1] **E** denotes the unit matrix.

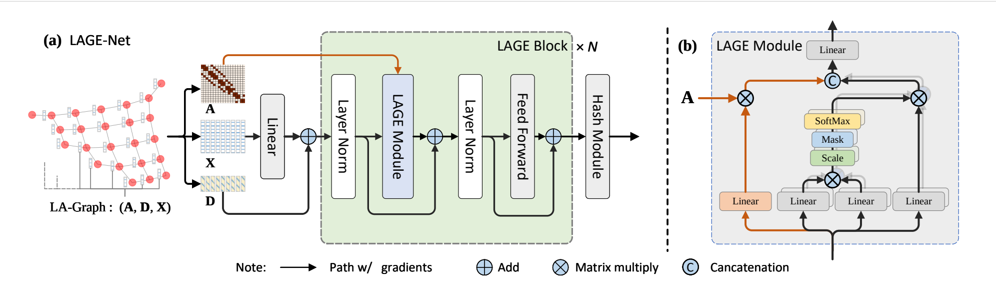


Figure 3: The location-aware graph encoding network (LAGE-Net) consists of a feature embedding layer, multiple stacked LAGE blocks, and a hash layer. It takes the location-aware graph as input and outputs a binary code that is used to index the graph for retrieval.

dure can be represented as

$$\mathbf{A}_{sa} = Softmax(\frac{\mathbf{X}_{sa}\mathbf{W}_q \cdot (\mathbf{X}_{sa}\mathbf{W}_k)^{\mathrm{T}}}{\sqrt{d_{att}}}), \tag{5}$$

$$\mathbf{H}_{sa} = \mathbf{A}_{sa}\mathbf{X}_{sa}\mathbf{W}_v, \tag{6}$$

where $\mathbf{W}_k, \mathbf{W}_q, \mathbf{W}_v \in \mathbb{R}^{d_e \times d_{att}}$ represents the weights for *Query*, *Key*, and *Value* branch of the self-attention module, respectively.

3) Information integration with linear transformation

Finally, the outputs of the graph convolution and self-attention are concatenated and then fed into a linear transformation layer, to integrate the patterns extracted from different aspects. The linear transformation layer is formulated as

$$\mathbf{H}_l = GELU([\mathbf{H}_{gc}; \mathbf{H}_{sa}] \cdot \mathbf{W}_l + \mathbf{b}_l), \tag{7}$$

where *GELU* represents the Gaussian error linear units function, $\mathbf{W}_l$ and $\mathbf{b}_l$ are the weights and bias.

It is notable that the graph convolution (Eq.2) and self-attention (Eq.6) share the same formulation. The main difference is that the message-passing matrix $\bar{\mathbf{A}}$ is generated from the natural adjacency relationship of graph nodes and fixed in the calculation, and $\mathbf{A}_{sa}$ is online generated based on the current state of each node regarding both the image content and the global location of the nodes. More generally, we extended the self-attention to the multi-head formulation to allow the LAGE module to observe more aspects of relationships behind the image content and the global location information.

3.2.2. Binary indexing with Hash function

In our method, the network is used to generate graph region indexes that are effective for data retrieval. To learn the representation of the graph for the hash function, a trainable token is concatenated to the initial graph embeddings referring to BERT and ViT, which can be formulated as

$$\mathbf{X}_e \leftarrow [\mathbf{x}_h^{\mathrm{T}}; \mathbf{X}_e^{\mathrm{T}}]^{\mathrm{T}}. \tag{8}$$

The learnable token is regarded as another node that is connected with all other nodes in the graph embedding calculation, for which the adjacency matrix $\mathbf{A}$ is correspondingly modified. Meanwhile, the token participates in all the self-attention computations.

To ensure the framework is applicable to the practical large-scale pathological database, we built a head layer with hash functions on the output of the last MLP. Supposing $\mathbf{z}_h =\in \mathbb{R}^{d_e}$ represents the final MLP output of learnable token, the hashing function is defined as

$$\mathbf{y}_h = \tanh(\mathbf{z}_h\mathbf{W}_h + \mathbf{b}_h), \tag{9}$$

where $\mathbf{W}_h \in \mathbb{R}^{d_e \times d_h}$ and $\mathbf{b}_h \in \mathbb{R}^{d_h}$ are the weights and bias for the hash functions, d_h is the dimension of binary codes, and tanh represents the hyperbolic tangent function. $\mathbf{y}_h \in (-1, 1)^{d_h}$ is the network outputs that can be simply converted into binary codes by equation $\mathbf{h} = sign(\mathbf{y}_h) \in \{-1, 1\}^{d_h}$. Letting $\mathbf{Y} \in (-1, 1)^{N_g \times d_h}$ denote the binary-like codes of N_g graphs, the objective function to minimize for training the LAGE-Net is defined as

$$J = \frac{1}{N_g} \|\frac{1}{d_h}\mathbf{Y}\mathbf{Y}^{\mathrm{T}} - \mathbf{C}\|_F^2 + \lambda\|\mathbf{W}_h^{\mathrm{T}}\mathbf{W}_h - \mathbf{E}\|_F^2 \tag{10}$$

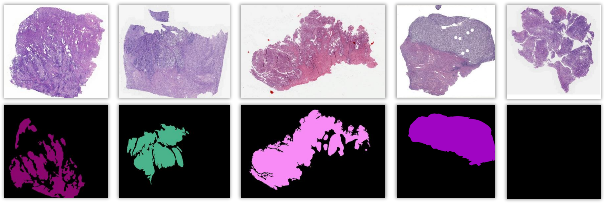
Algorithm 1: The algorithm of tissue graph construction.

```
Input:
$m_s \leftarrow$ The number of patches in the $s$-th WSI;
$\{\mathbf{x}_i | i = 1, 2, ..., m_s\} \leftarrow$ The feature vectors of patches;
$\mathbf{A}_s \in \{0, 1\}^{m_s \times m_s} \leftarrow$ The adjacency matrix of patches;
$\mathbf{\Phi}_s \in [0, 1]^{m_s \times m_s} \leftarrow$ The distance transformation matrix
for patch centers;
$\hat{g}_s \leftarrow$ The target number of graphs ($\hat{g}_s \le m_s$);
Output: $\mathcal{G}_s$
for $i = 1$ to $m_s$ do
    $C_i \leftarrow \{\mathbf{x}_i\}$;
end
$C \leftarrow \{C_i | i = 1, 2, ..., m_s\}$;
$g_s \leftarrow m_s$;
while $g_s > \hat{g}_s$ do
    $\mathcal{T} \leftarrow \emptyset$;
    for $(C_i, C_j)$ in
        $\{(C_i, C_j) | \exists \mathbf{x}_p \in C_i, \exists \mathbf{x}_q \in C_j, i \neq j, s.t. a_{pq} = 1\}$
     do
        $d_{ij} \leftarrow EES(C_i \cup C_j)$;
        $\mathcal{T} \leftarrow \mathcal{T} \cup \{d_{ij}\}$;
    end
    index $(p, q) \leftarrow \arg\min_{(i,j)}(\mathcal{T})$;
    $C_p \leftarrow C_p \cup C_q$;
    $C \leftarrow C \setminus C_q$;
    $g_s \leftarrow g_s - 1$;
end
$\mathcal{G}_s \leftarrow \emptyset$;
for $C_i$ in $C$ do
    $\mathbf{X}_i \leftarrow (\mathbf{x}_1, ..., \mathbf{x}_j, ..., \mathbf{x}_{|C_i|})_{\mathbf{x}_j \in C_i}$;
    $\mathbf{A}_i \leftarrow$ Seek $\mathbf{A}_s$ for the adjacent relationship of
      patches corresponding to $\mathbf{X}_i$;
    $\phi_i \leftarrow$ Seek $\mathbf{\Phi}_s$ for the distance values of patches
      corresponding to $\mathbf{X}_i$;
    $G_i \leftarrow (\mathbf{X}_i, \mathbf{A}_i, \phi_i)$;
    $\mathcal{G}_s \leftarrow \mathcal{G}_s \cup \{G_i\}$;
end
return $\mathcal{G}_s$;
```

where $\mathbf{C} \in \{-1, 1\}^{N_g \times N_g}$ is the pair-wise label matrix in which $c_{ij} = 1$ represents the i-th graph and the j-th graph are relevant and $c_{ij} = -1$ otherwise. λ is the weight coefficient of the orthogonal regularization and is empirically set to 0.01 in this paper. Finally, the proposed LAGE-Net is trained end-to-end from the input graph with CNN-features to the output $\mathbf{Y}$. For simplification, the LAGE-Net is represented as $\mathbf{h} = \mathcal{F}_{LAGE-Net}(G)$.

Table 1: Data allocation of the *Endometrium-2K* dataset.

Type Name	WDEA	MDEA	LDEA	SEIC	Normal	Total
Number	813	821	277	152	587	2650



(a) WDEA (b) MDEA (c) LDEA (d) SEIC (e) Normal

Figure 4: Instances in the endometrial WSI dataset, where (a) is Well-differentiated Endometrioid adenocarcinoma (WDEA), (b) is Moderately differentiated Endometrioid adenocarcinoma (MDEA), (c) is Low differentiated Endometrioid adenocarcinoma (LDEA), (d) is Serous endometrial intraepithelial carcinoma (SEIC), (e) is Normal and the ground-truth for the cancerous regions are provided on the second row.

3.3. *Efficient retrieval with binary codes*

For each WSI in the retrieval database, a set of binary codes that represent the graphs in the WSI can be obtained using the trained feature extraction model $\mathcal{F}_{CNN}$ and hash model $\mathcal{F}_{LAGE-Net}$. When retrieving, the region the pathologist queries is divided into patches and converted into binary codes using the same model. Then, the similarities between the query code and those in the database are measured using Hamming distance referring to Zhang et al. (2015b); Zhang and Metaxas (2016); Shi et al. (2018); Zheng et al. (2019). After ranking the similarities, the most relevant regions are retrieved and finally returned to the pathologist.

4. Experiments

4.1. *Experimental setting*

The experiments were mainly conducted on a public dataset and an in-house dataset of histopathological whole slide images. The profiles of the datasets are provided as follows.

- *Endometrium-2K* contains 2650 WSIs of endometrium histopathology from 2650 patients collected by Tianjin Fifth Central Hospital of China. These WSIs were scanned under a lens of 20× and categorized to 5 types of endometrial pathology, including Well-differentiated Endometrioid adenocarcinoma (WDEA), Moderately differentiated Endometrioid adenocarcinoma (MDEA), Low differentiated Endometrioid adenocarcinoma (LDEA), Serous endometrial intraepithelial carcinoma (SEIC), and Normal. All the cancerous regions were annotated by expert pathologists. The WSI instances are shown in Fig. 4, and data allocation is given in Table 1. In the experiment, 30% WSIs were randomly selected as the testing dataset (to generate query regions), and the remainders were used to train the retrieval models and establish the retrieval database.

- *ACDC-LungHP* (Li et al. (2021)) contains 150 WSIs within lung cancer regions annotated by pathologists.[2] In the evaluation, 30 WSIs were randomly selected as the testing dataset, and the remainders were used to train the retrieval models and establish the retrieval database.

All the WSIs were divided into square patches under lenses of 20× following the sliding window paradigm. The step of the window was set half of the length of the patch side. DenseNet Huang et al. (2017) was employed as the CNN structure to extract patch features. The global average pooling (GAP) layer of the DenseNet structure was used as the feature extractor. The patch size was set 224×224 to fit the input of DenseNet. Graphs were constructed for each WSI using the algorithm provided in Algorithm 1. For convenience, the graphs for establishing the retrieval database are represented as a set $\mathcal{D}$ and the query graphs are correspondingly represented as $\mathcal{Q}$.

We first conducted experiments to determine hyperparameters of models involved in our method on the training set. Then, the retrieval performance was evaluated on the testing set and compared with the state-of-the-art methods.

In the evaluation, the graphs that contain more than 10% cancerous pixels referring to the pathologists' annotations were defined as *Cancerous Graph*, the graphs containing none cancerous pixels were regarded as *Cancer-free Graph* and the remainders were not counted in the evaluation. For the *Endometrium-2K* dataset, the label of a *Cancerous Graph* is set the same as the WSI to which graph belongs. Correspondingly, only the returned graphs that share the same label with the query one were considered as relevant in both the training and evaluation stage. The average precision of retrieval $P@k$ for top-k-returned regions and the mean average precision mAP are used as the metrics. Letting $r_{ik} = 1$ indicates that the k-th returned result shares the same label with the i-th query graph and $r_{ik} = 0$ otherwise, $P@k$ and mAP can be defined by equations

$$P@k = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} p_i(k),$$

$$mAP = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \frac{\sum_{k=1}^{|\mathcal{D}|} p_i(k) \cdot r_{ik}}{\sum_{k=1}^{|\mathcal{D}|} r_{ik}},$$

where $|\cdot|$ denotes the length of the set and

$$p_i(k) = \frac{\sum_{j=1}^{k} r_{ij}}{k}$$

represents the retrieval precision of the top-k returned results for the i-th query instance. The higher the metrics, the better the retrieval performance.

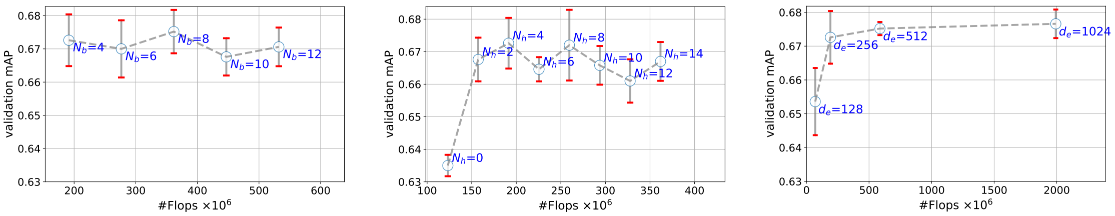
All the experiments were conducted in python with pytorch and run on a computer cluster with 10 available GPUs of Nvidia Geforce 2080Ti.

The Adam optimizer was employed to train the model. The initial learning rate for training the LAGE-Net is 3×10^{-4}

4.2. The structure of feature extraction CNN

The CNN used for patch feature extraction was trained via a subtype classification task. Specifically, patches in size of 224×224 pixels were sampled from the training WSIs. The patches containing above 75% percentage of cancerous pixels were labeled as positive samples, the

[2] The dataset is accessible at https://acdc-lunghp.grand-challenge.org. Since the annotations of the testing part of the data set are not yet published, only the 150 training WSIs of the data were used in this paper.



(a) Number of LAGE blocks N_b

(b) Number of heads N_h

(c) Dimension of embedding d_e

Figure 5: The $mAP - \#FLOPs$ curves as functions of the hyper-parameters of the LAGE-Net, where the average mAP of the 5-fold cross-validation is presented by each data point and the standard variance of the 5 trials is drawn with red bar.

patches involving none cancerous pixels were regarded as negative samples, and the other patches were not used.

The depth of DenseNet was tuned within the training set in the scope suggested in Huang et al. (2017). The best depth was determined according to the mean classification error of five-fold cross-validation on within the training slides. Finally, the depth of the CNN was determined as 121 according to the validation error, for which the dimension of the patch features (d_f) is 1024.

4.3. The structure of LAGE-Net

The body of the LAGE-Net is stacked by multiple LAGE blocks and each block is composited of a graph convolution head and multiple self-attention heads. The number of blocks N_b, the number of attention heads N_h and the dimension of the embeddings d_e determine the computational complexity of the encoding process and the performance of retrieval. These hyper-parameters were tuned over a wide range and selected based on the best mAP obtained through five-fold cross-validation in the training set. Note that the division of the data for the cross-validation here was the same with that in the CNN training stage.

The mAP and the number of floating-point operations (#FLOPs) as functions of hyper-parameter settings for the *Endometrium-2K* dataset are presented in Fig. 5. The other hyper-parameters were set fixed when one hyper-parameter was tuned.

1) The number of blocks N_b determines the depth of the LAGE-Net. A larger N_b helps extract higher level of information from tissue graphs but would also increase the risk of over-fitting. N_b was tuned from 4 to 12 with a step of 2 and the results (as shown in Fig. 5a) indicate $N_b = 8$ is optimum for the dataset.

2) The number of heads N_h determines the width of the network. More heads for self-attention enable the network to build node relations in more aspects and therefore generate better graph representations for retrieval. N_h was tuned from 0 to 14 with a step of 2. Note that $N_h = 0$ means self-attention operations along with the global location information are entirely omitted and the graph representation learning is only based on the local adjacency information. The results in Fig. 5b show that $N_h = 4$ and $N_h = 8$ achieved comparable retrieval performance. Finally, we decided to use 4 self-attention heads in each LAGE module in pursuit of lower computation.

3) The dimension of embedding d_e affects the total floating-point operations of the linear transformation and the multi-head self-attention after N_b and N_h are decided. The computational complexity of LAGE-Net is in direct proportion to $O(N_b N_h d_e^2)$ when the input graph is fixed. In this experiment, we tune d_e from 128 to 1204. As shown in Fig. 5c, the mAP steady increases as d_e enlarges, but the computational amount suffers from quadratic augmentation. To ensure the retrieval can be quickly completed and meanwhile maintain a high retrieval precision, we set $d_e = 512$ in the following experiment.

4.4. Ablation study

The patch features $\mathbf{X}$ are the essential information that is required to be encoded by the LAGE-Net. Besides, The internal adjacency information described by $\mathbf{A}$ and the global location information in the distance embedding $\mathbf{D}$ are also proposed to be important in this paper. The former is mainly modeled by the graph convolution in the LAGE module and the latter is modeled by the self-attention operation along with the patch features. We conducted ablation experiments to certify the effectiveness of the two factors. The ablation models are as follows.

- *LAGE-Net w/o dist*. The distance embedding $\mathbf{d}_{\bar{\varphi}_j}$ is replaced with a common positional embedding $\mathbf{d}_j$ that is associated to the patch index j. As a result, the global location information of the graph is removed.
- *LAGE-Net w/o adj*. The adjacency matrix in the graph convolution path is set to $\mathbf{A} = \mathbf{0}$. Consequently, the internal structure constraint is not considered in the graph encoding process.
- *LAGE-Net w/o dist* & *adj*. Both the above two types of ablation are performed.

The retrieval performance is compared in Table 2. The average precision and mAP apparently decreased as the internal adjacency or the global location information of the graph was discarded. The experiment has verified the effectiveness of the two components. Especially, LAGE-Net w/o dist suffers a 3.6% drop in P@5 and 3.3% drop in mAP when the distance embedding is not considered. It indicates that the depth of a region to the border of the tissue matters to the recognition of tumor types. Moreover, we visualized the graph representation $\{\mathbf{z}_h\}$ for the training graphs (the retrieval database) in the 2-dimensional space with t-SNE Maaten and Hinton (2008). Figures 6(a-b) illustrates the averaged border distance for each graph. It is obvious that, in Fig. 6b, the graphs sharing a similar depth to the tissue border tends to cluster in the feature space. That is one of the reasons that LAGE-Net is significantly superior to LAGE-Net w/o dist. In contrast, the clustering phenotype became less obvious when the distance embedding was removed from the encoding. It also demonstrates the effectiveness of the proposed distance embedding approach.

Table 2: Results for the ablation study, where the metrics on the test set are compared and the best values are shown in bold.

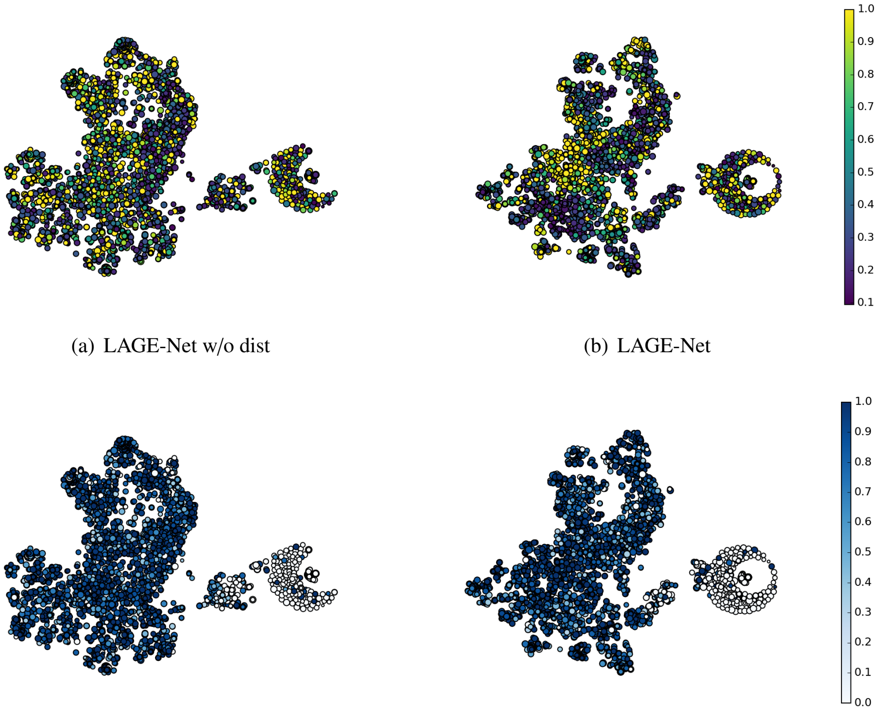
Networks	P@5	P@20	mAP
LAGE-Net w/o dist & adj	0.561	0.553	0.626
LAGE-Net w/o dist	0.561	0.559	0.634
LAGE-Net w/o adj	0.571	0.563	0.652
LAGE-Net	**0.587**	**0.583**	**0.667**

4.5. Comparison with the state-of-the-art

The proposed method is compared with 6 state-of-the-art-methods proposed by Ma et al. (2018); Jimenez-del Toro et al. (2017); Zheng et al. (2018b, 2019); Yan et al. (2020); Dosovitskiy et al. (2021). These methods can be categorized into two groups. The first group designs distance between the sets of features of two sub-regions to measure their similarity for retrieval. This group includes the following four methods:

- Jimenez-del Toro et al. (2017). The retrieval is achieved based on both the WSIs and the text information of the cases. In the experiment, only the part for WSIs retrieval was implemented for the meta-information of the datasets is not available. Specifically, the distances between all pairs of patch features across two sub-regions were calculated and the mean value of the distances was used as the similarity measurement.
- Zheng et al. (2018b). The patches in the sub-regions are encoded into binary codes. When retrieving, a set of proposal graphs is first retrieved through table lookup operation based on patch codes. Then, the distances between the query graph and the proposal graphs are calculated under specific similarity measurement and then the most similar graphs are returned.
- Ma et al. (2018). The features for a sub-region are quantified through a max-pooling operation. Then, the obtained representation is converted into binary codes based on latent Dirichlet allocation (LDA) Blei et al. (2003) followed by supervised hashing. Finally, the similarity between two graphs is computed based on binary codes.

The second group trains deep learning models that take sub-region features as input and outputs uniform representations for the sub-regions. Then, the retrieval can be completed by measuring the distance of these uniform representations. We compare three typical methods from this group. Note that the proposed method belongs to the second group.



(a) LAGE-Net w/o dist

(b) LAGE-Net

(c) LAGE-Net w/o dist

(d) LAGE-Net

Figure 6: The 2-dimensional visualization of the graph representation output by the last MLP ($\mathbf{z}_h$) for the *Endometrium-2K* retrieval database, where a dot represents a graph, the color of the dots in (a) and (b) presents the averaged normalized distance for each graph to the border of the tissue (ϕ_j), and the color of the dots in (c) and (d) indicates the ratio of tumor occupation in the graph referring to the color bar on the right of the figure. (Only a part of the database graphs are randomly selected and plotted for the purpose of clear display.)

- Zheng et al. (2019). Graphs are constructed to describe the sub-regions in the WSIs and fed into a GCN with diffpool module Ying et al. (2018) to extract the graph representation. A hash layer is built to the end of the GCN to convert the representation into binary codes.

- Yan et al. (2020). The patch features are fed into a two multi-layer bi-directional LSTM to exchange the contextual information. Then the outputs of the LSTM are merged by an average pooling layer for similarity measurement.[3]

- Dosovitskiy et al. (2021). The vision transform (ViT) model is applied to encode the graph and the classification head of ViT is replaced with a hash layer to generate the binary codes for retrieval.

For a fair comparison, the feature extractors (or backbones) of the compared methods were the same DenseNet-121 structure.

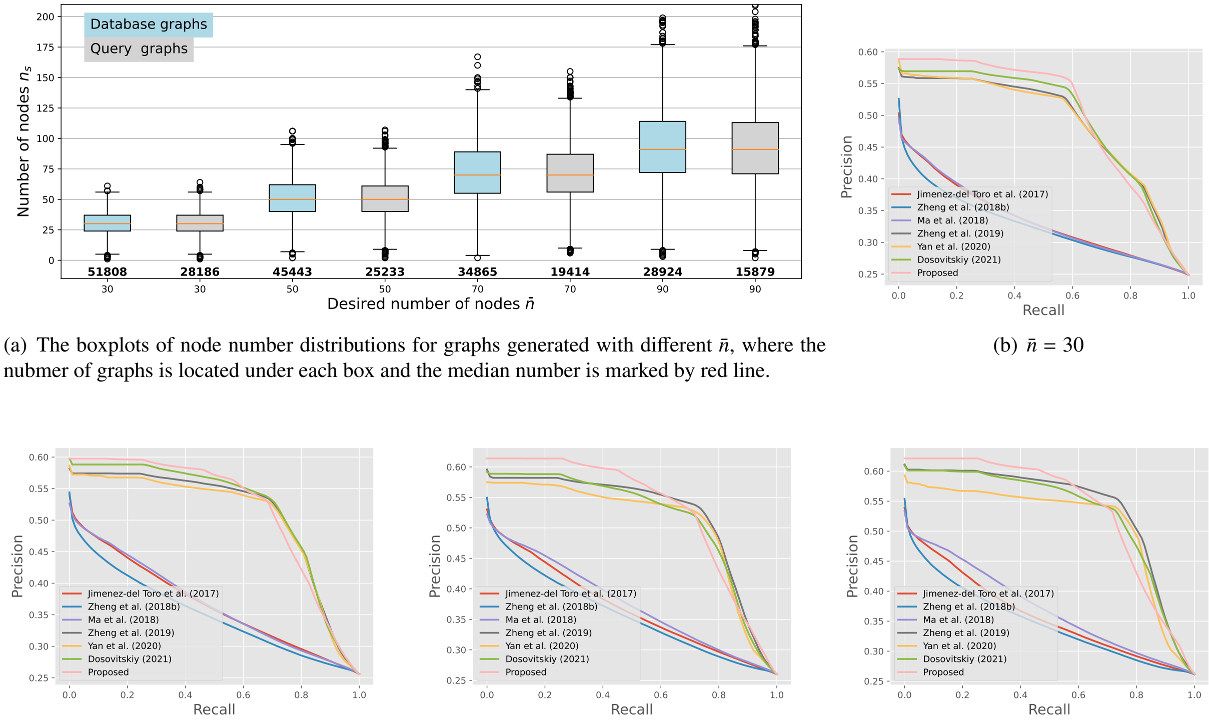
4.5.1. Comparison of retrieval precision

We conducted experiments for sub-regions in different scales to evaluate the retrieval performance of the com-

[3] The outputs of the LSTM are concatenated to generate the regional representation in the original method. For that the number of the features for a sub-region in our experiment was not consistent, we used a average pooling layer as a substitute for the concatenation.

Table 3: Retrieval performance for the state-of-the-art methods on the *Endometrium-2K* dataset, where the results for different size allocations of graphs (determined by $\bar{n}$) are compared.

Methods	$\bar{n}$ = 30 P@5 / P@50 / mAP	$\bar{n}$ = 50 P@5 / P@50 / mAP	$\bar{n}$ = 70 P@5 / P@50 / mAP	$\bar{n}$ = 90 P@5 / P@50 / mAP	Retrieval Complexity
Jimenez-del Toro et al. (2017)	0.492 / 0.472 / 0.362	0.532 / 0.514 / 0.392	0.527 / 0.512 / 0.393	0.526 / 0.509 / 0.383	$O(a^2b)$
Zheng et al. (2018b)	0.502 / 0.469 / 0.337	0.530 / 0.504 / 0.364	0.541 / 0.513 / 0.372	0.542 / 0.507 / 0.359	$O(a^2b)$
Ma et al. (2018)	0.486 / 0.466 / 0.361	0.526 / 0.510 / 0.392	0.522 / 0.507 / 0.401	0.522 / 0.505 / 0.398	$O(b)$
Zheng et al. (2019)	0.563 / 0.558 / 0.594	0.581 / 0.561 / 0.618	0.573 / 0.584 / 0.633	0.603 / 0.599 / 0.646	$O(b)$
Yan et al. (2020)	0.568 / 0.567 / 0.594	0.574 / 0.564 / 0.597	0.540 / 0.569 / 0.595	0.589 / 0.589 / 0.593	$O(b)$
Dosovitskiy et al. (2021)	0.569 / 0.569 / 0.633	0.587 / 0.587 / 0.645	0.590 / 0.588 / 0.653	0.604 / 0.603 / 0.665	$O(b)$
Proposed	**0.588 / 0.581 / 0.667**	**0.593 / 0.593 / 0.673**	**0.611 / 0.610 / 0.686**	**0.619 / 0.617 / 0.692**	$O(b)$



(a) The boxplots of node number distributions for graphs generated with different $\bar{n}$, where the nubmer of graphs is located under each box and the median number is marked by red line.

(b) $\bar{n}$ = 30

(c) $\bar{n}$ = 50

(d) $\bar{n}$ = 70

(e) $\bar{n}$ = 90

Figure 7: Comparison of interpolated precision-recall curves of different retrieval methods on the *Endometrium-2K* dataset, where (a) provides the distributions of number of graph nodes obtained with different $\bar{n}$, and (b-e) present the interpolated precision-recall curves for different settings of $\bar{n}$, respectively.

pared method. Specifically, $\bar{n}$ was set from 30 to 90 with a step of 20 and the retrieval database and the number of clusters (i.e. the target number of graphs $\hat{g}_s$ for each WSI is determined by equation $\hat{g}_s = [m_s/\bar{n}]$. The graphs for each setting of $\bar{n}$ were obtained based on Algorithm 1. The allocation of graph node numbers is visualized by boxplots in Fig. 7. The experimental results are summarized in Table 3. Correspondingly, the interpolated precision-recall curves are illustrated in Fig. 7 (b-e).

Overall, the proposed method has achieved the best retrieval performance in the quantitative evaluation. The retrieval methods in the first group, including Jimenez-del Toro et al. (2017); Ma et al. (2018); Zheng et al. (2018b), depend on the similarity measurement of patch features. In these methods, the patch features were regarded as equally informative in the mean average and max-pooling operations. Both the interaction and the location information of the patches were not considered in the sub-region encoding process. These issues make it challenging to identify the subtle patterns in different subtypes of tumors, resulting in a gap in P@5 about 6.1% – 11.3% to the methods in the second group.

The methods Zheng et al. (2019) and Yan et al. (2020) in the second group have modeled the internal adjacency information among sub-region patches. The former applied the adjacency matrix to connecting features in 2D planar space and the latter constructed a sequence to describe the 1D adjacency information. Then, the graph neural networks and recurrent neural networks were trained end-to-end based on the region labels to generate uniform region representations. The end-to-end training strategy delivered a high mean average precision in the retrieval evaluation, as shown in Fig. 7(b-2). Meanwhile, the precision for the top-returned was significantly improved. However, the gap of mAP for Yan et al. (2020) to the other methods in the second group gradually enlarges as the size of the sub-region increases. The main reason is that the adjacency information is modeled by RNN. The communication of the patch features requires traversing the entire sequence and therefore is weakened when the sequence lengthens.

ViT Dosovitskiy et al. (2021) achieved comparable retrieval performance with our method. The superiority benefits from the self-attention mechanism, which enables a weighted communication among patch features during the encoding process. The essential difference of ViT from our method is that ViT utilizes trainable embedding indexed by the tensor positions rather than the border distances. However, the semantic for a tensor position is not fixed for the sub-region because the object in the histopathology image is non-rigid and non-directional. In this case, the trainable embedding could not find consistent meaning for a certain position and therefore could not be beneficial to the region encoding. In contrast, the proposed LAGE-Net equips the explicit embedding that is indexed by the distance of the patch to the tissue border. This difference brings an improvement of 1.9% – 2.1% in P@5 and 2.7% – 3.4% in mAP to our method.

4.5.2. Comparison of computational complexity

Efficiency is equally important for CBHIR system. In the online retrieval stage, the computational complexity mainly derives from the strategy of retrieval, which is relevant to the pixel size of query region a, the scale of the database b. The O notation for the compared methods are given in Table 3.

In our method, the retrieval is completed by measuring Hamming distances, which is irrelevant to the size of the query region after the graph encoding. Therefore, the complexity is $O(b)$. The average time for querying the database containing 51,808 graphs is 0.752 ms in our experimental environment. Moreover, benefiting from the binary encoding, the similarity measurement is time-saving than those based on float-type high-dimensional features (e.g., Jimenez-del Toro et al. (2017)). When the order of magnitudes of WSI in the database increases, a hash table can be pre-established. Then, the retrieval can be easily achieved by a table-lookup operation, for which the complexity of retrieval is potentially reduced to $O(1)$.

4.6. Comparison on ACDC-LungHP dataset

The same evaluations were completed on the ACDC-LungHP dataset. The hyper-parameters of the LAGE-Net were tuned within the 120 training WSIs and were finally determined as $(N_b, H_h, d_e) = (4, 4, 512)$. Then the training WSIs were encoded to construct the retrieval database. The 30 testing WSIs were used to generate the query graphs. The metrics of retrieval for different settings of $\bar{n}$ are compared in Table 4. As for the dataset only provides binary annotation, the training and evaluation were completed with binary labels, i.e., *Tumor vs. Normal*. The results have shown that all the compared methods achieved applicable retrieval performance. The P@5 is better than 0.779, and the mAP is above 0.701 for different sizes of query regions. Overall, the proposed retrieval framework with the LAGE-Net achieved the best performance. The experiment results are consistent with those obtained on the Endometrial-2K dataset.

4.7. Visualization

To further validate the qualitative performance of the proposed retrieval framework, we drew the graph struc-

Table 4: Retrieval performance for the state-of-the-art methods on the ACDC-LungHP dataset, where the results for different size allocations of graphs (determined by $\bar{n}$) are compared.

Methods	$\bar{n}$ = 30	$\bar{n}$ = 40	$\bar{n}$ = 50	$\bar{n}$ = 60	$\bar{n}$ = 70	$\bar{n}$ = 80	$\bar{n}$ = 90
	P@50 / mAP	*P*@50 / mAP	*P*@50 / mAP	*P*@50 / mAP	*P*@50 / mAP	*P*@50 / mAP	*P*@50 / mAP
Jimenez-del Toro et al. (2017)	0.779/ 0.708	0.777/ 0.709	0.772/ 0.709	0.779/ 0.708	0.765/ 0.705	0.786/ 0.707	0.782/ 0.707
Zheng et al. (2018b)	0.797/ 0.702	0.788/ 0.704	0.789/ 0.703	0.794/ 0.703	0.780/ 0.701	0.790/ 0.703	0.793/ 0.704
Ma et al. (2018)	0.783/ 0.715	0.783/ 0.719	0.779/ 0.719	0.788/ 0.718	0.783/ 0.717	0.789/ 0.718	0.786/ 0.718
Zheng et al. (2019)	0.801/ 0.862	0.811/ 0.865	0.840/ 0.867	0.831/ 0.872	0.797/ 0.857	**0.845**/ **0.884**	0.858/ 0.881
Yan et al. (2020)	0.796/ 0.803	0.824/ 0.841	0.805/ 0.813	0.818/ 0.835	**0.821**/ 0.816	0.793/ 0.816	0.832/ 0.844
Dosovitskiy et al. (2021)	0.815/ 0.861	0.838/ 0.880	0.843/ 0.886	0.829/ 0.875	0.815/ 0.864	0.815/ 0.853	**0.887**/ 0.885
Proposed	**0.819**/ **0.869**	**0.860**/ **0.899**	**0.863**/ **0.901**	**0.848**/ **0.885**	0.820/ **0.868**	0.833/ 0.866	0.884/ **0.897**

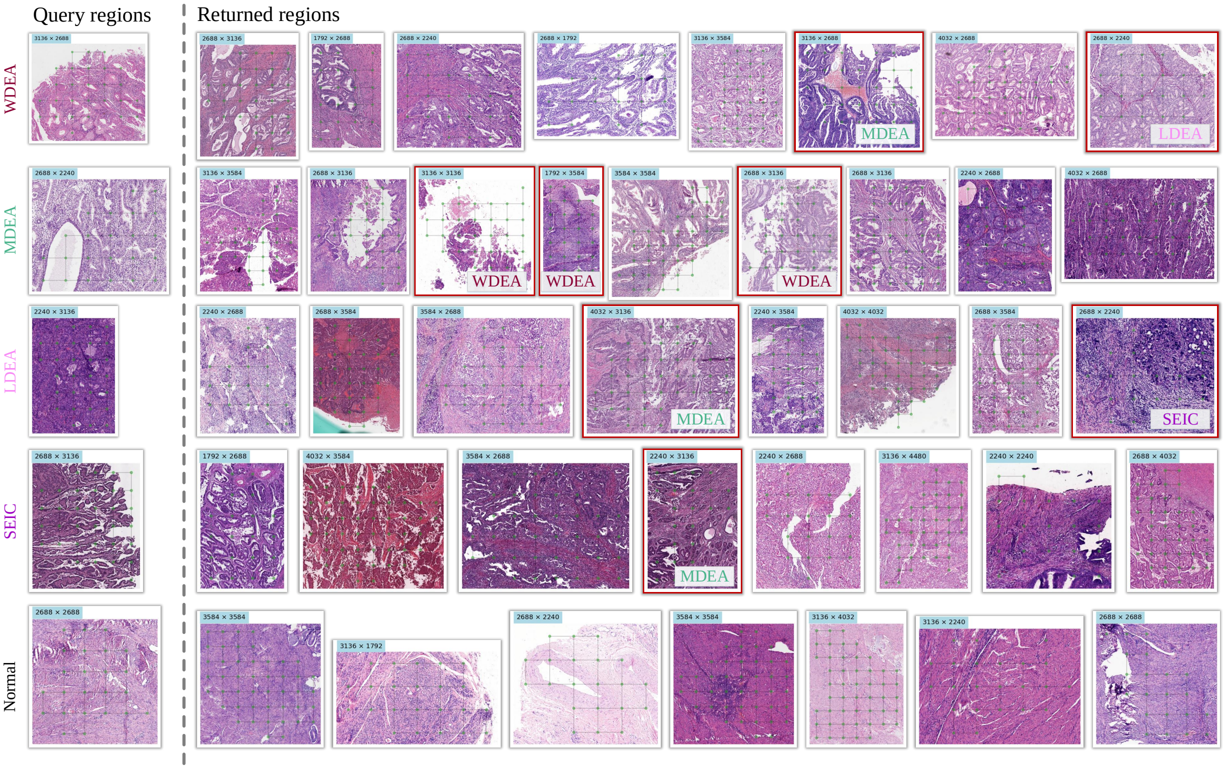


Figure 8: Visualization of the retrieval performance of the proposed method on the *Endometrium-2K* dataset, where the first column provides the 5 query regions in different type of lesions, the top-returned regions from the retrieval are ranked on the right, the irrelevant return regions (has different labels with the query graph) are framed in red and the pixel size of the regions are located on the leftop of the images. Please check the supplemental material for the high resolution version of the figure.

tures on the retrieved regions. The joint visualization of WSI regions and graphs for the retrieval instances in the *Endometrial-2K* dataset is provided in Fig. 8. It shows that the relevant regions in various shape and size for the query region are returned from the WSIs. It means that the LAGE-Net has learned the representations to identify the WDEA, MDEA, LDEA, and SEIC in endometrial histopathology.

4.8. Discussion

We have tried to use a trainable distance embedding indexed by $\bar{\varphi}_j$ as a substitute of the sine-cosine embedding in the LAGE module but observed a slight decrease in retrieval precision. Therefore, we finally applied this constant embedding strategy.

The parameter $\bar{n}$ affects the average size of the sub-regions that can be retrieved by the system. $\bar{n}$ is considered more of a *Control value* than a *Hyper-parameter* to be optimized in this framework. In this case, the proposed method was expected to be stable to $\bar{n}$. Higher values of $\bar{n}$ generate larger sub-regions. Larger regions contain more contextual information, which helps the model better identify the category of the sub-regions. That was the main reason that larger $\bar{n}$ value delivered better performance. Because the purpose of this study was to develop framework for fine-grained retrieval of sub-regions, we limited the value of $\bar{n}$ to 90 in the main experiment. We also tested the retrieval framework for the settings $\bar{n} = 180$ and $\bar{n} = 360$, and got a mAP of 0.703 and a mAP of 0.711 in the *Endometrium-2K* dataset, respectively, which were better than those obtained with $\bar{n} \leq 90$. However, the retrieval database were occupied by large WSI regions. Simultaneously, the computational amount of the self-attention module in LAGE-Net quadratically grows as the enlarge of $\bar{n}$, which increases difficulties to the training and inference of LAGE-Net.

Generally, the source of a query region, i.e. the organ a region comes from, is available in the digital pathology system, and the retrieval is usually performed within the WSIs from the same organ. Therefore, we did not mix the histopathology WSIs from different organs to establish the retrieval database. The proposed global location embeddings used to describe the location of the patches were calculated in the same resolution and then scaled based on the minimum and maximum distance values of the database. It ensures the same embedding represents consistent semantics. And when the retrieval system needs to be generalized to retrieval database containing WSIs from different organs or even in different resolutions, the distances to the border should be properly scaled to make sure the same embedding represents the equivalent actual distance.

The models in the proposed framework were trained based on supervised learning, which depends on the manual annotations of pathologists. Theoretically, the CNN and LAGE-Net are potentially trained based on the methodology of unsupervised learning, especially the contrast representation learning He et al. (2020); Grill et al. (2020); Chen et al. (2021). Therefore, one of the future works will focus on deploying the proposed framework without pathologists' manual annotations.

The encoding of regions in the proposed framework can be divided into three separate stages: feature extraction, graph construction, and graph encoding. Another future work will focus on combining the three stages into an integrated model that can be trained end-to-end and can simultaneously predict the representative regions in the WSI and encode these regions to establish the retrieval database.

5. Conclusions

In this paper, we propose a novel histopathological image retrieval framework for a large-scale WSI database based on location-aware graphs and deep hashing techniques. The sub-regions in the WSI are represented as graphs within image features, internal connection information and WSI location information. The graphs are encoded by the designed LAGE-Net and archived with binary codes for hash retrieval. The experimental results on an in-house endometrium dataset and a public lung dataset have demonstrated that the proposed method achieves state-of-the-art retrieval performance. The LAGE-Net is scalable to size and shape variations of query regions and can effectively retrieve relevant regions that contain similar content and structure of the tissue. It allows pathologists to create query regions by free-curves on the digital pathology platform. Benefited from hashing structure, the retrieval process is completed based on hamming distance, which is very time-saving. One future work is to build an integrated model for representative generation and indexing of whole slide images. Another future work will focus on developing unsupervised and weakly supervised retrieval frameworks.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China [grant number

61901018, 61771031, 61906058, and 61471016], partly supported by the China Postdoctoral Science Foundation [grant number 2019M650446], partly supported by Tianjin Science and Technology Major Project [grant number 18ZXZNSY00260], partly supported by the Anhui Provincial Natural Science Foundation [grant number 1908085MF210], partly supported by the Fundamental Research Funds for the Central Universities of China [grant number JZ2020YYPY0093], and partly supported by the 111 Project [grant number B13003].

References

Bejnordi, B.E., Balkenhol, M., Litjens, G., Holland, R., Bult, P., Karssemeijer, N., van der Laak, J.A., 2016. Automated detection of dcis in whole-slide h&e stained breast histopathology images. IEEE Transactions on Medical Imaging 35, 2141–2150. doi:10.1109/tmi.2016.2550620.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 318, 2199–2210.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research 3, 993–1022.

Caicedo, J.C., Gonzalez, F.A., Romero, E., 2008. A semantic content-based retrieval method for histopathology images, in: Asia Information Retrieval Conference on Information Retrieval Technology, pp. 51–60.

Caicedo, J.C., Izquierdo, E., 2010. Combining low-level features for improved classification and retrieval of histology images. Ibai Publishing 2, 68–82.

Chen, P., Shi, X., Liang, Y., Li, Y., Yang, L., Gader, P.D., 2020. Interactive thyroid whole slide image diagnostic system using deep representation. Computer Methods and Programs in Biomedicine 195, 105630.

Chen, X., Xie, S., He, K., 2021. An empirical study of training self-supervised vision transformers. arXiv preprint arXiv:2104.02057 .

Cheng, S., Wang, L., Du, A., 2019. Histopathological image retrieval based on asymmetric residual hash and dna coding. IEEE Access 7, 101388–101400.

Comaniciu, D., Meer, P., Foran, D., 1998a. Shape-based image indexing and retrieval for diagnostic pathology, in: Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on, IEEE. pp. 902–904.

Comaniciu, D., Meer, P., Foran, D., Medl, A., 1998b. Bimodal system for interactive indexing and retrieval of pathology images, in: Applications of Computer Vision, 1998. WACV. Proceedings. Fourth IEEE Workshop on, pp. 76–81.

Day, W.H.E., Edelsbrunner, H., 1984. Efficient algorithms for agglomerative hierarchical clustering methods. Journal of Classification 1, 7–24.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR 2021: The Ninth International Conference on Learning Representations.

Doyle, S., Hwang, M., S., N., MD., F., JE., T., A., M., 2007. Using manifold learning for content-based image retrieval of prostate histopathology., in: Medical Image Computing and Computer-Assisted Intervention.

Erfankhah, H., Yazdi, M., Babaie, M., Tizhoosh, H.R., 2019. Heterogeneity-aware local binary patterns for retrieval of histopathology images. IEEE Access 7, 18354–18367.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al., 2019. U-net: deep learning for cell counting, detection, and morphometry. Nature methods 16, 67.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent: A new approach to self-supervised learning, in: Advances in Neural Information Processing Systems, pp. 21271–21284.

Gu, Y., Jie, Y., 2018. Densely-connected multi-magnification hashing for histopathological image retrieval. IEEE journal of biomedical and health informatics 23, 1683–1691.

Gu, Y., Yang, J., 2019. Multi-level magnification correlation hashing for scalable histopathological image retrieval. Neurocomputing 351, 134–145.

Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B., 2009. Histopathological image analysis: a review. IEEE Reviews in Biomedical Engineering 2, 147–171.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738.

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

Hollon, T.C., et al., 2020. Near real-time intraoperative brain tumor diagnosis using stimulated raman histology and deep neural networks. Nature Medicine 26, 52–58.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Jia, Z., Huang, X., Eric, I., Chang, C., Xu, Y., 2017. Constrained deep weak supervision for histopathology image segmentation. IEEE transactions on medical imaging 36, 2376–2388.

Jiang, M., Zhang, S., Huang, J., Yang, L., Metaxas, D.N., 2016. Scalable histopathological image analysis via supervised hashing with multiple features. Medical Image Analysis 34, 3–12.

Kalra, S., Tizhoosh, H., Choi, C., Shah, S., Diamandis, P., Campbell, C.J., Pantanowitz, L., 2020. Yottixel – an image search engine for large archives of histopathology whole slide images. Medical Image Analysis , 101757.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks, in: Proceedings of Advances in Neural Information Processing Systems.

Li, Z., Zhang, J., Tan, T., Teng, X., Sun, X., Zhao, H., Liu, L., Xiao, Y., Lee, B., Li, Y., Zhang, Q., Sun, S., Zheng, Y., Yan, J., Li, N., Hong, Y., Ko, J., Jung, H., Liu, Y., cheng Chen, Y., wei Wang, C., Yurovskiy, V., Maevskikh, P., Khanagha, V., Jiang, Y., Yu, L., Liu, Z., Li, D., Schuffler, P.J., Yu, Q., Chen, H., Tang, Y., Litjens, G., 2021. Deep learning methods for lung cancer segmentation in whole-slide histopathology images—the acdc@lunghp challenge 2019. IEEE Journal of Biomedical and Health Informatics 25, 429–440.

Li, Z., Zhang, X., Müller, H., Zhang, S., 2018. Large-scale retrieval for medical image analytics: A comprehensive review. Medical image analysis 43, 66–84.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

Liu, W., Wang, J., Ji, R., Jiang, Y.G., 2012. Supervised hashing with kernels, in: Computer Vision and Pattern Recognition, pp. 2074–2081.

Ma, Y., Jiang, Z., Zhang, H., Xie, F., Zheng, Y., Shi, H., Zhao, Y., Shi, J., 2017. Breast histopathological image retrieval based on latent dirichlet allocation. IEEE Journal of Biomedical and Health Informatics 21, 1114–1123. doi:10.1109/JBHI.2016.2611615.

Ma, Y., Jiang, Z., Zhang, H., Xie, F., Zheng, Y., Shi, H., Zhao, Y., Shi, J., 2018. Generating region proposals for histopathological whole slide image retrieval. Computer methods and programs in biomedicine 159, 1–10.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9, 2579–2605.

Mehta, N., Alomari, R.S., Chaudhary, V., 2009. Content based sub-image retrieval system for high resolution pathology images using salient interest points., in: International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3719–22.

Peng, T., Boxberg, M., Weichert, W., Navab, N., Marr, C., 2019. Multi-task learning of a deep k-nearest neighbour network for histopathological image classification and retrieval, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 676–684.

Sapkota, M., Shi, X., Xing, F., Yang, L., 2018. Deep convolutional hashing for low-dimensional binary embedding of histopathological images. IEEE journal of biomedical and health informatics 23, 805–816.

Shi, X., Sapkota, M., Xing, F., Liu, F., Cui, L., Yang, L., 2018. Pairwise based deep ranking hashing for histopathology image classification and retrieval. Pattern Recognition 81, 14–22.

Shi, X., Xing, F., Xu, K., Xie, Y., Su, H., Yang, L., 2017. Supervised graph hashing for histopathology image retrieval and classification. Medical Image Analysis 42, 117.

Sparks, R., Madabhushi, A., 2011. Out-of-sample extrapolation using semi-supervised manifold learning (osessl): Content-based image retrieval for prostate histology grading. International Symposium on Biomedical Imaging , 734–737.

Sridhar, A., Doyle, S., Madabhushi, A., 2011. Boosted spectral embedding (bose): Applications to content-based image retrieval of histopathology, in: International Symposium on Biomedical Imaging, pp. 1897–1900.

Tizhoosh, H.R., Babaie, M., 2018. Representing medical images with encoded local projections. IEEE Transactions on Biomedical Engineering 65, 2267–2277.

Jimenez-del Toro, O., Otálora, S., Atzori, M., Müller, H., 2017. Deep multimodal case–based retrieval for large histopathology datasets, in: MICCAI 2018 Workshop on Patch-based Techniques in Medical Imaging, Springer. pp. 149–157.

Uijlings, J.R.R., Sande, K.E.A.V.D., Gevers, T., Smeulders, A.W.M., 2013. Selective search for object recognition. International Journal of Computer Vision 104, 154–171.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 5998–6008.

Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., et al., 2019. Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. Medical image analysis 54, 111–121.

Wetzel, A.W., Crowley, R., Kim, S., Dawson, R., Zheng, L., Joo, Y.M., Yagi, Y., Gilbertson, J., Gadd, C., Deerfield, D.W., 1999. Evaluation of prostate tumor grades by content-based image retrieval. Proceedings of SPIE - The International Society for Optical Engineering 3584, 244–252.

Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., Madabhushi, A., 2015. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. IEEE transactions on medical imaging 35, 119–130.

Xu, Y., Jia, Z., Wang, L.B., Ai, Y., Zhang, F., Lai, M., Chang, I.C., 2017. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. Bmc Bioinformatics 18, 281.

Xu, Y., Zhu, J.Y., Chang, I.C., Lai, M., Tu, Z., 2014. Weakly supervised histopathology cancer image segmentation and classification. Medical Image Analysis 18, 591–604.

Yan, R., Ren, F., Wang, Z., Wang, L., Zhang, T., Liu, Y., Rao, X., Zheng, C., Zhang, F., 2020. Breast cancer histopathological image classification using a hybrid deep neural network. Methods 173, 52–60.

Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W., Leskovec, J., 2018. Hierarchical graph representation learning with differentiable pooling, in: Advances in Neural Information Processing Systems, pp. 4800–4810.

Zhang, S., Metaxas, D., 2016. Large-Scale medical image analytics: Recent methodologies, applications and Future directions. Medical Image Analysis 33, 98–101.

Zhang, X., Dou, H., Ju, T., Xu, J., Zhang, S., 2015a. Fusing heterogeneous features from stacked sparse autoencoder for histopathological image analysis. IEEE Journal of Biomedical and Health Informatics 20, 1377–1383.

Zhang, X., Liu, W., Dundar, M., Badve, S., Zhang, S., 2015b. Towards Large-Scale Histopathological Image Analysis: Hashing-Based Image Retrieval. IEEE Transactions on Medical Imaging 34, 496–506. doi:10.1109/tmi.2014.2361481.

Zheng, L., Wetzel, A.W., Gilbertson, J., Becich, M.J., 2004. Design and analysis of a content-based pathology image retrieval system. IEEE Transactions on Information Technology in Biomedicine 7, 249–55.

Zheng, Y., Jiang, B., Shi, J., Zhang, H., Xie, F., 2019. Encoding histopathological wsis using gnn for scalable diagnostically relevant regions retrieval, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 550–558. doi:10.1007/978-3-030-32239-7_61.

Zheng, Y., Jiang, Z., Shi, J., Ma, Y., 2014. Retrieval of pathology image for breast cancer using plsa model based on texture and pathological features, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 2304–2308.

Zheng, Y., Jiang, Z., Xie, F., Zhang, H., Ma, Y., Shi, H., Zhao, Y., 2017. Feature extraction from histopathological images based on nucleus-guided convolutional neural network for breast lesion classification. Pattern Recognition 71, 14——25. doi:10.1016/j.patcog.2017.05.010.

Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Ma, Y., Shi, H., Zhao, Y., 2018a. Histopathological whole slide image analysis using context-based cbir. IEEE transactions on medical imaging 37, 1641–1652.

Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Ma, Y., Shi, H., Zhao, Y., 2018b. Size-scalable content-based histopathological image retrieval from database that consists of wsis. IEEE journal of biomedical and health informatics 22, 1278–1287.

Zhou, G., Jiang, L., 2004. Content-based cell pathology image retrieval by combining different features. Medical Imaging Pacs and Imaging Informatics 5371, 326–333.