Self-Supervised Learning with Graph Neural Networks for Region of Interest Retrieval in Histopathology

Yiğit Özen, Selim Aksoy
Department of Computer Engineering
Bilkent University
Ankara, 06800, Turkey
Email: saksoy@cs.bilkent.edu.tr

Kemal Kösemehmetoğlu, Sevgen Önder, Ayşegül Üner
Department of Pathology
Hacettepe University
Ankara, 06100, Turkey

***Abstract*—Deep learning has achieved successful performance in representation learning and content-based retrieval of histopathology images. The commonly used setting in deep learning-based approaches is supervised training of deep neural networks for classification, and using the trained model to extract representations that are used for computing and ranking the distances between images. However, there are two remaining major challenges. First, supervised training of deep neural networks requires large amount of manually labeled data which is often limited in the medical field. Transfer learning has been used to overcome this challenge, but its success remained limited. Second, the clinical practice in histopathology necessitates working with regions of interest (ROI) of multiple diagnostic classes with arbitrary shapes and sizes. The typical solution to this problem is to aggregate the representations of fixed-sized patches cropped from these regions to obtain region-level representations. However, naive methods cannot sufficiently exploit the rich contextual information in the complex tissue structures. To tackle these two challenges, we propose a generic method that utilizes graph neural networks (GNN), combined with a self-supervised training method using a contrastive loss. GNN enables representing arbitrarily-shaped ROIs as graphs and encoding contextual information. Self-supervised contrastive learning improves quality of learned representations without requiring labeled data. The experiments using a challenging breast histopathology data set show that the proposed method achieves better performance than the state-of-the-art.**

***Index Terms*—Digital pathology, histopathological image analysis, self-supervised learning, graph neural networks, content-based image retrieval**

I. Introduction

Histopathology image analysis aims to serve as an important tool for helping pathologists with the diagnostic process. It can relieve the workload on pathologists and offer more objective analysis of histopathology images. In addition to the classifier systems providing diagnostic predictions or grading scores, content-based image retrieval (CBIR) has also been investigated for decision support in many clinical applications [1]–[4]. Given an image database, CBIR methods aim to retrieve images with morphological characteristics most relevant to and consistent with the query image. CBIR can also be used for classification purposes by considering the majority diagnosis of the retrieved images as the most likely diagnosis.

In the general CBIR pipeline, feature extraction methods are employed to represent each image with a feature vector. Depending on the feature extraction method, the image representations can be directly compared for similarity, or a ranking-based model can be learned on top of the image features. In the query phase, the learned or constructed similarity model is used to retrieve the most similar images. The retrieval results can be provided to users for further analysis.

Although the current CBIR methods achieved successes in generic image retrieval problems, how to tackle the retrieval in histopathology image databases is still a challenging topic [5]. The size of histopathology images can be extremely large. For example, the whole slide images (WSI) that are obtained by digitizing biopsy slides at high magnification can include more than $100{,}000 \times 100{,}000$ pixels. WSIs often contain many regions of interest (ROI) that can belong to different diagnostic categories and can carry different levels of relevance for the final slide-level diagnosis. Furthermore, the pathologists do not have any restrictions on the ROI shapes and sizes when they evaluate the slides, and can select and study the regions at any size and magnification deemed suitable. The complex imaging parameters (e.g., staining procedures, machine properties), microanatomic differences, and interactions between different structures result in a more complex analysis compared to natural images. The relevant changes of histopathology images require both cell-level and contextual analysis.

Earlier works used hand-crafted features, most notably the bag-of-features based on SIFT descriptors [6], to represent histopathology images, and focused on similarity measures [3], [4], [7]. We believe tackling these challenges requires more effective representation learning methods. In recent years, deep learning-based approaches, in particular convolutional neural networks (CNN), have been shown to be successful in visual representation learning in various domains including digital pathology [8]. As the mainstream CNN architectures typically require fixed-sized inputs, their common use in the digital pathology domain has also been in the classification of fixed-sized histopathology image patches. The generally studied setting has been to aggregate the feature representations of

fixed-sized patches cropped from these images to obtain an image-level representation. Aggregation methods typically include fixed rules such as averaging features or class scores, or weighted averaging with learnable weights [9]–[12]. More recently, graph neural network (GNN) architectures that deal with size and shape variation of ROIs and encode contextual information via message passing [13], [14] are used instead of CNN. GNN-based methods formulate the ROI classification problem as a graph classification problem. Regardless of the model architecture, after the training phase of the classification model, the "head", i.e., the last one or more fully-connected layers, is removed and the remaining model is used as a feature representation extractor. Then, the extracted features are used for CBIR.

State-of-the-art results have been achieved using representation learning through classification. However, their power is bounded by the amount of manually labeled training data. Annotation of histopathology images by expert pathologists is a costly operation. Better performance can be achieved using deeper and wider neural networks but training larger models in a supervised setting requires more labeled data. Training deep neural networks from scratch using small amount of labeled data can easily result in overfitting [15]. Pre-training the models on other settings and fine-tuning on target histopathology images, and unsupervised neural networks such as auto-encoders have also been explored, but their success in large-scale images remain limited [16]–[18].

In this paper, we propose a self-supervised method to learn visual representations of arbitrarily-shaped ROIs without reference diagnostic information. The method employs a GNN that models each ROI as a graph where vertices denote the patches sampled from the ROI and edges represent the spatial proximity of those patches. A self-supervised contrastive learning method recently proposed by [19] and achieved the state-of-the-art results in natural image data sets is adapted to histopathology images. The graph structure implicitly encodes the spatial relationships across the patches, which can be used to tackle fine-grained representation learning in a holistic manner. Self-supervised learning improves the quality of learned representations while allowing utilization of large amounts of unlabeled histopathology image data. Our experimental results show that the proposed method performs better than the state-of-the-art.

In the following, we first introduce the breast pathology data set used in the paper, then, describe the proposed method, and finally, present the experimental results.

II. Data Set

We constructed a new breast histopathology data set using 78 WSIs that were digitized from specimens collected from 63 patients. The haematoxylin and eosin stained specimens were selected from the archives of the Department of Pathology at Hacettepe University based on their slide-level diagnoses. The WSIs were acquired at $40\times$ magnification by using an Olympus slide scanner. The resulting average image size was $170{,}000 \times 132{,}000$ pixels. $1{,}030$ ROIs were annotated by experienced pathologists in free form with no shape and size restrictions. The resulting annotations were collected into 4 diagnostic classes: *benign* (including samples containing non-proliferative changes, apocrine metaplasia, usual ductal hyperplasia, columnar cell hyperplasia, flat epithelial hyperplasia, and intraductal papilloma without atypia), *atypia* (including samples containing atypical ductal hyperplasia, atypical lobular hyperplasia, and intraductal papilloma with atypia), *in situ* carcinoma (including both ductal carcinoma in situ and lobular carcinoma in situ), and *invasive* carcinoma. The per-class ROI size statistics are shown in Table I.

TABLE I
ROI size statistics per diagnostic class in number of pixels at $10\times$ magnification. Rows show the average ROI size, the standard deviation of ROI sizes, and the ratio of the largest ROI size to the smallest one, respectively.

	Benign	Atypia	In Situ	Invasive
Average	$1308K$	$473K$	$2815K$	$12568K$
Standard deviation	$2510K$	$711K$	$4948K$	$17822K$
Max-min ratio	977.2	210.8	941.1	762.5

Since the specimens were stained at different times following different procedures, there is a great variation in their color distributions. To eliminate the effects of staining differences in model learning, we performed stain normalization as a pre-processing step. We first applied color deconvolution [20] to estimate the stain matrix of each slide. To make the estimation more robust, haematoxylin stain vector estimation considered only the pixels inside nucleus masks which were automatically generated using a pre-trained convolutional network, and eosin stain vector estimation considered the remaining regions excluding high luminosity regions which correspond to background. Then, the histograms of haematoxylin and eosin channels of each slide are matched to a target slide chosen from the data set [21].

Finally, we partitioned the data set into four folds by using ROI-level diagnosis labels. The split is constrained to make all WSIs and ROIs from the same patient fall under the same fold. We employed a genetic algorithm to find a good split that achieves similar slide-level and ROI-level class distributions among the folds. During model learning, we use two folds as the training set, one fold as the validation set, and one fold as the test set. The assignment of folds to data subsets are random. The resulting ROI-level and slide-level class distributions of the three sets are given in Table II.

III. Proposed Method

In our framework, the regions of interest (ROI) with arbitrary shapes are represented by undirected graphs where vertices correspond to fixed-size patches and edges correspond to spatial proximity relation between the patches. We propose learning a graph neural network (GNN) that encodes ROI graphs into representations using a contrastive loss function in a self-supervised setting. Then, a content-based retrieval system is constructed using the trained GNN to extract representations from ROIs and Euclidean distance between the extracted representations to measure the similarity of ROIs.

TABLE II
CLASS DISTRIBUTION OF SLIDES AND ROIS IN TRAINING, VALIDATION, AND TEST SETS. NOTE THAT A SLIDE CAN CONTAIN MULTIPLE ROIS CORRESPONDING TO DIFFERENT DIAGNOSTIC LABELS, RESULTING IN A MULTI-LABEL SETTING FOR EACH SLIDE. THUS, THE NUMBERS OF SLIDES FOR EACH DIAGNOSTIC CLASS IN THE TABLE DO NOT SUM UP TO THE TOTAL NUMBER OF SLIDES FOR A GIVEN SET.

		Benign	Atypia	In Situ	Invasive	Total
Slide	Training Set	30	16	16	13	39
	Validation Set	15	7	8	6	18
	Test Set	16	8	9	6	21
	Total	61	31	33	25	78
ROI	Training Set	226	55	154	102	537
	Validation Set	109	25	56	50	240
	Test Set	105	30	69	49	253
	Total	440	110	279	201	1030

A. Graph Construction

First, patches of size 224×224 are extracted from the ROI. We use scrambled Sobol sequences for dense and low discrepancy sampling of 2-dimensional coordinates that correspond to the centers of the patches. If more than half of the pixels of a patch are within the ROI, the patch is included in the graph. Using low discrepancy sampling introduces irregularity to the distribution while preventing large gaps between patch clusters that can occur in random sampling. Then, patch features are extracted using a ResNet-50 model [22] trained on ImageNet, although any suitable architecture including unsupervised models can be used in this step. A vertex for each patch is added to the ROI graph where the position of the vertex is the center coordinates of the patch relative to the region and the feature representation of the vertex is the extracted feature vector. An edge is added to the graph between two vertices if the distance between their positions is less than or equal to 448, i.e., twice the patch size.

B. Region of Interest Representation Learning

We adopt the SimCLR framework [19] for learning representations of ROIs. SimCLR has recently achieved state-of-the-art results on ImageNet with an architecture simpler than its alternatives such as Contrastive Predictive Coding [23].

SimCLR framework comprises a stochastic data augmentation module, a neural network encoder, a neural network projection head, and the normalized temperature-scaled cross entropy loss as the contrastive loss function. The data augmentation module takes as input a data sample, and outputs two different augmented views of this example. The encoder extracts feature vectors from the augmented data points. Finally, the projection head maps the feature vectors to the space where contrastive prediction loss is calculated.

During the end-to-end training of the model, a random mini-batch of size M is fed to the augmentation module, resulting in $2M$ data points. Given the positive pair of a data point which originated from the same example, the other $2(M-1)$ data points are treated as negatives. The loss function between a positive pair of data points (i, j) is defined as

$$l_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\Sigma_{k=1}^{2M} \mathbb{I}_{[k \neq i]} \exp(sim(z_i, z_k))/\tau} \tag{1}$$

where $\mathbb{I}$ is the 0-1 indicator function, z_i, z_j are the outputs of the prediction head, sim is the cosine similarity function, and τ is the adjustable temperature parameter. The loss function in (1) is computed across all positive pairs in a mini-batch.

In our application, the data augmentation module removes a random subset of vertices from the ROI graph. The resulting two views represent the same ROI with different graph structures and vertex features. A GNN is employed as the encoder. As the projection head, we use a 2-layer multi-layer perceptron (MLP) with ReLU nonlinearity. This way, the GNN encoder is forced to learn high-level contextual features to maximize the agreement between the two views. The process is illustrated in Figure 1.

C. Graph Neural Network Architectures

Three common building blocks of GNN models are neighborhood aggregation, local pooling, and global pooling [24]. Neighborhood aggregation enables encoding contextual information, local pooling makes the learned representations hierarchical similarly to the pooling in convolutional neural networks, and global pooling aggregates vertex representations into a graph representation. In its general form, neighborhood aggregation can be defined as

$$X' = \Phi(A, X; \Theta) \tag{2}$$

where $X \in \mathbb{R}^{N \times C}$ is the input feature matrix of N vertices with C input channels, A is the adjacency matrix, Θ is the set of trainable parameters, and $X' \in \mathbb{R}^{N \times F}$ is the output feature matrix with F channels.

In our method, we consider three architectures: stacked graph convolutional network (GCN) [25] followed by global pooling, DiffPool [26], and GraphConv [27].

1) GCN-based Architecture: Our GCN-based architecture is flat, i.e., does not have local pooling that constructs a hierarchy. A single layer of GCN is defined as

$$X' = ReLU(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X \Theta) \tag{3}$$

where $\Theta \in \mathbb{R}^{C \times F}$ is the matrix of filter parameters, X' is the convolved output, $\tilde{A} = A + I_N$ is the adjacency matrix with inserted self-loops, and $\tilde{D}_{ii} = \Sigma_j \tilde{A}_{ij}$. A number of GCN layers are stacked to enlarge the receptive field of the neurons. Finally, vertex features are averaged across the vertex dimension to produce a single graph feature vector G as

$$G = \frac{1}{N} \sum_{i=1}^{N} X'_i. \tag{4}$$

2) DiffPool-based Architecture: DiffPool learns differentiable dense cluster assignments for vertices at each layer, mapping each vertex to a cluster. Each cluster becomes the input vertex for the next layer. The pooling operation is defined as

$$X' = softmax(S)^T \cdot X \tag{5}$$

$$A' = softmax(S)^T \cdot A \cdot softmax(S) \tag{6}$$

where S is the learned assignments and A' is the coarsened adjacency matrix. The number of clusters in the pooling

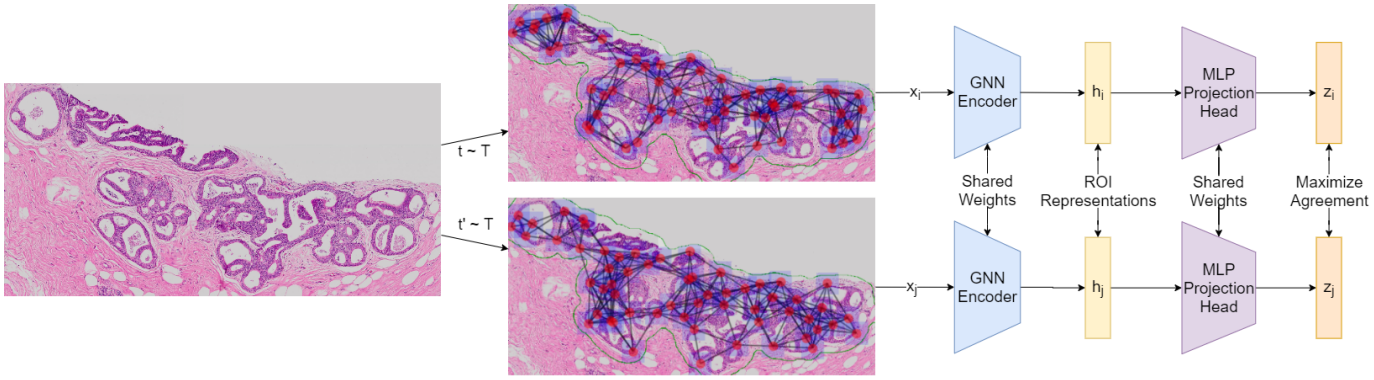

Fig. 1. The SimCLR framework applied to a region of interest in a breast biopsy image. Two separate vertex dropout augmentations ($t \sim T$ and $t' \sim T$) are applied to obtain two separate views of the same ROI. The GNN encoder and the MLP projection head are trained to maximize the agreement between the representations using the contrastive loss. After training is completed, the GNN encoder is used to extract ROI representations for the retrieval task.

operation must be predetermined. At the end, features of final clusters are averaged across the cluster dimension to produce a single graph feature vector G as given in (4).

3) GraphConv-based Architecture: The GraphConv-based architecture consists of the top-k pooling operator where vertices are dropped based on a learnable projection score as described in [28]. The neighborhood aggregation function for vertex i is defined as

$$X_i' = \Theta_1 X_i + \sum_{j \in N(i)} \Theta_2 X_j \tag{7}$$

where Θ contains the learned weights and $N(i)$ is the set of neighbors of vertex i. The aggregation followed by pooling is repeated k times which forms k subgraphs. Then, the features from k subgraphs are produced by mean and max-pooling of their vertex features. Finally, the subgraph features are concatenated and fed to a 2-layer MLP as described in [27]. The output of the MLP forms the graph representation.

IV. Experiments

A. Model Configurations

The GCN-based architecture has two layers of GCN convolutions with ReLU activation. The DiffPool-based architecture has two pooling operations that divide the network into three hierarchical layers. The first two layers have one subnetwork for vertex features and one subnetwork for cluster assignments. The last layer has only one subnetwork for vertex features. Each subnetwork consists of two consecutive SAGE convolution layers [29]. The GraphConv-based model has three hierarchical layers, each with one graph convolution and one top-k pooling. All models are trained using a variant of Adam optimizer with weight decay proposed in [30]. The parameters of the optimizer, the temperature parameter in the loss function, the hidden layer sizes, and pooling ratios are determined through hyperparameter optimization based on the validation set performance. The best values are chosen independently for each model.

B. Evaluation

The proposed method is compared to the supervised classification learning method which is the state-of-the-art. For a fair comparison, the vertex dropout augmentation is applied to the training set in supervised learning. Similar to the proposed method, a two-layer MLP that follows the GNN encoder and outputs four class scores is used in the classification method. In all methods, the output of the GNN encoder is used as the ROI representation for content-based image retrieval.

In all experiments, the same training, validation, and test sets are used for model training, hyperparameter optimization through random search and model selection, and retrieval performance evaluation, respectively. The training and validation sets are included neither in the gallery nor in the query sets in the CBIR setup for a realistic evaluation.

The test set is randomly split into query and gallery subsets according to 1-to-4 ratio and the same split is used in all experiments. For each ROI in the query set, regions in the gallery set are retrieved in a ranked order based on the Euclidean distance between their representations. Regions that have the same class label with the query are considered relevant results while others are considered irrelevant. Mean Average Precision (mAP) metric is used for quantitative evaluation of retrieval performance. Average Precision (AP) is the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight as

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{8}$$

where P_n is the precision and R_n is the recall at the n'th threshold. Then, the mAP is defined as the mean of average precisions over all queries as

$$mAP = \frac{1}{Q} \sum_{q=1}^{Q} AP(q). \tag{9}$$

AP@K is defined as the AP calculated using the top K retrieved items. MAP@K is calculated using AP@K. Considering the size of the test set, MAP@10 and MAP@25 are calculated and reported in the experiments. Combinations of

TABLE III
ROI RETRIEVAL RESULTS FOR DIFFERENT METHODS AND TRAINING SETTINGS.

Method	Supervision	Architecture	MAP@10	MAP@25
GNN-LR [13]	Supervised	DiffPool	0.62	0.59
		GraphConv	0.73	0.64
		GCN	0.80	0.76
Ours	Self-Supervised	DiffPool	0.78	0.70
		GraphConv	0.82	0.75
		GCN	**0.86**	**0.80**

different GNN architectures and training settings are compared. The quantitative results are summarized in Table III. Example retrieval results are presented in Figure 2. GNN-LR [13] is the state-of-the-art method considered for classification-based learning. The only modification we made is to use dense patch sampling instead of non-overlapping sliding window approach used in the paper. The dense sampling method performed better than the sliding window approach in both supervised and self-supervised settings.

C. Discussion

Overall, the proposed self-supervised learning method achieved the best performance despite not utilizing class labels. This result is consistent with our view that contrastive learning can improve the quality of learned representations.

GCN architecture performed better then DiffPool and GraphConv in our experiments, although DiffPool performs better in popular GNN benchmark data sets [31]. We argue that due to the large variation of graph sizes in our data set, it is difficult to choose cluster sizes in the DiffPool architecture that are meaningful for all ROI graphs and this degrades its performance. In contrast, the top-k pooling operations in the GraphConv architecture require ratio hyperparameters instead of fixed numbers. A simpler architecture like our GCN-based architecture still encodes contextual information through neighborhood aggregation. Therefore, the choice of GNN architecture should be treated as a hyperparameter in the proposed learning method.

V. CONCLUSIONS

In this paper, we proposed a novel histopathology region of interest retrieval learning method. The regions are represented by graphs, and graph neural networks are trained using contrastive loss. The proposed method, without using class labels, has achieved better retrieval performance in a realistic breast histopathology data set than its alternatives that use the same amount of data with class labels. Thus, the method allows utilizing the vast amount of unlabeled histopathology image data. The method enables cheaply training histopathology retrieval systems that can process arbitrarily-shaped queries.

ACKNOWLEDGMENT

This work was supported in part by the Scientific and Technological Research Council of Turkey (grant 117E172).

REFERENCES

[1] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applicationsclinical benefits and future directions," *International Journal of Medical Informatics*, vol. 73, no. 1, pp. 1–23, 2004.

[2] H. C. Akakin and M. N. Gurcan, "Content-based microscopic image retrieval system for multi-image queries," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 758–769, 2012.

[3] X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang, "Towards Large-Scale Histopathological Image Analysis: Hashing-Based Image Retrieval," *IEEE Transactions on Medical Imaging*, vol. 34, no. 2, pp. 496–506, Feb. 2015.

[4] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, Y. Ma, H. Shi, and Y. Zhao, "Histopathological Whole Slide Image Analysis Using Context-Based CBIR," *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1641–1652, Jul. 2018.

[5] Z. Li, X. Zhang, H. Mller, and S. Zhang, "Large-scale retrieval for medical image analytics: A comprehensive review," *Medical Image Analysis*, vol. 43, pp. 66–84, Jan. 2018.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, Y. Ma, H. Shi, and Y. Zhao, "Size-Scalable Content-Based Histopathological Image Retrieval From Database That Consists of WSIs," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1278–1287, Jul. 2018.

[8] Litjens, G., et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, December 2017.

[9] K. Roy, D. Banik, D. Bhattacharjee, and M. Nasipuri, "Patch-based system for classification of breast histology images using deep learning," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 90–103, 2019.

[10] K. Das, S. P. K. Karri, A. G. Roy, J. Chatterjee, and D. Sheet, "Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification," in *IEEE International Symposium on Biomedical Imaging*, 2017, pp. 1024–1027.

[11] Y. Feng, L. Zhang, and J. Mo, "Deep manifold preserving autoencoder for classifying breast cancer histopathological images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[12] M. Y. Lu, R. J. Chen, J. Wang, D. Dillon, and F. Mahmood, "Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding," *arXiv preprint arXiv:1910.10825*, 2019.

[13] Y. Zheng, B. Jiang, J. Shi, H. Zhang, and F. Xie, "Encoding histopathological wsis using gnn for scalable diagnostically relevant regions retrieval," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 550–558.

[14] J. Wang, R. J. Chen, M. Y. Lu, A. Baras, and F. Mahmood, "Weakly supervised prostate tma classification via graph convolutional networks," *arXiv preprint arXiv:1910.13328*, 2019.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] T. Schlegl, J. Ofner, and G. Langs, "Unsupervised pre-training across image domains improves lung tissue classification," in *International MICCAI Workshop on Medical Computer Vision*. Springer, 2014, pp. 82–93.

[17] B. Kieffer, M. Babaie, S. Kalra, and H. R. Tizhoosh, "Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks," in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2017, pp. 1–6.

[18] L. Hou, V. Nguyen, A. B. Kanevsky, D. Samaras, T. M. Kurc, T. Zhao, R. R. Gupta, Y. Gao, W. Chen, D. Foran *et al.*, "Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images," *Pattern Recognition*, vol. 86, pp. 188–200, 2019.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *arXiv:2002.05709 [cs, stat]*, Feb. 2020, arXiv: 2002.05709. [Online]. Available: http://arxiv.org/abs/2002.05709

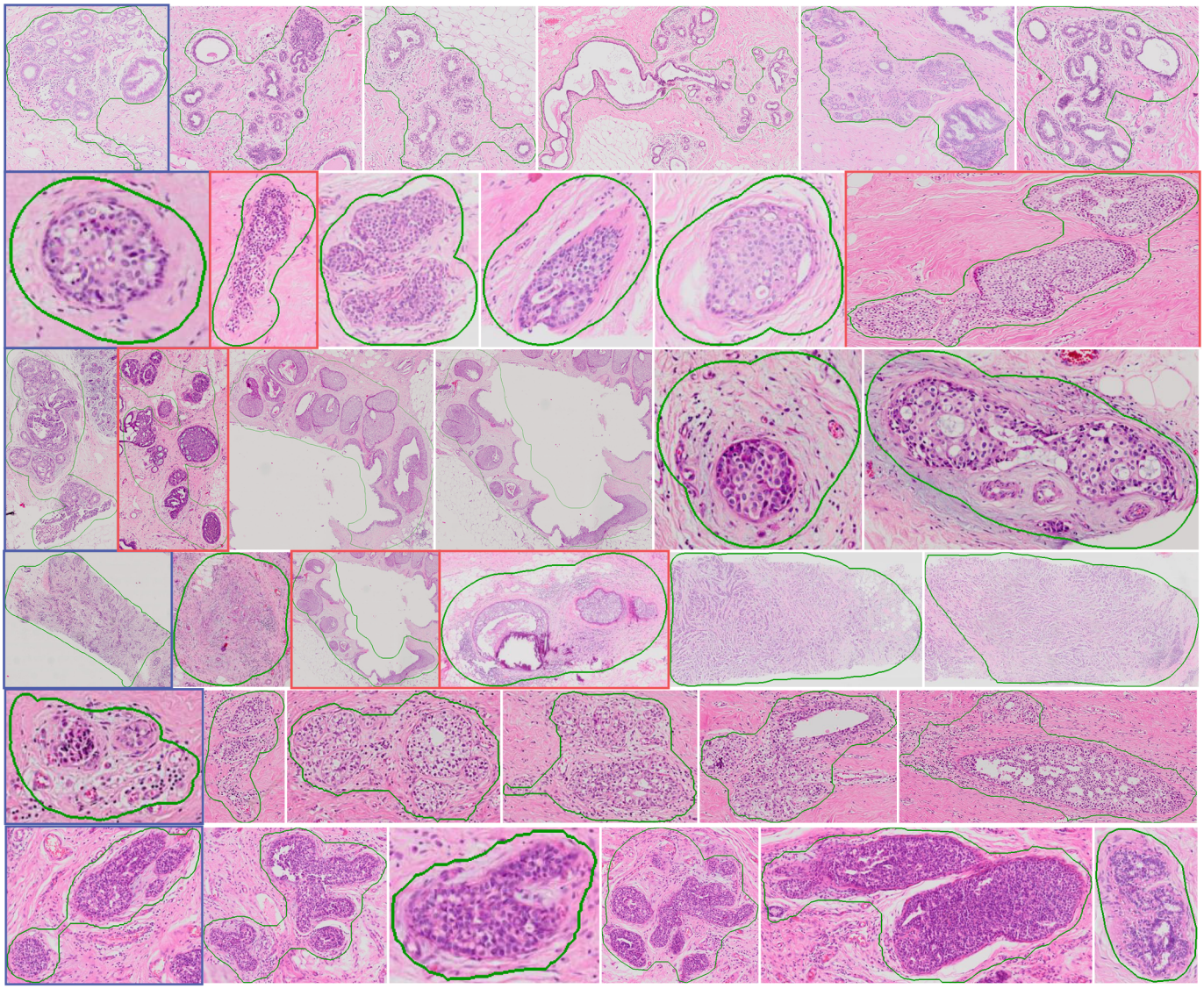

Fig. 2. Examples of ROI retrieval using the best model, showing the top 5 retrieved items for each query in separate rows. Query images are marked in blue. Among the retrieved images, the irrelevant ones are marked in red. The ROI boundaries are marked in green. Query ROI classes from top to bottom are: benign, in-situ, in-situ, invasive, atypia, and benign.

[20] A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology*, vol. 23, no. 4, pp. 291–299, 2001.

[21] A. Basavanhally and A. Madabhushi, "EM-based segmentation-driven color standardization of digitized histopathology," in *SPIE Medical Imaging Symposium, Digital Pathology Conference*, vol. 8676, 2013.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[23] O. J. Hnaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. v. d. Oord, "Data-Efficient Image Recognition with Contrastive Predictive Coding," *arXiv:1905.09272*, Dec. 2019. [Online]. Available: http://arxiv.org/abs/1905.09272

[24] D. Bacciu, F. Errica, A. Micheli, and M. Podda, "A Gentle Introduction to Deep Learning for Graphs," *arXiv:1912.12693*, Dec. 2019. [Online]. Available: http://arxiv.org/abs/1912.12693

[25] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *arXiv:1609.02907 [cs, stat]*, Feb. 2017, arXiv: 1609.02907. [Online]. Available: http://arxiv.org/abs/1609.02907

[26] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical Graph Representation Learning with Differentiable Pooling," *arXiv:1806.08804 [cs, stat]*, Feb. 2019, arXiv: 1806.08804. [Online]. Available: http://arxiv.org/abs/1806.08804

[27] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, "Weisfeiler and leman go neural: Higher-order graph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4602–4609.

[28] B. Knyazev, G. W. Taylor, and M. Amer, "Understanding attention and generalization in graph neural networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 4204–4214.

[29] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," *arXiv:1706.02216 [cs, stat]*, Sep. 2018, arXiv: 1706.02216. [Online]. Available: http://arxiv.org/abs/1706.02216

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[31] F. Errica, M. Podda, D. Bacciu, and A. Micheli, "A fair comparison of graph neural networks for graph classification," *arXiv preprint arXiv:1912.09893*, 2019.