This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Node-aligned Graph Convolutional Network for Whole-slide Image Representation and Classification

Yonghang Guan[1]*, Jun Zhang[2]*, Kuan Tian[2], Sen Yang[2], Pei Dong[2],
Jinxi Xiang[2], Wei Yang[2], Junzhou Huang[2], Yuyao Zhang[1]†, Xiao Han[2]†
[1]ShanghaiTech University, [2]Tencent AI Lab

Abstract

The large-scale whole-slide images (WSIs) facilitate the learning-based computational pathology methods. However, the gigapixel size of WSIs makes it hard to train a conventional model directly. Current approaches typically adopt multiple-instance learning (MIL) to tackle this problem. Among them, MIL combined with graph convolutional network (GCN) is a significant branch, where the sampled patches are regarded as the graph nodes to further discover their correlations. However, it is difficult to build correspondence across patches from different WSIs. Therefore, most methods have to perform non-ordered node pooling to generate the bag-level representation. Direct non-ordered pooling will lose much structural and contextual information, such as patch distribution and heterogeneous patterns, which is critical for WSI representation. In this paper, we propose a hierarchical global-to-local clustering strategy to build a Node-Aligned GCN (NAGCN) to represent WSI with rich local structural information as well as global distribution. We first deploy a global clustering operation based on the instance features in the dataset to build the correspondence across different WSIs. Then, we perform a local clustering-based sampling strategy to select typical instances belonging to each cluster within the WSI. Finally, we employ the graph convolution to obtain the representation. Since our graph construction strategy ensures the alignment among different WSIs, WSI-level representation can be easily generated and used for the subsequent classification. The experiment results on two cancer subtype classification datasets demonstrate our method achieves better performance compared with the state-of-the-art methods.

1. Introduction

Histopathology plays an essential role in the clinical diagnosis and understanding of the underlying reasons for specific treatments being deployed [35]. Whole-slide imaging, as the technique that translates the tissue specimens on glass slides into digital format without losing the tissue information, provides a comprehensive view of individual diseases and their effects on human tissues. Nowadays, the massive amount of whole-slide images (WSIs) makes the field of computational pathology an important application scenario for deep-learning-based computer aided diagnostic systems [14, 17, 18, 23, 27, 36, 39].

WSI classification is a fundamental task in digital pathology. However, the characteristics of pathological images pose unique challenges for deep-learning-based approaches. For example, the ultra-high resolution of WSIs prevents them from being directly fed into deep neural networks due to the huge memory consumption. In addition, since manual annotations require non-trivial effort and domain knowledge from pathologists, usually only slide-level labels are available and pixel- or region-level annotations are missing. To tackle these challenges, current methods often adopt a two-stage multiple-instance learning (MIL) paradigm. In the first stage, a group of disjoint or overlapping "tissue patches" cropped from WSIs are encoded into semantic features using an encoder. Then an aggregation algorithm is designed to integrate these instance-level features to obtain a bag-level (slide-level) representation.

There exist different ways of aggregating patch-level features. A simple yet efficient strategy is to directly pool the patch features in one WSI [12, 42]. However, WSIs contain tissues of varying morphologies and types. Brute force pooling of all patch features will dilute distinctive features, thus resulting in an inadequate representation of WSI. Bag of visual words (BOVW) is another classical method to describe images, which has been widely used in image representation, classification and retrieval [13, 44]. It provides an intuitive way to build descriptions for unstructured image data. Compared with pooling-based methods, BOVW can well delineate the global representation of WSI.

Despite the concise and straightforward definition, both pooling-based and BOVW-based methods are pre-defined and non-trainable. Recent approaches typically employ at-

*Equal contribution. Y. Guan interned at Tencent AI Lab.
†Corresponding authors.

tention mechanisms to aggregate patch-level features, in which a trainable neural network is used to fit the weights of instances and then take the weighted average sum of all instance features as WSI representation. However, attention-based approach is essentially a weighted linear combination of instances, which lacks the ability to reflect structural and contextual information of WSIs. Besides, since one WSI usually contains thousands of patches, the vast number of redundant patches will make attention-based approaches computationally costly.

Graph-based MIL approach is another significant branch for weakly-supervised classification of WSIs, which employs graphs to model the instance relationship and depict WSI representation. Graph convolution network (GCN) provides a powerful analytical paradigm for MIL and histopathology image, which first selects the important instances using a sampling strategy such as feature similarity, cell density or attention mechanisms, and then constructs a graph where graph nodes are selected instances and edges are the intrinsic relationship among instances [1, 37]. Adopting graphs to represent WSIs can reflect the complex contextual information, establish the dependency relationship between patches and reduce the gap between different data, which can be important for WSI diagnosis. However, existing graph-based WSI classification methods cannot ensure the correspondence of graphs nodes derived from different WSIs, which results in performing non-ordered pooling for global representation and classification.

In order to retain local structural information as well as global distribution, we propose a hierarchical global-to-local clustering strategy to build a Node-Aligned GCN (NAGCN[§]) for whole-slide image representation and classification. First, to filter out redundant information and select discriminative instances, we borrow the idea from BOVW and construct a codebook by leveraging a global clustering operation to instance features in the dataset. The codebook is comprised of amounts of visual words, where each visual word corresponds to a specific tissue type. Through the global clustering, we can divide instances from WSI bags into distinct sub-bags (each sub-bag corresponds to a visual word) and build correspondence across different WSIs at the sub-bag level. Second, we perform a local clustering-based sampling strategy to select typical instances within sub-bags for each WSI and use them as graph nodes. Finally, different from BOVW which only uses a non-trainable frequency histogram to represent WSIs, we deploy the node-aligned graph to achieve trainable WSI embeddings. Since our graph construction strategy ensures the alignment among different WSIs, WSI-level representation can be easily generated, which can be used for the subsequent classification.

We summarize our technical contributions as follows:

1. We introduce a novel node-aligned GCN for WSI representation and classification with only slide-level annotations. Compared with other graph-based MIL methods that have to perform non-ordered pooling to generate slide-level representation, our aligned graphs can establish node correspondence among WSIs, thus having more options to get the global representation, such as flattening the nodes.

2. We leverage global clustering to divide instances from WSI bags into distinct sub-bags and achieve correspondence at the sub-bag level across different WSIs. The pre-built codebook can well distinguish the pathological structures and partially reflect the tissue distribution.

3. We propose a local-clustering-based sub-bag generation strategy, which can sparsely sample typical instances within sub-bags. Combined with global clustering, the hierarchical global-to-local clustering can retain both local structural information and global distribution.

2. Related Work

Depending on at which level the classifier is adopted, MIL-based methods can be divided into two paradigms: instance-level and embedding-level [2]. Instance-level methods first utilize instance classifiers to predict the instance label, then aggregate these labels to generate the bag-level label following standard multiple-instance (SMI) assumption [2]. For embedding-level MIL methods, instances are first encoded to get semantic features, then the extracted features are aggregated to obtain the bag-level representation. In this section, only embedding-level MIL methods are discussed since instance-level approaches only focus on local information, whereas the global representation of WSIs is essential for the slide-level classification.

2.1. Conventional Bag Representation Approaches

Conventional bag representation methods adopt handcrafted aggregators. Pooling is one of the most common strategies. Pooling-based approaches aggregate all instances within a bag indiscriminately, which leverages maximum or average operation along the feature dimension to get the integrated feature. Obviously, pooling-based representation lacks the ability to distinguish instances and makes the obtained global representation inadequate. Bag of visual words (BOVW) [33] provides an intuitive way to build descriptions for unstructured image data and has been widely used for pathological image classification and retrieval [11, 13, 44]. Akin to multiple-instance learning, BOVW takes each WSI as a bag containing many instances. Specifically, BOVW constructs a codebook by clustering in-

[§]GitHub repository: https://github.com/YohnGuan/NAGCN

stances in the dataset, then assigns each instance to a specific visual word through nearest neighbor. The frequency histogram of the visual words is then used as the global representation. BOVW is insensitive to image scale, which makes it well suited for ultra-high resolution WSIs. However, BOVW only considers the frequency of words occurrence, which is inadequate for accurate WSI representation. Yet, all handcrafted aggregation methods are not trainable, which cannot well fit the complex application scenarios.

2.2. Attention-based MIL Approaches

Recently, leading-edge techniques adopt attention mechanism into MIL to tackle the weakly-supervised classification problem [5, 15, 19, 25, 26, 32, 40]. In essence, attention-based approach is to identify the contribution of each patch in the bag to the collective representation using a trainable neural network. Then the contributions of patches (are also called attention scores) are used to perform a weighted sum to get the global representation. For example, Ilse *et al*. [19] use a neural net to fit the importance of instance features as the attention score and obtain the bag-level representation by weighted sum. Li *et al*. also propose a dual-stream attention architecture [25], where the first stream selects the critical instance and the second stream determines the attention score of each instance through its distance to the critical instance. The outputs of two streams are averaged to generate the global representation of WSI. Xie *et al*. [40] cluster tiles of WSIs into K parts. The proposed method samples a single instance from each part, and uses the sampled K instances to represent one WSI. In [32], Sharma *et al*. adopt a local clustering operation to expose the model to diverse discriminative patches. The model performs a local clustering for each WSI, and samples instances from the clustering centroids before every epoch. However, local clustering performing on one WSI cannot guarantee the correspondence of clustering centroids among different WSIs. In this work, we combine global clustering and local clustering operation together to ensure the correspondence relationship and sample typical instances, which can be used to build the node-aligned graph. For all these methods, the weighted linear combination of instances limits the ability to obtain global contextual and structural information.

2.3. Graph-based MIL Approaches

Graph-based approaches have been widely utilized in computational pathology for multiple tasks [1, 4, 30, 37, 41, 45]. In computational pathology, a WSI can be abstracted as a graph, where the graph nodes represent biological structures (cells or tissue patches), and the graph edges reflect the internal relationships among biological structures. Due to the good property in relation-aware representations, graph provides a powerful tool to represent the biopsy slides in non-Euclidean space. According to the node level, WSI

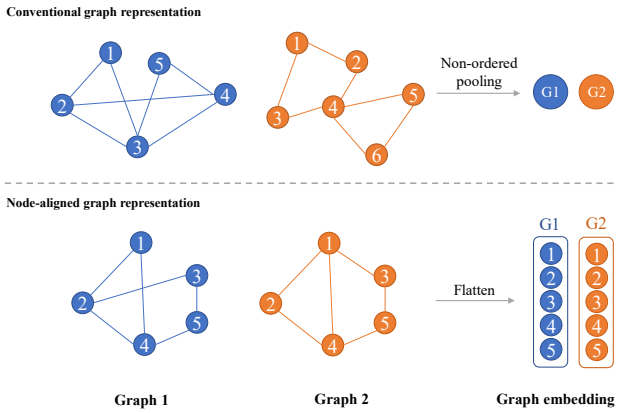

Figure 1. Illustration of correspondence in WSI graph representation. **Top**: Conventional tissue-graph representation. **Bottom**: Our proposed graph construction strategy.

graphs can be divided into cell-graphs and tissue-graphs. Cell-graph approaches [20, 34, 46] first detect the nuclei in the WSI using a detection network, followed by edge building based on spatial distance, which can well depict the cell micro-environment. For tissue-graph methods [45], the graph nodes are tissue patch features and the graph edges are connected based on the underlying relationships (such as feature similarity and spatial distance). [37] firstly combined graph neural network (GNN) with MIL to model the structural information among instances. The proposed algorithm treats each bag as a graph where instances are taken as graph nodes and edge connection is based on Euclidean distance. Zhao *et al*. [45] proposed a GCN-based MIL framework for lymph node metastasis prediction. The framework adopts a feature selection module based on histogram and maximum mean discrepancy to select the most relevant instances. Then, a spectral-GCN [10] followed by SAGPool [24] is employed to generate the slide-level representation. However, all these approaches do not consider the node correspondence in different WSI graphs. As illustrated in Figure 1, in those methods nodes from different WSI graphs may have different amounts and topological orders. To map these graphs to a fixed-length embedding, all these methods will inevitably employ non-ordered graph pooling operations, which will result in a loss of representation performance. In the light of BOVW, we build a node-aligned graph using a pre-built codebook, where the correspondence is ensured by the pre-formulated visual words.

3. Method

In this section, we present our overall framework for weakly-supervised multiple-instance learning, illustrated in Figure 2. The proposed framework consists of three components: instance sampling and encoding, hierarchical global-

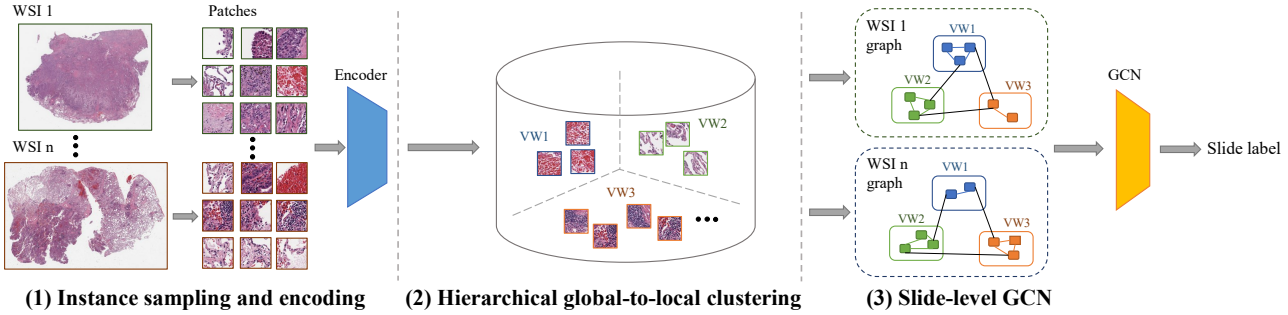

Figure 2. Overview of NAGCN. The proposed framework consists of 1) instance sampling and encoding, 2) hierarchical global-to-local clustering, 3) slide-level representation and classification. Patches are extracted from WSIs and fed into an encoder to generate the instance-level features. Then, the hierarchical global-to-local clustering is performed using the instance-level features to construct the node-aligned graphs. We leverage a novel graph construction strategy and a GCN to model the slide-level representation.

to-local clustering, and WSI-level graph representation.

3.1. Problem Formulation

Conventional MIL problem follows the standard multiple-instance (SMI) assumption [2]: A bag label is positive if and only if the bag contains at least one positive instance (i.e., an instance belongs to a certain target positive class), otherwise the bag label is negative. However, some of the WSI-level prediction problems are not applicable to the conventional SMI assumption, which rely on both the global and local representation of the entire slide.

In the context of weakly-supervised pathology image classification problem, we can consider each WSI as a bag consisting of multiple patches. Specifically, let $W = \{p^1, p^2, ..., p^n\}$ be a WSI which can be cut into a multitude of patches p^j (n is the number of sampled patches and p^j denotes j-th patch), the corresponding slide-level label is Y while the patch-level labels are unknown. The procedure of embedding-based MIL paradigm for pathology image classification is involved in the following steps:

1. A patch extraction strategy that samples patches $\{p^1, p^2, ..., p^n\}$ from WSI W and an instance encoder to generate instance features $f^j = \epsilon(p^j)$ and form the WSI bag $B = \{f^1, f^2, ..., f^n\}$ in the encoding space.

2. An instance embedding aggregation strategy to integrate instance-level features and generate the global representation $g = \rho(B)$. For graph-based MIL, the aggregation strategy first constructs graph G for the bag, and then embeds the graph-level representation g through GCN and graph pooling.

3. A bag-level classifier for the aggregated global representation $\hat{Y} = \xi(g)$.

3.2. Instance Encoding&Codebook Construction

Since the background in WSI occupies a significant portion and does not contribute to the WSI prediction, a foreground segmentation method is first adopted for the slide thumbnail to extract the tissue. WSI format data are usually stored as image pyramids, with each layer corresponding to a specific magnification and physical resolution of the digital biopsy sections. We extract the non-overlapping foreground patches with a fixed size at a particular magnification (e.g., 20x) as the multiple instances of the WSI. Each instance represents a local tissue region in the WSI, and the integration of these instances generates the slide-level representation of the WSI.

Let $D = \{W_1, W_2, ..., W_N\}$ be the training dataset that includes N WSIs. After pre-processing, each WSI W_i (W_i denotes i-th WSI in the dataset) is cropped into non-overlapping instances, where $W_i = \{p_i^1, p_i^2, ..., p_i^n\}$, and n is the number of sampled patches (n can vary for different WSIs.). A pre-trained deep neural network is then adopted to encode each instance p_i^j (p_i^j denotes j-th instance in W_i) into a fixed dimensional vector $f_i^j \in \mathbb{R}^{L\times 1}$ to capture the semantic information of the patch. After instance encoding, we transform the dataset into the encoding space $D_f = \{B_1, B_2, ..., B_N\}$, where $B_i = \{f_i^1, f_i^2, ..., f_i^n\} \in \mathbb{R}^{n\times L}$. Then we generate a global codebook $C = (vw_1, vw_2, ..., vw_{K_G})$ by performing K-means clustering to all the encoding features in D_f. In this work, we use the term "visual word" vw to represent the global cluster category. The constructed codebook divides the feature space into K_G sub-spaces, each corresponding to a visual word used to represent biopsy tissues with different structural, textural, and pathological properties. Note that only training dataset instances are used for codebook construction to prevent data leakage.

3.3. Node-aligned Graph MIL

3.3.1 Graph Construction

In this work, we propose a heuristic tissue-graph construction strategy in the light of BOVW model. Differ-

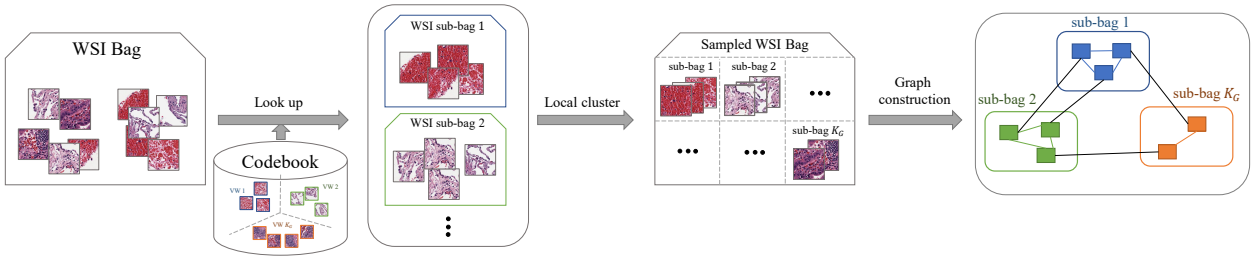

Figure 3. Illustration of the proposed graph construction process. First we look up the global codebook to divide the WSI bag into K_G sub-bags where each sub-bag denotes a specific visual word. Then local clustering is performed within each sub-bag to select a certain number of typical instances. The instances in the sampled WSI bag are taken as graph nodes and we connect the instance nodes using both inner-sub-bag edges (colored lines) and outer-sub-bag edges (black lines).

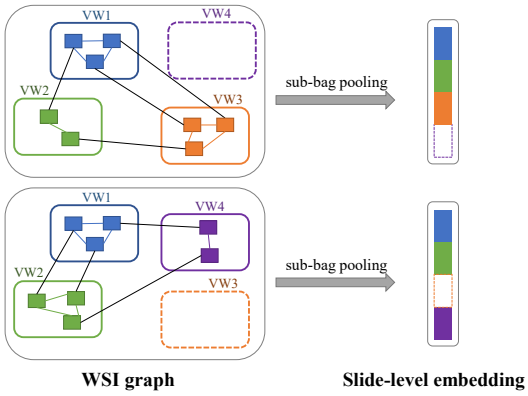

Figure 4. Illustration of the sub-bag pooling module. After GCN we obtain the encoded WSI graphs aligned at the sub-bag level. To generate slide-level representation, we pool the features within each sub-bag and connect the pooled sub-bag level features in the visual word order. Zero vectors (indicated as dashed lines) are used to fill in the positions corresponding to empty sub-bags.

ent from existing state-of-the-art graph-based MIL methods [37, 45], NAGCN can ensure feature alignment from different WSI graphs through a hierarchical global-to-local clustering strategy and an adjacency relationship construction mechanism, as illustrated in Figure 3.

Our proposed hierarchical global-to-local clustering strategy can screen out discriminative instances as graph nodes and achieve correspondence among different WSI graphs. It consists of a global clustering part and a local clustering part. Global clustering operation is the above-mentioned codebook construction, which performs clustering to all instance features in the training dataset. BOVW represents the WSI as the frequency histogram of visual words, where each bin/bucket of the histogram corresponds to a visual word, and the count of each bin corresponds to the number of instances in the bag belonging to this visual word. Similar to BOVW, we first look up the pre-built codebook and match each instance feature in the WSI bag with its nearest visual word. Through the codebook, the WSI bag is divided into K_G sub-bags, where each sub-bag contains the instances belonging to the same visual word. Since all WSI bags in the dataset share the same codebook, the graph node correspondence can be achieved at the sub-bag level. Clustering only at the global level can probably result in the loss of fineness due to millions of instances in our dataset. A local clustering sampling strategy is then performed to sample a fixed number of distinct instances inside each sub-bag. As shown in Figure 3, after global clustering a WSI bag B is divided into K_G sub-bags $B = \{sub_1, sub_2, ..., sub_{K_G}\}$ and each sub-bag sub_k (sub_k denotes the k-th sub-bag in B) contains instances belonging to the same visual word vw_k. For local clustering sampling, K-means clustering is conducted independently within each sub-bag sub_k to divide these patches into S_k bins and we randomly select one instance for each bin. Specifically, S_k is set to a random number within $[5, 25]$ to augment the data during training. In inference, S_k is set to a fixed value. After that in each WSI bag $B \in \mathbb{R}^{n \times L}$, we extract $V = \sum_{k=1}^{K_G} S_k$ instances to generate the sampled WSI bag $X \in \mathbb{R}^{V \times L}$ for graph construction.

We propose two kinds of edges to form the graph connection relationship: inner-sub-bag edges and outer-sub-bag edges. All instances belonging to the same sub-bag are interconnected to formulate the inner-sub-bag edges. For outer-sub-bag edges, each node is connected to its N_e nearest nodes through Euclidean distance. Inner-edge enables node feature communication within the sub-bag, while the outer-edge enables the information propagation among instances belonging to different visual words, which models the heterogeneous pattern. Since global clustering is usually sparse, which can leave many sub-bags empty in a WSI, null nodes (zero vector) are used to indicate that the corresponding sub-bag is empty and we do not connect null nodes to any others.

After that, we generate the input graph G for each WSI bag B as:

$$G = Graph(X, A), \tag{1}$$

where $X \in \mathbb{R}^{V \times L}$ represents node feature matrix (the sampled WSI bag) and $A \in \{0,1\}^{V \times V}$ denotes the adjacency matrix of G.

3.3.2 Graph Convolution and Bag Representation

Through the above-mentioned graph construction strategy, graphs from different WSIs can achieve correspondence relationships at the sub-bag level. The constructed graphs are fed into a GCN to conduct information passing over the graph. The output graph of GCN has the same node counts and orders as the input graph, which can be denoted as:

$$Z = GCN(Graph(X, A)), \tag{2}$$

where $Z \in \mathbb{R}^{V \times L'}$ is the node feature matrix of the output graph and L' is the output feature dimension. As illustrated in Figure 4, after GCN we adopt a sub-bag pooling module, which performs pooling operation to the node features within the same sub-bag. The pooled graphs are then flattened in the visual word order to get a fixed-length vector $g \in \mathbb{R}^{K_G L' \times 1}$, which can be used as the slide-level representation for the WSI classification.

4. Experiments

4.1. Dataset

The public pathology datasets greatly facilitate the development of computational pathology. In this work, the experiments are in whole based upon The Cancer Genome Atlas (TCGA) repository[§].

To evaluate the proposed approach, we construct two representative and clinically meaningful weakly-supervised cancer subtype classification datasets based on TCGA: TCGA non-small cell lung cancer (TCGA-NSCLC) and TCGA renal cell carcinoma (TCGA-RCC). NSCLC is one of the most common primary lung cancer types, which contains two cancer subtypes: Lung Adenocarcinoma (LUAD) [6] and Lung Squamous Cell Carcinoma (LUSC) [28]. We select 873 WSI digital slides from TCGA to build the TCGA-NSCLC dataset, which consists of 451 LUAD slides and 422 LUSC slides. RCC contains three common cancer subtypes: Kidney Chromophobe Renal Cell Carcinoma (KICH) [8], Kidney Renal Clear Cell Carcinoma (KIRC) [7] and Kidney Renal Papillary Cell Carcinoma (KIRP) [29]. Our dataset consists of a total of 726 WSI digital slides, with 118 KICH slides, 390 KIRC slides and 218 KIRP slides. During dataset construction, we removed all the frozen section digital slides and only preserved formalin-fixed paraffin-embedded hematoxylin and eosin (H&E) slides due to the poor quality of frozen slides.

[§] https://www.cancer.gov/tcga

4.2. Implementation Details

4.2.1 Patch Extraction

First, we use Otsu's method for foreground tissue extraction. To capture detailed information of the images, we extract a series of non-overlapping patches at $20\times$ with size 256×256 which contains more than 50% foreground tissue. Since some WSI slides in TCGA may not contain the pyramid layer at $20\times$, we first extract patches at $40\times$ with size 512×512, then resize them to 256×256 with bicubic interpolation. Finally, each WSI contains an average of 13904 patches for NSCLC and 14116 patches for RCC.

4.2.2 Networks

In our proposed framework, ResNet-50 [16] pre-trained by ImageNet1024 is adopted as the instance-level feature extraction backbone, which encodes 256×256 instance patches into 1024-dimensional vectors. All the clustering operations are performed upon the 1024-dimensional vector. For codebook construction, mini-batch K-means algorithm is performed with K-means++ [3] center initialisation. We deploy a three-layer graph convolutional network [22], with each layer followed by a ReLU activation, for feature propagation among nodes. The sub-bag pooling module is a pooling layer used to pool the node features within the same sub-bag. After sub-bag pooling the graph is flattened to generate the slide-level representation. A two-layer fully connected net is adopted as the classifier head.

4.2.3 Training and Evaluation

Cross Entropy loss and Adam [21] were adopted to optimise our model. The batch size was 256 and we set the learning rate as 5×10^{-5} with a linear decay. Our experiments used L2-regularization 5×10^{-3} and dropout rate 0.4 to mitigate model overfitting. To prevent information leakage, codebook construction is only performed on instances in the training dataset, and WSI graph construction is based upon this pre-built codebook during both training and inference. In our experiment, K_G is set to 100. During training, S_k is set to a random number within $[5, 25]$ and in inference S_k is set to 25. We set the number of outer-sub-bag edges N_e to 5 for each graph node. The feature dimension of output graph nodes L' is 16. After sub-bag pooling and flattening we can get a 1600 dimension vector for graph-level representation. All experiment results are obtained through 5-fold cross validation.

Accuracy and Area under the Curve of ROC (AUCROC) are used to evaluate the classification performance of different approaches. All the experiments are implemented using PyTorch [31] on a workstation with two Nvidia V100 GPUs and an Intel Xeon Gold 6133 CPU.

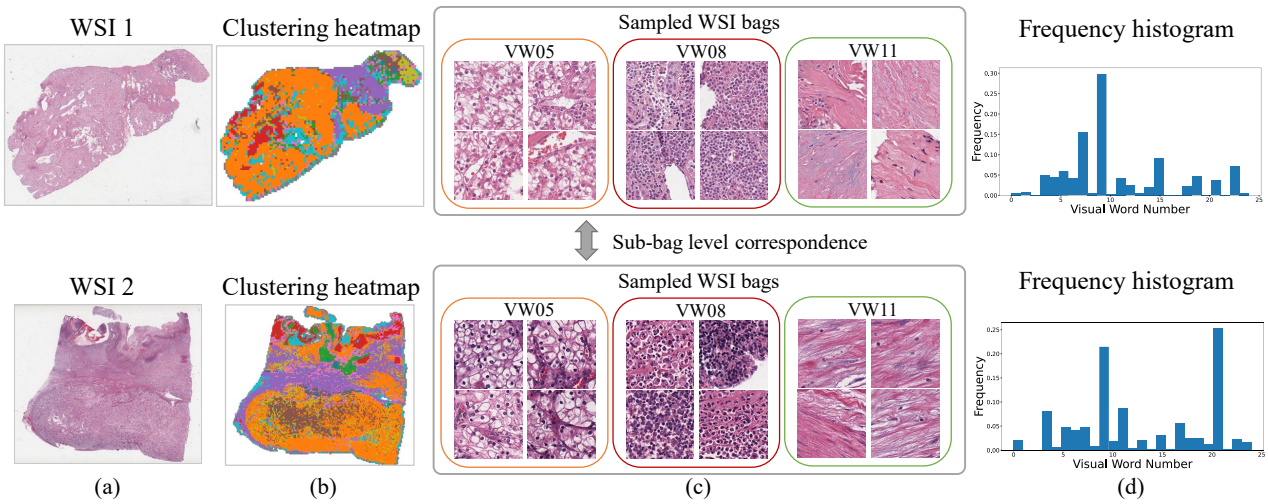

Figure 5. Visualisation of hierarchical clustering sampling strategy. Total two WSIs are presented. For each WSI, (**a**) thumbnail, (**b**) clustering heatmap, (**c**) examples of the sampled WSI sub-bags, (**d**) frequency histogram. For (**c**), different colored boxes represent sub-bags containing typical patches corresponding to specific tissue types (VW05: tumor cells, VW08: immune cells, VW11: stromal regions).

5. Results and Discussions

5.1. Visualisation of Hierarchical Global-to-local Clustering Strategy

Figure 5 shows the examples of hierarchical global-to-local clustering sampling strategy. To visualize the performance of global clustering, we overlay the visual word labels computed for each instance to form a heatmap. The frequency histograms are also presented to describe the overall distribution of instances. In addition, we select some typical visual words (sub-bags) to visualize the sampled WSI bags, which are also presented in Figure 5.

Overall, we observe that the hierarchical global-to-local clustering sampling strategy can reflect global distribution as well as local structures in pathological images and achieve instance alignment in the sub-bag level. From the heatmap, we can see that the global clustering result shows a high correspondence with the WSI tissue structures, which indicates that the pre-built codebook in the embedding space can well distinguish the pathological structures of WSIs. We observe that the clustering-based sampling strategy can divide WSI bags into distinct phenotype groups, where each visual word corresponds to a specific tissue type. Hence, the sampled instances can achieve correspondence among different WSI bags. Besides, extensive studies [9, 43] have validated that cancer development is highly correlated with abnormal stromal development and immune cell growth. Our sampled WSI bags can reflect global-wise and local-wise heterogeneous patterns, and thus better represent the slide-level information of WSI.

5.2. Comparison with State-of-the-art Methods

We conduct a comprehensive comparison of NAGCN with several state-of-the-art MIL approaches used in computational pathology on two representative cancer subtype classification datasets. The comparison methods include: (1-2) Max&Mean pooling, (3) Bag of Visual words (BOVW), (4) ABMIL [19], (5-6) CLAM with single-branch (CLAM-SB) and multi-branch (CLAM-MB) [26], (7) DSMIL [25], (8) C2C [32]. For fairness of comparison, we use ResNet50 [16] as the instance-level feature encoder for all methods. All approaches are evaluated using the same 5-fold cross-validation splits.

As shown in Table 1, NAGCN achieves the overall best performance. Our method improves over 8.7% in accuracy and 6.8% in AUCROC compared to conventional MIL methods (pooling and BOVW), which demonstrates the importance of trainable global representations for WSI classification tasks. Both CLAM and C2C outperforme ABMIL, which illustrates the effectiveness of the clustering-based sampling strategy. Our approach combines global clustering and local clustering together, while realizing both node alignment and feature filtering, thus improving about 4.0% in accuracy and 2.0% in AUCROC. DSMIL had poorer performance compared with our method since the attention mechanism lacks the ability to reflect structural and contextual information.

5.3. Ablation Study

5.3.1 Effect of Modules in NAGCN

Our proposed work can be divided into three main parts: (1) Global-clustering-based codebook construction (GC), (2)

Table 1. The comparison between NAGCN and SOTA methods.

	TCGA-NSCLC		TCGA-RCC	
	Accuracy	AUCROC	Accuracy	AUCROC
Max pooling	0.815	0.884	0.887	0.959
Mean pooling	0.753	0.805	0.814	0.943
BOVW	0.774	0.823	0.847	0.957
ABMIL	0.860	0.921	0.906	0.977
CLAM-SB	0.861	0.936	0.900	0.988
CLAM-MB	0.862	0.932	0.914	0.989
DSMIL	0.874	0.949	0.911	0.986
C2C	0.873	0.938	0.919	0.987
NAGCN	**0.902**	**0.952**	**0.954**	**0.992**

Local-clustering-based sub-bag generation strategy (LC), (3) graph convolution network and sub-bag level node feature pooling (GCN). Ablation study was performed to validate the effectiveness of these components using the TCGA-NSCLC dataset. During the ablation study, we set K_G as 100, S_k as 15 and N_e as 5.

To verify the criticality of node alignment for graph representation, we perform local clustering for each WSI bag respectively to generate unaligned sub-bags as an alternative to global codebook construction. As shown in Table 2 (A) and (C), node alignment strategy can improve accuracy with 4.1% and AUCROC with 2.6%. Besides, the comparison between (B) and (A) indicates that local-clustering-based sub-bag generation can bring improvement to the results. We also directly use the sampled WSI bags without graph construction and replace GCN with (D) attention modules and (E) max pooling to confirm the power of graph to model the instance correlations and structural information. As shown in Table 2, graph and GCN can improve accuracy by 5.2% and 8.3%, AUCROC by 3.9% and 6.1%, respectively.

Table 2. Effects of model modules in NAGCN.

Modules	Accuracy	AUCROC
(A) GC + LC + GCN (**Ours**)	0.896	0.946
(B) GC + Random sampling + GCN	0.889	0.945
(C) Unaligned sub-bags + GCN	0.855	0.920
(D) GC + LC + attention	0.844	0.907
(E) GC + LC + max pooling	0.813	0.885

5.3.2 Effect of Model Hyper-parameters

We now systematically explore the effect of model hyper-parameters in NAGCN on TCGA-NSCLC dataset. We will discuss the impact of the following model parameters: (1) number of visual words and (2) number of outer-sub-bag edges for each node.

As shown in Figure 6(a), more visual words will allow global clustering to divide the embedding space more finely, thus improving the model performance. As the number of visual words increases, the model performance growth tends to level off, which indicates that the codebook is sufficient to characterize patches in the dataset. From Figure 6(b) we can see that the increase in the number of outer-sub-bags leads to a slight decrease in performance, which is caused by the over-smoothing during graph convolution.

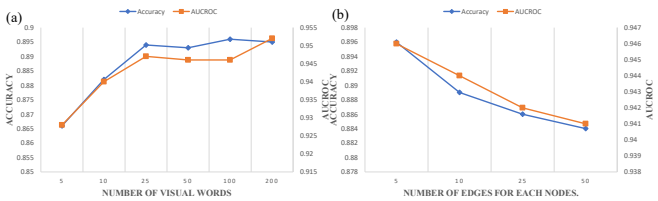

Figure 6. Performance of cancer subtype classification with different model hyper-parameters on TCGA-NSCLC dataset. Comparison results with the different number of (**a**) visual words and (**b**) outer-sub-bag edges for each node.

5.3.3 Improvement with Self-supervised Learning

To further explore the potential of self-supervised learning for NAGCN performance improvement, we implement a self-supervised model for pathological images according to [38] and adopt it as the instance-level encoder. We compare the performance using both ImageNet pre-trained model and self-supervised model on the TCGA-NSCLC dataset. With the same parameter settings, the self-supervised model achieves higher performance, including the accuracy of 0.928 and AUC of 0.972. The self-supervised model can improve over 3% in accuracy and AUC, which further validates the scalability of NAGCN.

6. Conclusion

In this work, we present the Node-Aligned GCN (NAGCN) for weakly-supervised WSI representation and classification. Our approach proposes a novel graph construction strategy based on a hierarchical global-to-local clustering, which not only retains both local structural information and global distribution, but also enables the alignment of different WSIs, thus avoiding non-ordered pooling to obtain the bag-level representation. A limitation in our current study is that the performance of our method relies heavily on the effectiveness of codebook construction. When the instance-level representation is not sufficiently robust or the region of interest is less occupied, the representational ability of WSI graphs may be compromised. Future work would focus on how to achieve joint feature extraction and global clustering.

References

[1] David Ahmedt-Aristizabal, Mohammad Ali Armin, S. Denman, C. Fookes, and L. Petersson. A survey on graph-based deep learning for computational histopathology. *ArXiv*, abs/2107.00272, 2021. 2, 3

[2] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013. 2, 4

[3] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics. 6

[4] Bulut Aygüneş, Selim Aksoy, Ramazan Gökberk Cinbiş, Kemal Kösemehmetoğlu, Sevgen Önder, and Ayşegül Üner. Graph convolutional networks for region of interest classification in breast histopathology. In *Medical Imaging 2020: Digital Pathology*, volume 11320, pages 134 – 141. SPIE, 2020. 3

[5] Gabriele Campanella, Matthew Hanna, Luke Geneslaw, Allen Miraflor, Vitor Silva, Klaus Busam, Edi Brogi, Victor Reuter, David Klimstra, and Thomas Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25:1, 08 2019. 3

[6] Eric Collisson, Barry Taylor, Joshua Campbell, Angela Brooks, Alice Berger, Juliann Chmielecki, Gad Getz, Peter Hammerman, Bryan Hernandez, Carrie Sougnez, Andrew Cherniack, Mara Rosenberg, Matthew Meyerson, Stacey Gabriel, Kristian Cibulskis, Jaegil Kim, Chip Stewart, Lee Lichtenstein, Eric Lander, and Neil Hayes. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature*, 511:543–550, 07 2014. 6

[7] Chad Creighton, Margaret Morgan, Preethi Gunaratne, David Wheeler, Richard Gibbs, Gordon Robertson, Andy Chu, Rameen Beroukhim, Kristian Cibulskis, Sabina Signoretti, Fabio Wu, Ben Raphael, Roel Verhaak, Pheroze Tamboli, Wandaliz Torres-García, Rehan Akbani, John Weinstein, Victor Reuter, James Hsieh, and Heidi Sofia. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499, 06 2013. 6

[8] Caleb F Davis, Christopher J Ricketts, Min Wang, Lixing Yang, Andrew D Cherniack, Hui Shen, Christian Buhay, Hyojin Kang, Sang Cheol Kim, Catherine C Fahey, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*, 26(3):319–330, 2014. 6

[9] Karin E De Visser, Alexandra Eichten, and Lisa M Coussens. Paradoxical roles of the immune system during cancer development. *Nature Reviews Cancer*, 6(1):24–37, 2006. 7

[10] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. NIPS'16, Red Hook, NY, USA, 2016. Curran Associates Inc. 3

[11] Meghana Dinesh Kumar, Morteza Babaie, Shujin Zhu, Shivam Kalra, and H. R. Tizhoosh. A comparative study of CNN, BoVW and LBP for classification of histopathological images. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, 2017. 2

[12] Ji Feng and Zhi-Hua Zhou. Deep MIML network. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1884–1890. AAAI Press, 2017. 1

[13] Antonio Foncubierta-Rodríguez, Alba García Seco de Herrera, and Henning Müller. Medical image retrieval using bag of meaningful visual words: Unsupervised visual vocabulary pruning with PLSA. In *Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare*, page 75–82, New York, NY, USA, 2013. Association for Computing Machinery. 1, 2

[14] Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16549–16559, June 2021. 1

[15] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6, 7

[17] Mahdi S. Hosseini, Lyndon Chan, Gabriel Tse, Michael Tang, Jun Deng, Sajad Norouzi, Corwyn Rowsell, Konstantinos N. Plataniotis, and Savvas Damaskinos. Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[18] Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M. Kurc, Rajarsi R. Gupta, and Joel H. Saltz. Robust histopathology image analysis: To label or to synthesize? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[19] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136, 10–15 Jul 2018. 3, 7

[20] Guillaume Jaume, Pushpak Pati, Behzad Bozorgtabar, Antonio Foncubierta, Anna Maria Anniciello, Florinda Feroce, Tilman Rau, Jean-Philippe Thiran, Maria Gabrani, and Orcun Goksel. Quantifying explainers of graph neural networks in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8106–8116, 2021. 3

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[22] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 6

[23] Jeroen Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature Medicine*, 27:775–784, 05 2021. 1

[24] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *International Conference on Machine Learning*, pages 3734–3743. PMLR, 2019. 3

[25] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, June 2021. 3, 7

[26] Ming Lu, Drew Williamson, Tiffany Chen, Richard Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:1–16, 06 2021. 3, 7

[27] Sam Maksoud, Kun Zhao, Peter Hobson, Anthony Jennings, and Brian C. Lovell. SOS: Selective objective switch for rapid immunofluorescence whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[28] Matthew Meyerson, Stephen Baylin, Ramaswamy Govindan, Rehan Akbani, Ijeoma Azodo, David Beer, Ron Bose, Lauren A.Byers, David Carbone, Li-Wei Chang, Derek Chiang, Andy Chu, Elizabeth Chun, Eric Collisson, Leslie Cope, Chad Creighton, Ludmila Danilova, Li Ding, Gad Getz, and Olga Potapova. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489:519–25, 09 2012. 6

[29] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of papillary renal-cell carcinoma. *New England Journal of Medicine*, 374(2):135–145, 2016. 6

[30] Yigit Ozen, Selim Aksoy, Kemal Kösemehmetoğlu, Sevgen Önder, and Ayşegül Üner. Self-supervised learning with graph neural networks for region of interest retrieval in histopathology. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6329–6334, 2021. 3

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 6

[32] Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A. Moskaluk, Sana Syed, and Donald Brown. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, 2021. 3, 7

[33] Josef Sivic and Andrew Zisserman. Video google: a text retrieval approach to object matching in videos. *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, 2003. 2

[34] Mookund Sureka, Abhijeet Patil, Deepak Anand, and Amit Sethi. Visualization for histopathology images using graph convolutional neural networks. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 331–335, 2020. 3

[35] Junko Tanizaki, Hidetoshi Hayashi, Masatomo Kimura, Kaoru Tanaka, Masayuki Takeda, Shigeki Shimizu, Akihiko Ito, and Kazuhiko Nakagawa. Report of two cases of pseudoprogression in patients with non–small cell lung cancer treated with nivolumab—including histological analysis of one case after tumor regression. *Lung Cancer*, 102:44–48, 2016. 1

[36] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[37] Ming Tu, Jing Huang, Xiaodong He, and Bowen Zhou. Multiple instance learning with graph neural networks, 06 2019. 2, 3, 5

[38] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 186–195, 2021. 8

[39] Bingzhe Wu, Shiwan Zhao, Guangyu Sun, Xiaolu Zhang, Zhong Su, Caihong Zeng, and Zhihong Liu. P3SGD: Patient privacy preserving SGD for regularizing deep CNNs in pathological image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[40] Chensu Xie, Hassan Muhammad, Chad M Vanderbilt, Raul Caso, Dig Vijay Kumar Yarlagadda, Gabriele Campanella, and Thomas J Fuchs. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning. In *Medical Imaging with Deep Learning*, pages 843–856. PMLR, 2020. 3

[41] Haili Ye, Da-Han Wang, Jianmin Li, Shunzhi Zhu, and Chenyan Zhu. Improving histopathological image segmentation and classification using graph convolution network. In *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, ICCPR '19, page 192–198, 2019. 3

[42] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1

[43] Cédric Zeltz, Irina Primac, Pugazendhi Erusappan, Jahedul Alam, Agnes Noel, and Donald Gullberg. Cancer-associated fibroblasts in desmoplastic tumors: emerging role of integrins. In *Seminars in Cancer Biology*, volume 62, pages 166–181. 7

[44] Xiaofan Zhang, Wei Liu, Murat Dundar, Sunil Badve, and Shaoting Zhang. Towards large-scale histopathological image analysis: Hashing-based image retrieval. *IEEE Transactions on Medical Imaging*, 34(2):496–506, 2015. 1, 2

[45] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, and Jianhua Yao. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 5

[46] Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. CGC-Net: Cell graph convolutional network for grading of colorectal cancer histology images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 3