HAT-Net: A Hierarchical Transformer Graph Neural Network for Grading of Colorectal Cancer Histology Images

Yihan Su*[1,2]
yhsu@bupt.edu.cn
Yu Bai*[1,2]
by@bupt.edu.cn
Bo Zhang‡[1,2]
zbo@bupt.edu.cn
Zheng Zhang[2]
zhangzheng@bupt.edu.cn
Wendong Wang‡[1,2]
wdwang@bupt.edu.cn

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
Beijing 100876, China

[2] Beijing University of Posts and Telecommunications
Beijing 100876, China

Abstract

Graph-based learning methods have gained more attention in colorectal adenocarcinoma cancer (CRA) grading tasks for encoding the tissue structure information, which patch-wise CNN based methods fail to. Graph-based methods usually involve extracting nuclei features in the histology images as cell-graph node features and modeling the connections between nodes to construct cell-graphs. However, it is infeasible to directly train a classification model to extract nuclei features as we normally do in nature images since different types of nuclei often cluster together. We propose a Masked Nuclei Patch (MNP) approach to train a ResNet-50 as a strong feature encoder to extract more representative nuclei feature for enhancing the overall performance. Graph Neural Networks (GNNs) are often used to train cell-graphs for different tasks. But GNN may struggle to capture the long-range dependency due to its underlying recurrent structure. Therefore, we propose a new network architecture named **H**ier**A**rchical **T**ransformer Graph Neural **Net**work (HAT-Net), which merits both GNN and Transformer, as a strong competitor for CRA grading tasks. We have achieved the state-of-the-art results on two publicly available CRA grading datasets: the colorectal cancer (CRC) dataset (98.55%) and the extended colorectal cancer (Extended CRC) dataset (95.33%).

1 Introduction

According to Colorectal Cancer Statistics 2020, colorectal cancer (CRC) is the second cause of cancer death worldwide [32] [34]. A well-adapted CRA grading guideline [17] classifies the adenocarcinomas into 3 levels: normal, low grade, and high grade based on the large

* Equal co-authorship, ‡ Corresponding author email: {wdwang, zbo}@bupt.edu.cn .

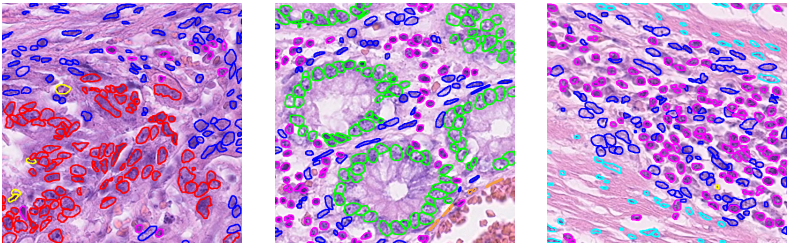


Figure 1: Regions cropped from CoNSeP dataset, different colors of nuclei boundary denote different nuclei types. It is ambiguous to assign a nuclei label to a single image, which makes it infeasible to train a CNN encoder like ResNet50 for nuclei feature extraction. ■ Malignant/dysplastic epithelial, ■ Normal epithelial, ■ Inflammatory, ■ Fibroblast, ■ Muscle, ■ Endothelial, ■ Miscellaneous.

histology images, known as Whole Slice Images (WSIs), which often have larger resolution than natural images, some of them even have ten billion pixels per image.

Due to the advancement of deep learning, many automatic algorithms are invented to help pathologists to avoid inter-observer and intra-observer variability [7]. The common CNN based approaches [3, 4, 24, 25, 31] for cancer grading tasks often perform a patch-wise prediction, and finally aggregate all the predictions into an image-level prediction. But these approaches fail to capture the tissue structure information due to the small patch size. Graph-based methods [2, 22, 27, 40, 46] are proposed to alleviate this issue, they construct cell-graphs from histology images to encode the tissue structure information, where nuclei are the nodes of a cell-graph, and interactions between different nuclei are modeled as graph edges. For the CRA grading task, Zhou *et al*. [46] propose a network architecture CGC-Net to fuse the multi-level cell graph information, 16 hand-crafted nuclei features and spatial features are used as the node features for cell-graph construction, also shows that more informative nuclei features can contribute more to the overall performance, similar results can be observed in [40]. However, none of the current works train a CNN model like ResNet-50 [19] to extract informative nuclei feature as we normally do in nature images [20, 30], because different types of nuclei often cluster together, see Fig 1. GNN is often used to train the cell-graphs, but GNN may struggle to model long-range dependencies due to the recurrent structure [15].

To alleviate the above issues, we propose a Masked Nuclei Patch (MNP) method to train a ResNet-50 as a strong nucleus feature extractor. This method crops nuclei patches for each nucleus with the given nucleus at the center of each patch, and mask out nuclei with the different types than the one in the center to make sure only a single type of nuclei is presented in a single patch. We train our model on these cropped patches to learn the abstraction of each nuclei type. Our MNP approach requires nuclei coordinates and shape information for cropping, so we utilize a public colorectal nuclei instance segmentation dataset CoNSeP [14] to train our model. We also propose a new graph neural network architecture: **H**ier**A**rchical **T**ransformer Graph Neural **Net**work (HAT-Net), which builds upon Graph Isomorphism Network (GIN) [42] to better distinguish similar graphs. We add MinCutPool [5] layer after GIN to cluster graph nodes and learn the hierarchical representation of graphs. Furthermore, we add the Transformer [13] in the graph neural network, which can model long-distance features. The main contributions are as follows:

- We develop a Masked Nuclei Patch (MNP) approach to enable us to directly train a

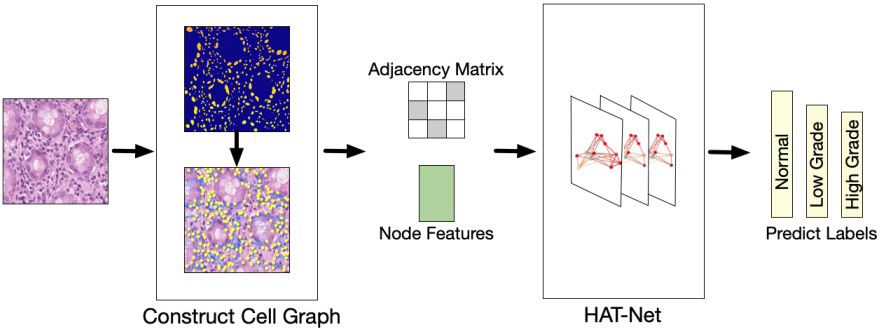


Figure 2: Overview of the proposed framework.

ResNet-50 model on histology images to extract more comprehensive nuclei features for better overall performance.

- We propose a hierarchical network HAT-Net, which combines GIN module and MinCutPool module to distinguish similar graphs more accurately, and adds Transformer module to learn long-distance dependencies.

- We have done extensive experiments, showing that the proposed approach has achieved the state-of-the-art result on the Colorectal Cancer (CRC) dataset [9] (98.55%) and the extended colorectal cancer (Extended CRC) dataset [31] (95.33%).

2 Related Work

Cell Graph: A cell graph is a structural representation of the tissue in the histology image where the nuclei and interactions between them are modeled as a graph. [1, 2, 16, 40, 46] construct cell graphs from the histology images, convert the grading tasks into graph classification tasks. Since different nuclei often clustered together, different methods are used to extract nuclei features, hand-crafted method [27, 46], weakly supervised approach [40], K-means [41] combines with image thresholding [16], etc. However, using a CNN encoder to extract the nuclei feature is still needed. [22] explores the explainability of cell-graph.

Graph Neural Network: GNN is designed to processing structure data like graphs [18, 23, 39, 42], it recursively aggregates neighborhood node features for each iteration. [18] continuously aggregates the information of neighboring nodes, generalizes to unseen nodes or graphs, but can not distinguish isomorphic graphs well. [42] is much more powerful for distinguish isomorphic graphs by aggregating nodes through summation. [39] uses a self-attention approach to aggregates only neighboring nodes. Graph pooling [6, 44] approaches map the graph into a compact form to generate a meaningful graph representation.

Transformer: Transformer is proposed in [38] to solve machine translation problem, and has been proven effective to model long-range dependencies in many NLP tasks [8, 12, 28, 29]. [10, 13, 36] apply Transformer to different vision tasks. Transformer is also used for medical image analysis [3, 37, 45].

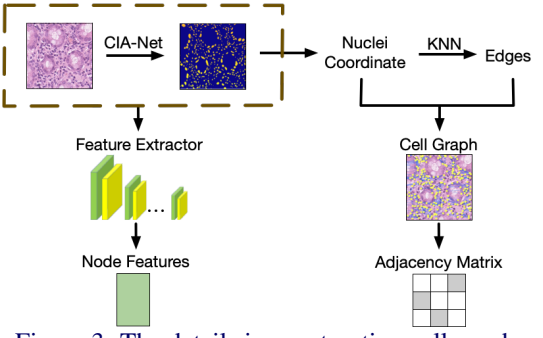


Figure 3: The details in constructing cell graph.

3 Proposed Approach

Our approach is illustrated in Fig.2. We generate the training data (adjacency matrix and node features) through a cell graph construction procedure (Sec. 3.1), then use the generated data to train our proposed HAT-Net (Sec. 3.2), so that we can get the prediction as the result.

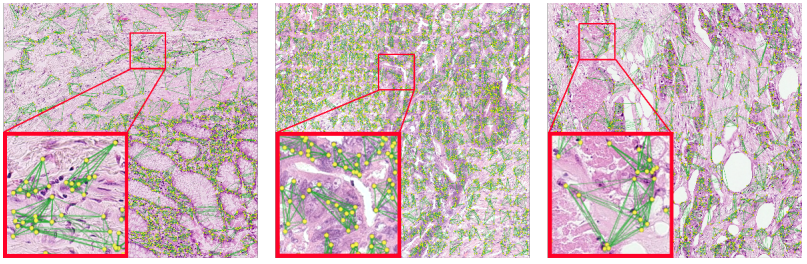
3.1 Construct Cell Graph

We construct cell-graph from histology images to encode tissue structure information. A cell-graph is composed of the adjacency matrix and nuclei features, we describe how to generate the adjacency matrix here, then elaborate on our nuclei feature extraction method.

Adjacency matrix generation: We regard nuclei as the nodes of the cell graph, so we use CIA-Net [47] to perform nuclei instance segmentation on the histology images to obtain nuclei coordinates and shape information. Furthermore, we calculate the Euclidean distance between nuclei in the same cell graph and connect them as graph edges, the process of building cell graphs from histology images is shown in Fig. 3. To represent the cell graph with the smallest adjacency matrix possible, we keep each node's edges connect to its k-nearest neighbors and define a maximum Euclidean distance (d_{max}) to reduce the number of edges. The calculation process of the adjacency matrix is as follows:

$$A[i,j] = \begin{cases} 1 & if\ j \in KNN(i)\ and\ D(i,j) < d_{max} \\ 0 & otherwise \end{cases} \tag{1}$$

CNN-based nuclei features extraction: CGC-Net adopts 16 different hand-crafted features and spatial features as the nuclei node features [46], but we argue that the hand-crafted features are not as effective as the learned features, such as CNN features. However, nuclei of different types often clustering together, makes it hard to train a CNN model. Thus, we proposed a Masked Nuclei Patches (MNP) approach to enable us directly train a CNN encoder for feature extraction. CoNSeP dataset [14] is a publicly available dataset with nuclei segmentation labels and center pixel coordinates. Based on the nuclear center pixel coordinates, we crop out each nucleus from the images with a patch size of 64×64 pixels. If the nuclear's height or width is greater than 64 pixels, we crop out the entire nucleus instead of 64 pixels. To alleviate the class imbalance problem of CoNSeP, we group 7 classes of CoNSeP into 3 classes. Nuclei of different classes often cluster together, so we mask out different classes of nuclei other than the one in the center, make sure that a single patch only contains one class of nuclei, to avoid confusing the network during training. We split in a



(a) Normal (b) Low grade (c) High grade

Figure 4: The cell graph of different histology images: (a) normal level, (b) low level and (c) high level.

ratio of 9:1 into training and testing sets, resulting in 21899 patches for training and 2433 patches for testing, a total of 24332 patches. We use an ImageNet pre-trained ResNet-50 as our feature extractor, fine-tuned on our nuclei classification dataset. The first max-pooling layer is removed since our input image is small (64×64). The extra global average pooling is applied for dimension reduction during feature extraction. During feature extraction, we crop out the nuclei region based on the instance segmentation results from CIA-Net, resize to 64x64 pixels, then feed into the ResNet-50 network. The output features right before the MLP layer will be used as our nuclei features since it is more linearly separable. Additional parameter-free channel dimension reduction will be applied through global average pooling.

3.2 Hierarchical Network Architecture

The recurrent structure of GNN makes it hard to model long-range dependencies, we combine Transformer with GNN to alleviate this issue. We propose a Hierarchical Transformer Graph Neural Network (HAT-Net), which contains three stages and each stage is an aggregation of different modules, as shown in Fig. 5. The input of the HAT-Net is the cell-graph, represented by the adjacency matrix and node features, the output is the classification results of histology images. Our network builds upon GIN, which can distinguish similar graphs more accurately, we add the MinCutPool module into our network to cluster similar nodes. Due to the large computational complexity of Transformer[38], we only add the Transformer in the second stage and the third stage to reduce the computational cost. We apply a LSTM-based jumping knowledge [43] to fuse hierarchical information from different layers. The MLP layer is applied to produce the final prediction.

Notations Definition: Given a graph $G = (V,E)$, where V represents a set of nodes, E represents a set of edges. The adjacency matrix $A \in \mathbb{R}^{n\times n}$ is used to describe the graph edges, and the matrix $\mathbf{h} \in \mathbb{R}^{n\times d}$ is used to describe the node features, where d is the feature dimensions of the nodes. We use $A_i \in \mathbb{R}^{n_i\times n_i}$ and $\mathbf{h}_i \in \mathbb{R}^{n_i\times d_i}$ to represent adjacency matrix and node feature matrix in the ith layer. We have two GIN at each stage before the pooling layer, we denote the output from one GIN as embedding matrix $M_i \in \mathbb{R}^{n_i\times d_i}$, and output from another GIN as assignment matrix $S_i \in \mathbb{R}^{n_i\times n_{i+1}}$, n_{i+1} is the number of cluster, also the number of nodes after pooling layer. And for a node u, we use $h^i(u)$ to represent node embedding in the ith layer, use $N(u)$ to denote the set of neighborhood nodes of u. What's more, MSA represents Multi-Head Attention and LN represents Layer Norm [5].

GNN Module: GIN [42] has a provably strong power to learn the discriminative graph features. Specifically, we apply two GIN in each stage, one is to calculate M_i, the other is to

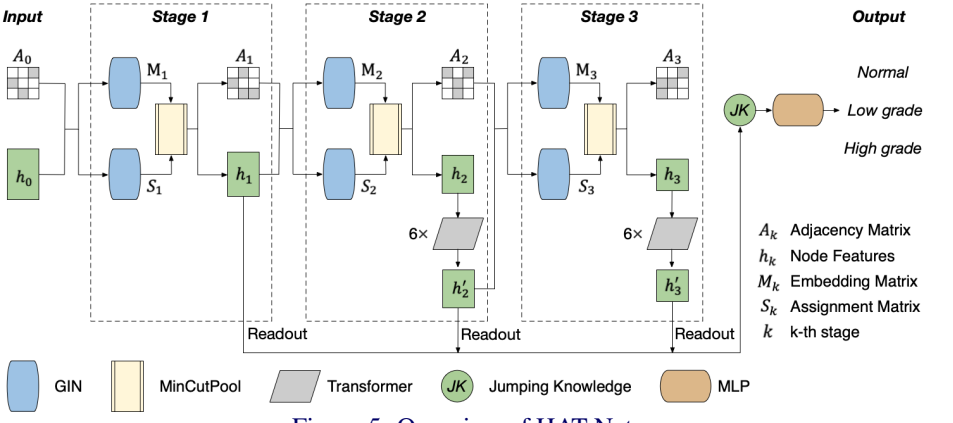


Figure 5: Overview of HAT-Net.

calculate S_i. First, we compute node embedding $h^i(u)$ through $\mathbf{h_{i-1}}$, see Eq. 2.

$$\mathbf{h^i(u)} = \mathbf{MLP}(\mathbf{h^{i-1}(u)} + \sum_{\mathbf{v \in N(u)}} \mathbf{h^{i-1}(v)}) \tag{2}$$

Then we can get the graph level embedding by concatenating all node embeddings by Eq. 3.

$$\mathbf{M_i} = \mathbf{CONCAT}(\sum_{\mathbf{u \in G}} \mathbf{h^i(u)} | \mathbf{i} = \mathbf{0}, \mathbf{1}) \tag{3}$$

In addition to changing the embedding dim d, we can also obtain S_i in the same way.

Pooling Module: We use MinCutPool [6] as the pooling module, which is a method of spectral clustering. After we obtain S_i and M_i by GIN, we will feed them into MinCutPool to generate the new node features of the next layer. Then the feature matrix $\mathbf{h}_{i+1}$ of the new node can be calculate by Eq. 4, and the corresponding adjacency matrix A_{i+1} can be calculated with Eq. 5. What's more, two losses are generated: the orthogonality loss and the minCUT loss, please refer [6] for more details.

$$\mathbf{h_{i+1}} = \mathbf{softmax}(\mathbf{S_i})^{\mathbf{T}} \cdot \mathbf{M_i} \tag{4}$$

$$\mathbf{A_{i+1}} = \mathbf{softmax}(\mathbf{S_i})^{\mathbf{T}} \cdot \mathbf{A_i} \cdot \mathbf{softmax}(\mathbf{S_i}) \tag{5}$$

Transformer architecture: We integrate the Transformer [13] into our network. Through the self-attention mechanism in the Transformer, a node can pay attention to the features of nodes far away from it. Thus we can use Transformer to learn the long-distance features of nodes. We stack 6 Transformer layers together to enhance the performance. We use Eq. 6 and Eq. 7 to generate the new features in each Transformer layer:

$$\mathbf{h'_{l+1}} = \mathbf{MSA}(\mathbf{LN}(\mathbf{h_l})) + \mathbf{h_l} \tag{6}$$

$$\mathbf{h_{l+1}} = \mathbf{MLP}(\mathbf{LN}(\mathbf{h'_{l+1}})) + \mathbf{h'_{l+1}} \tag{7}$$

Jumping Knowledge Technique: We add an LSTM-based jumping knowledge [43] in our network to fuse hierarchical features from different layers. For each layer, we use a Max Readout to obtain the graph representation of each layer. We feed all the graph representations into a LSTM-based jumping knowledge to fuse the different graph representations.

Classification Module: In order to output the final prediction, we feed the output of the jumping knowledge module to the linear layer for a 3-class classification prediction, a cross-entropy loss is applied after the linear layer.

4 Experiments

We choose CoNSeP dataset [14] to train our nuclei segmentation network[47] and nuclei feature extractor, then evaluate our proposed method on CRC dataset [4] and a larger dataset Extended CRC dataset [31].

Colorectal cancer(CRC) dataset: The CRC dataset is a commonly used dataset for histology image grading which contains 139 H&E stained histology images with an average size of 4548 $\times$ 7520 at 20$\times$ magnification. Annotated as follows: 71 normal, 33 low grade, and 35 high grade. All 139 images are taken from 37 different patients.

Extended colorectal cancer(Extended CRC) dataset: The Extended CRC dataset is an extension of CRC dataset. It contains 300 images, including 120 normal, 120 low grade, and 60 high grade images, with the size of 4548 $\times$ 7548 and 5000 $\times$ 7300 pixels, extracted at 20$\times$ magnification.

For the above two datasets, we extract patches with a size of 1792 $\times$ 1792 from images to construct cell graphs. In our experiments, we use the majority voting method to generate the image-level prediction. The image-level accuracy metric is used to evaluate our method, which refers to the percentage of correctly classified images.

Colorectal nuclear segmentation and phenotypes (CoNSeP) dataset: The CoNSeP dataset consists of 41 H&E stained image tiles, each of size 1000 $\times$ 1000 pixels at 40$\times$ objective magnification. It has 24319 annotated nuclei with 7 associated classes. We also merge 7 classes into 4 classes[14], then we discard the miscellaneous/others class due to the small number of nuclei, resulting in 3 classes, same as [46].

4.1 Experimental Setup

We normalize all node features according to their mean and standard deviation. We implement our proposed approach based on PyTorch[26] framework. It takes 8 hours on a server equipped with 1 NVIDIA GeForce RTX 3090 GPU to train our model on the CRC dataset.

Dataset Split: We extract the patches of size 1792 $\times$ 1792 with a stride of 224 from images, resulting in 31500 patches for the CRC dataset and 114243 patches for the Extended CRC dataset. We split the CRC dataset and the Extended CRC dataset into 3 folds respectively for cross-validation. To conduct fair experiments, we use the same split as in [46] and [31] to guarantee the consistency of experiment data. Patches extracted from the same case are placed into the same fold, to avoid the patches from the same patient appearing in both the training set and test set, same as[31, 46].

Hyper-parameter: We use Adam optimization[21] in our experiment, setting the learning rate to 0.001 by fine-tuning our learning rate through a grid search. We train our model for 30 epochs with a batch size of 20.

4.2 Ablation Studies

We conduct ablation experiments on the CRC dataset to test the influence of the different components of our proposed approach.

Impact of putting Transformer into different stages: In order to assess the impact of Transformer, we construct an ablation experiment to evaluate the impact of placing the Transformer in different stages. We first train the Hierarchical Graph Neural Network (HANet) without Transformer, then we add Transformer to the 2nd stage, the 3rd stage, and both the 2nd and 3rd stages to verify the effectiveness of the Transformer. The experimental

results can be seen in Tab. 1. Without Transformer, we can only achieve 97.76% for image-level accuracy. If the Transformer is added to a single stage, the image-level accuracy only has a relatively small promotion (0.03%). But if the Transformer is added into both stage-2 and stage-3 at the same time, the image-level accuracy will be greatly improved to 98.55%. The results prove that the Transformer module can learn more efficient node features for the graph neural network.

Network	Stage	Image Accuracy (%)
HA-Net	-	97.76 ± 2.27
+Transformer	2	97.79 ± 2.17
+Transformer	3	97.79 ± 2.17
+Transformer	2,3	$\mathbf{98.55 \pm 1.26}$

Table 1: Ablation study on various stage.

Impact of different ways in extracting node features: The performance of GNN mainly depends on the selection of initial node features [18]. To analyze the influence of the initial node features on the accuracy of grading histology images, we conducted experiments on the different node feature extract methods.

- Hand-crafted extracted node features: As stated in [46], we can extract hand-crafted nuclei features to initialize the graph nodes, the hand-crafted nuclei features contain 16 morphological features, that are mean nuclei intensity, average fore-/background difference, standard deviation of nuclei intensity, skewness of nuclei intensity, mean entropy of nuclei intensity, GLCM of dissimilarity, GLCM of homogeneity, GLCM of energy, GLCM of ASM, eccentricity, area, maximum length of axis, minimum length of axis, perimeter, solidity and orientation. In addition, the nuclei features also use centroid coordinates as spatial features.

- CNN-based extracted node features: In order to obtain more feature information, we use the modified CoNSeP dataset which has the ground truth of nuclei segmentation and classification to fine-tune the ResNet-50 model pretrained by ImageNet. After getting the feature extractor model, we apply the CRC dataset to this model, and we can obtain 2048-dimensional feature data. In order to be consistent with the hand-crafted extracted feature dimensions, we use global average pooling to reduce the dimensions, and finally, obtain 16-dimensional data as the morphological feature of the nuclei. And we also use centroid coordinates of nuclei to represent spatial features.

The experimental results can be seen in Tab. 2. With the hand-crafted extracted node features, we can only achieve 97.04% for image-level accuracy, but the result will be improved by 1.51%, to 98.55%, with the CNN-based extracted node features. This shows that the CNN-based method can obtain more comprehensive morphological information.

Nuclei Features	Image Accuracy (%)
Hand-crafted features	97.04 ± 2.57
CNN-based features	$\mathbf{98.55 \pm 1.26}$

Table 2: Ablation study on the method of extracting node features.

Impact of different pooling technique: We investigate the impact of two stage-of-the-art pooling layers: DIFFPOOL [44] and MinCutPool [6] on the grading accuracy.

The experimental results can be seen in Tab. 3. We can achieve 98.52% for image-level accuracy with DIFFPOOL, and 98.55% with MinCutPool. Therefore, we use MinCutPool as our pooling layer.

Pooling Layer	Image Accuracy (%)
DIFFPOOL	98.52 ± 1.28
MinCutPool	$\mathbf{98.55 \pm 1.26}$

Table 3: Ablation study on various pooling techniques.

Impact of different jumping knowledge technique: In order to adapt to different subgraph structures in the various layer, we use the jumping knowledge technique in our network inspired by [43]. Different jumping knowledge techniques also have a different impact on the experimental results. We use three common jumping knowledge techniques, max-pooling jumping knowledge, concatenation jumping knowledge and LSTM jumping knowledge to explore better experimental models. The experimental results can be seen in Tab. 4. The LSTM-based approach achieves the highest result (98.55%).

Jumping Knowledge Technique	Image Accuracy (%)
Max-pooling	97.76 ± 0.06
Concatenation	97.79 ± 2.17
LSTM	$\mathbf{98.55 \pm 1.26}$

Table 4: Ablation study on various jumping knowledge technique.

How does the nuclei segmentation accuracy affect the overall result: Lower nuclei segmentation accuracy often leads to fewer nuclei being detected. Therefore, we randomly drop p percent number of nuclei in the training and test set to mimic the situation of low nuclei segmentation accuracy rate, then train the network from scratch, to test how the nuclei segmentation accuracy affects the final predictions. Results on the CRC dataset are shown in Tab. 5. We observe a significant performance drop when setting p=50%, showing that our method is more robust with smaller variations (p<=30%) in nuclei segmentation results than with larger (p=50%).

Node Drop Rate (%)	Image Accuracy (%)	Average #Nodes
p=0	**98.55±1.26**	3367
p=10	97.76±0.06	3030
p=30	97.79±2.17	2356
p=50	91.75±0.48	1683

Table 5: Network performance using different node drop rates, Average #Nodes denotes that the average number of nodes in each cell graph during training and test phase.

4.3 Comparison with Existing Methods

We verify the advancement of the method proposed in this paper by comparing it with CRA methods. BAM[4] first segment gland from histology images, then proposed a metric method

named Best Alignment Metric (BAM) to capture the shape difference between the segmented gland and normal gland they expect, BAM-1 computes the average BAM and BAM entropy, BAM-2 additionally computes the Regularity Index. Context-G[33] is a method based on context-aware learning, which first captures the context features by down-sampling from images at different magnification levels, then use a CNN model to classify the histology images. ResNet50 [19], Inception [35], MobileNet [21], and Xception [11] network can be used to train LR-CNN, which can directly do classify tasks. CA-CNN[31] first encodes the local representation of a histology image into high dimensional features, then aggregates the features by considering their spatial organization to make a final prediction. CGC-Net [46] also constructs cell graph first, and it uses a GNN network (Adaptive GraphSage) to classify the graph. Since our experimental data are split in the same way as those in [31] and [46], we directly obtain their experimental results from respective paper for comparison. VIT [13] is a pure Transformer network. To verify that our network is better than a Transformer network, we also use VIT for the grading task both in the CRC dataset and the Extended CRC dataset. The results are shown in Tab. 6. We can observe that our proposed HAT-Net outperforms former SOTA by a large margin in both the CRC dataset and the Extended CRC dataset.

Method	CRC (%)	Extended CRC (%)
BAM-1 [4]	87.79 ± 2.32	-
BAM-2 [4]	90.66 ± 2.45	-
Context-G [33]	89.96 ± 3.54	-
MobileNet [21]	91.37 ± 3.55	84.33 ± 3.30
ResNet50 [19]	92.08 ± 2.08	86.33 ± 0.94
Inception [35]	92.78 ± 2.74	84.67 ± 1.70
Xception [11]	92.09 ± 0.98	86.67 ± 0.94
CA-CNN [31]	95.70 ± 3.04	86.67 ± 1.70
VIT [13]	96.28 ± 3.45	86.67 ± 4.04
CGC-Net [46]	97.00 ± 1.10	93.00 ± 0.93
Ours (HAT-Net)	$\mathbf{98.55 \pm 1.26}$	$\mathbf{95.33 \pm 0.58}$

Table 6: The image accuracy comparison results with other methods on two datasets.

5 Conclusion

Different from nature images, it is impractical to directly train a nucleus feature extractor on histology images since different types of nuclei often cluster together. In this paper, we introduce a MNP approach to enable us to train a strong CNN encoder to extract more morphological information of cell nuclei, to enhance the overall performance. We also propose a new graph neural network combined with Transformer, named Hierarchical Transformer Graph Neural Network (HAT-Net), which exploits the micro-environmental information and the long-range dependency in histology images. We have conducted various experiments on the two datasets, achieved 98.55% and 95.33% classification accuracies on CRC and extended CRC datasets, respectively. Our results outperform the former SOTA by a large margin (+1.55%, +2.33%).

Acknowledgements

This work was supported in part by the Beijing Natural Science Foundation-Haidian Original Innovation Joint Fund Project (No.L182034), the National Natural Science Foundation of China (No.61802022 and No.61802027), and the Fundamental Research Funds for the Central Universities (No.2019XD-A12 and No.2020RC07).

References

[1] Sahirzeeshan Ali, Robert Veltri, Jonathan A Epstein, Christhunesa Christudass, and Anant Madabhushi. Cell cluster graph for prediction of biochemical recurrence in prostate cancer patients from tissue microarrays. In *Medical Imaging 2013: Digital Pathology*, volume 8676, page 86760H. International Society for Optics and Photonics, 2013.

[2] Deepak Anand, Shrey Gadiya, and Amit Sethi. Histographs: graphs in histopathology. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 11320, 2020.

[3] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.

[4] Ruqayya Awan, Korsuk Sirinukunwattana, David Epstein, Samuel Jefferyes, Uvais Qidwai, Zia Aftab, Imaad Mujeeb, David Snead, and Nasir Rajpoot. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Scientific reports*, 7(1):1–12, 2017.

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv: Machine Learning*, 2016.

[6] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International Conference on Machine Learning*, pages 874–883. PMLR, 2020.

[7] W K Blenkinsopp, S Stewart-Brown, L Blesovsky, G Kearney, and L P Fielding. Histopathology reporting in large bowel cancer. *Journal of Clinical Pathology*, 34 (5):509–513, 1981.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[9] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[10] Mark Chen, Alec Radford, Rewon Child, Jeffrey K Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML 2020: 37th International Conference on Machine Learning*, volume 1, pages 1691–1703, 2020.

[11] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2018.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

[15] Fangda Gu, Heng Chang, Wenwu Zhu, Somayeh Sojoudi, and Laurent El Ghaoui. Implicit graph neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 11984–11995, 2020.

[16] Cigdem Gunduz, Bülent Yener, and S Humayun Gultekin. The cell graphs of cancer. *Bioinformatics*, 20(suppl_1):i145–i151, 2004.

[17] Stanley R Hamilton, Lauri A Aaltonen, et al. *Pathology and genetics of tumours of the digestive system*, volume 2. IARC press Lyon:, 2000.

[18] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *arXiv preprint arXiv:1706.02216*, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 2020.

[21] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[22] Guillaume Jaume, Pushpak Pati, Behzad Bozorgtabar, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, 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, 2020.

[23] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[24] Caner Mercan, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. From patch-level to roi-level deep feature representations for breast histopathology classification. In *Medical Imaging 2019: Digital Pathology*, volume 10956, page 109560H. International Society for Optics and Photonics, 2019.

[25] Kamyar Nazeri, Azad Aminpour, and Mehran Ebrahimi. Two-stage convolutional neural network for breast cancer histology image classification. In *International Conference Image Analysis and Recognition*, pages 717–726. Springer, 2018.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NIPS Autodiff Workshop*, 2017.

[27] Pushpak Pati, Guillaume Jaume, Lauren Alisha Fernandes, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Daniel Riccio, Maurizio Di Bonito, et al. Hact-net: A hierarchical cell-to-tissue graph neural network for histopathological image classification. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pages 208–219. Springer, 2020.

[28] Ofir Press, Noah A. Smith, and Omer Levy. Improving transformer models by reordering their sublayers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2996–3005, 2020.

[29] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[31] Muhammad Shaban, Ruqayya Awan, Muhammad Moazam Fraz, Ayesha Azam, Yee-Wah Tsang, David Snead, and Nasir M Rajpoot. Context-aware convolutional neural network for grading of colorectal cancer histology images. *IEEE transactions on medical imaging*, 39(7):2395–2405, 2020.

[32] Rebecca L Siegel, Kimberly D Miller, Ann Goding Sauer, Stacey A Fedewa, Lynn F Butterly, Joseph C Anderson, Andrea Cercek, Robert A Smith, and Ahmedin Jemal. Colorectal cancer statistics, 2020. *CA: a cancer journal for clinicians*, 70(3):145–164, 2020.

[33] Korsuk Sirinukunwattana, Nasullah Khalid Alham, Clare Verrill, and Jens Rittscher. Improving whole slide segmentation through visual context-a systematic study. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 192–200. Springer, 2018.

[34] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[36] Hugo Touvron, Matthieu Cord, Douze Matthijs, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML 2021: 38th International Conference on Machine Learning*, pages 10347–10357, 2021.

[37] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M. Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46, 2021.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[39] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[40] Jingwen Wang, Richard J. Chen, Ming Y. Lu, Alexander Baras, and Faisal Mahmood. Weakly supervised prostate tma classification via graph convolutional networks. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 239–243, 2020.

[41] Manchek A Wong and JA Hartigan. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[42] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[43] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, pages 5453–5462. PMLR, 2018.

[44] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *arXiv preprint arXiv:1806.08804*, 2018.

[45] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24, 2021.

[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 Workshops*, pages 0–0, 2019.

[47] Yanning Zhou, Omer Fahri Onder, Qi Dou, Efstratios Tsougenis, Hao Chen, and Pheng-Ann Heng. Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation. In *International Conference on Information Processing in Medical Imaging*, pages 682–693. Springer, 2019.