How GNNs Facilitate CNNs in Mining Geometric Information from Large-Scale Medical Images

Yiqing Shen, Bingxin Zhou, Xinye Xiong, Ruitian Gao, Yu Guang Wang

1Johns Hopkins University, Baltimore, MD, USA
2Shanghai Jiao Tong University, Shanghai, China

Abstract—Gigapixel medical images are a rich source of information containing both morphological textures and spatial information. However, existing deep learning solutions primarily rely on convolutional neural networks (CNNs) for global pixel-level analysis, ignoring the underlying local geometric structure. Since the topological structure in medical images is closely related to tumor evolution, graphs can be utilized to characterize it. To obtain a more comprehensive representation for downstream analysis, a fusion framework is proposed to enhance the global image-level representation captured by CNNs with the geometry of cell-level spatial information learned by graph neural networks (GNN). Two fusion strategies have been developed: one with MLP, which is simple but efficient through fine-tuning, and the other with TRANSFORMER, which excels in fusing multiple networks. The proposed fusion strategies have been evaluated on histology datasets from large patient cohorts of colorectal and gastric cancers for three biomarker prediction tasks. Both models outperform plain CNNs or GNNs, achieving a consistent AUC improvement of more than 5% on various network backbones. Importantly, the experimental results demonstrate the necessity of combining image-level morphological features with cell spatial relations in medical image analysis. Codes are available at https://github.com/yiqings/HEGnnEnhanceCnn.

Index Terms—Graph Neural Networks, Convolutional Neural Networks, Histology, Model Fusion.

I. INTRODUCTION

Large-scale medical images, such as histology, provide a wealth of complex patterns for deep learning algorithms to mine. The standard practice for feature extraction from these images is to train convolutional neural networks (CNNs), which focus on morphology and texture. These networks have been employed in various diagnostic and prognostic tasks, such as gene mutation identification, molecular biomarker classification from Hematoxylin and Eosin (H&E) stained histology whole-slide images (WSIs) [1]. However, due to the convolutional kernel’s limited capability to analyze fixed connectivity between local areas, these approaches only extract geometric behaviors are crucial for downstream diagnostics. To this end, we propose an efficient strategy to integrate structure features from CNNs with image features from CNNs for H&E slide analysis. To fuse the graph-level representation learning with image-level embeddings, we train the GNN in parallel with the CNN. This integration takes place in a learnable fusion layer that incorporates the morphology features of the image with the geometric representation of cell graphs. In this way, insights into the spatial structure are gained for a specific staining image, such as the distribution of cells, interaction of cancer and healthy cells, and tumor microenvironment. In summary, the contributions are three-fold. (1) We develop two fusion schemes, based on MLP and TRANSFORMER, for integrating the features extracted from CNNs and GNNs. Moreover, we present and adapt the proposed framework on histology analysis, where cell geometric behaviors are crucial for downstream diagnostics. (2) Experiments on three real-world and one synthetic datasets yield that geometric and image-level representations are complementary. (3) We release the constructed graph-image paired datasets, which can serve as a benchmark for future research of image-graph bimodality.

II. FUSING CNN WITH GNN

A. Geometric Feature Representation

We denote the corresponding graph (e.g., cell graph in histology) to the image $X_G$ as $G = (V, E)$, where $V$ is the collection of nodes, $E$ is the set of all edges $e_{ij}$ with the attribute $w_{ij}$ describing the pair-wise node interaction. For notation simplicity, we use a matrix pair $(X_G, A_G)$ to represent the node attributes and weighted edges, respectively. We name
$A_G$ as an adjacency matrix with its element $(A_G)_{ij} = w_{ij}$. The $\ell$th layer of the GNN finds the hidden representation of the graph by $H_G^{(\ell)} = \text{ReLU}(\text{GraphConv}(A_G, H_G^{(\ell-1)}))$, where $H_G^{(0)} = X_G$. We consider spatial-based convolutions for GraphConv, which usually follow the message-passing [15] form. For the $i$th node of a graph, its representation $H_{g,i}^{(\ell)}$ at the $\ell$th convolutional layer reads $H_{g,i}^{(\ell)} = \gamma \left( \sum_{j \in \mathcal{N}(i)} \phi(H_{g,j}^{(\ell-1)}, H_{g,j}^{(\ell-1)}, A_{ij}) \right)$ with some differentiable operators $\gamma, \phi$ (e.g., MLP) and permutation invariant aggregation function $\Box$ (e.g., average or summation). The $H_{g,j}^{(\ell-1)}$ denotes node $j$’s hidden representation at the $(\ell - 1)$th layer, where $j \in \mathcal{N}(i)$ is a 1-hop neighbor of node $i$, i.e., $A_{ij} = A_{ji} \neq 0$. The representation $H_{g}^{(\ell)}$ embeds spatial topological structures of the underlying graph, which is usually sent to a readout layer, such as a linear layer, before eventually being fed into the fusion layer. We term this linear layer as the alignment layer, which helps to align the feature dimensions of the GNN with the parallel CNN output.

![Diagram](image.png)

Fig. 1. A schematic illustration for fusing the representations of local geometry and global image features. An image input $X_I$ is encoded to a global image-level presentation $H_I$ by CNNs. Meanwhile, the geometric information $(X_G, A_G)$ from the image, which is first transformed into a cell graph with attributes, is embedded by a GNN (i.e., $H_G$). The two sets of hidden representations are then fed into a fusion layer for adaptive integration (i.e., $H_o$). The output $H_o$ is eventually sent to a prediction head (e.g., classification or regression head) for training assignment.

### B. Image-level Feature Representation

The image-level feature representation is directly extracted from histology patches by CNNs. For instance, denote $\{H_I^{(1)}, \ldots, H_I^{(\ell-1)}\}$ the output of the first $(\ell - 1)$ blocks after convolution layers. We can use different convolutional modules for the CNN. For all CNN blocks, we assign the input feature $H_I^{(0)}$ by staining normalized histology image patches. In the same fashion as geometric feature representation, the final image representation is fed to a learnable fully connected layer to adjust the embedding feature dimensions.

### C. Learnable Feature Fusion Layer

Denote the output representations from image and graph as $H_I \in \mathbb{R}^{d_I}$ and $H_G \in \mathbb{R}^{d_G}$. We then train the fusion layer to learn the optimal integration between them. In particular, we consider two candidate structures: MLP and TRANSFORMER [16] for feature fusion, as shown in Fig. 1. The former scheme approaches the fused representation $H_o$ with MLPBlocks formulated as $H_o = \text{MLPBlock}(\ldots(\text{MLPBlock}(H_i)))$, where $H_e = \text{concat}(H_I, H_G) \in \mathbb{R}^{d_I+d_G}$, and

$$MLPBlock(H) = \text{Dropout}(\text{ReLU}(\text{Linear}(H))).$$

The TRANSFORMER fusion scheme formulates $H_o$ by $H_o = \text{Pooling}(\text{Transformer}(\ldots(\text{Transformer}(H_i))))$, where $H_e = \text{stack}(H_T, H_G) \in \mathbb{R}^{2x d}$. The stack operation requires an identical dimension of $H_T$ and $H_G$, thus feature shape alignment with additional linear layer is required i.e., $d = d_T = d_G$.

### D. Cell Graph Construction

#### Cell Graph

We establish a cell graph for each patch image as visualized in Fig. 2. The graph nodes ($v_i$, with its subscript $i$ representing the node index) in a cell graph are biologically determined by the nuclei regions. As the morphological signals are believed relative to cell-cell interplay, the cell-specific features $X_G$, which include the nuclei coordination, optical, and representations, then characterize the cell-level morphological behavior. We then calculate the pair-wise Euclidean distance between nuclei centroids to establish edges of a cell graph [7] to quantify the interplay between cells in a patch. To be precise, for arbitrary two nuclei nodes $v_i$ and $v_j$, with their associated centroid Cartesian coordinates $(x_i, y_i)$ and $(x_j, y_j)$, the edge weight $w_{ij}$ for the interaction between two nodes reads

$$w_{ij} := \begin{cases} 
\frac{d_c}{d(v_i, v_j)}, & d(v_i, v_j) \leq d_c \text{ pixels}, \\
0, & \text{otherwise},
\end{cases}$$

where $d(v_i, v_j)$ regards the Euclidean distance between $v_i$ and $v_j$. From the clinical observations, two cells do not exert mutual influence with their centroid distance exceeding $d_c$ [3]. Thus, the critical distance $d_c$ depicts the range where a cell can interact with another. Note that the precise value of $d_c$ depends on the tissue structure, image category, and magnification of the WSI. An edge $e_{ij}$ exists between $v_i$ and $v_j$ if and only if the weight $w_{ij} > 0$.

![Image](image.png)

Fig. 2. Visualization of the segmented cells and the generated graphs from an arbitrary patch sample of GIST-PDL1. The four subgraphs from left to right are the raw patch image, the segmented cells masks, the patch image with overlaid segmentation masks, and the generated graph.

![Image](image.png)

Fig. 3. An illustrative example of annotated histology patch from GIST-PDL1 for training the nuclei segmentation network. The four subgraphs from left to right are the raw patch image, the generated semantic nuclei masks, the generated semantic nuclei edge, and the annotated instance nuclei mask (ground truth).
Node Feature Extraction. In the node construction process, we employ CA$^2$-Net [17]–[19] as the backbone for nuclei segmentation. For illustration purposes, we pick one annotated sample and show it in Fig. 3. For an arbitrary patch, a graph is generated where nodes represent cells and the weighted edges reveal the Euclidean distance between nodes. Next, we select 94 features from pathomics, i.e., a pre-defined feature library for medical image analysis [20] that describes the location, first-order statistics, and the gray-level textural features of each segmented cell. The proposed GNN and CNN fusion scheme for the downstream patient-level prediction from histology is presented in Fig. 4.

III. EXPERIMENTS

A. Dataset

CRC-MSI and STAD-MSI. The two public datasets focus on the prediction of distinguishing the microsatellite instability (MSI) from microsatellite stability (MSS) in H&E stained histology. In particular, CRC-MSI contains H&E stained histology slides of 315 colorectal cancer patients, and STAD-MSI includes H&E slides of 360 gastric cancer patients. For both datasets, a WSI concerning a patient is tessellated into non-overlapped patches/images with a resolution of $224 \times 224$ pixels at the magnification of $20 \times$. The patches from 70% patients are used for training and the remaining patches from 30% patients are left for validation.

GIST-PDL1. This in-house dataset predicts programmed death-ligand 1 (PD-L1) status from gastric cancer histology slides. This dataset collects 129 well-annotated H&E stained histology slides of gastric cancer patients between the year 2020 and the year 2021 from Ruijin Hospital, with ethical approval. Each whole-slide image (WSI) corresponds to one patient, where the patient is labeled as either positive (CPS $\geq 5$) or negative (CPS $< 5$) determined by its PD-L1 combined positive score (CPS) tested from the immunohistochemistry (IHC) test. The patch-level annotation inherits the associated patient/WSI-level label. The resolution of a WSI is around $10,000 \times 10,000$ pixels, which is split into non-overlapping images (patches) of $512 \times 512$ pixels at the magnification $20 \times$, and afterward resized to $224 \times 224$ to get aligned with the two public datasets. Background patches are excluded for downstream analysis with the OTSU algorithm. Each patch comprises approximately 200 cells i.e., nodes. We summarize the data descriptions in Table I.

| Dataset | GIST-PDL1 | CRC-MSI | STAD-MSI |
|---------|-----------|---------|----------|
| # Patients | 129 | 315 | 360 |
| # Training Images | 7,676 | 93,408 | 100,570 |
| Training Positive Rate | 41.10% | 50.0% | 50.0% |
| # Test Images | 2,471 | 99,904 | 118,008 |
| Test Positive Rate | 47.71% | 29.4% | 23.6% |
| Magnification | $20 \times$ | $20 \times$ | $20 \times$ |
| Original Patch Size | $512 \times 512$ | $224 \times 224$ | $224 \times 224$ |

B. Comparison on the Number of Trainable Parameters

Table II reports the number of trainable parameters of all the models we evaluated in Table III. The scale of the trained model is jointly determined by the choice of modules in CNN, GNN, and fusion layers, where we highlight different options by color. The choice of colors aligns with the associated modules visualized in Fig. 1. ‘N/A’ indicates an absence of such layers in the framework. The numbers are reported in millions ($1 \times 10^6$).

| Dataset | GIST-PDL1 | CRC-MSI | STAD-MSI |
|---------|-----------|---------|----------|
| GNN Fusion | N/A | MOBILENET V3 | DENSENET | RESNET |
| N/A | N/A | - | 1.7865 | 7.2225 | 11.3110 |
| GCN MLP TRANSFORMER | 0.0665 | 1.8506 | 7.2866 | 11.3751 |
| | - | 8.4943 | 13.9303 | 13.1231 |
| GIN MLP TRANSFORMER | 0.0712 | 1.8552 | 7.2912 | 11.3797 |
| | - | 8.4989 | 13.9349 | 13.1277 |

C. Results and Analysis

Image-level performance is evaluated with two metrics, namely test accuracy (ACC) and area-under-curve (AUC). Similarly, we evaluate patient-level prediction with AUC (denoted as AUC$\text{patient}$). As shown in Table III, the fused learning schemes consistently achieve more than 5% performance gain over plain CNNs. The improvement is more significant at
the patient level at up to 23%. The additional performance boost suggests that our design of the integrated scheme has better potential to overcome the disturbance of heterogeneous patches for patient-level overall diagnosis. The main takeaways include: 1) An individual GNN fails to achieve satisfactory performance. But as a parallel layer, GNNs can enhance the learning capability of CNN by a learnable fusion layer. 2) MLP, though simple, serves as a good fusion layer. 3) General speaking, the TRANSFORMER integrator outperforms the simple MLP scheme. However, one can not tell whether MLP or TRANSFORMER is a universally better fusion solution. 4) All the integrated models outperform the plain CNNs or GNNs. 5) For the choice of a GNN module, GCN and GIN do not present a significant advantage one over the other.

### Ablation Study.
In the MLP fusion scheme where feature alignment is not necessary, we found the performance is sensitive to the dimension of $H_T$, $H_G$. In Figure 5, we present ablations on $d_G/d_T$ with CRC-MSI, where we use GIN and RESNET18 as the backbones and consequently $d_T$ is fixed to 512 (see supplementary for details). Thus, $d_G/d_T$ needs to be carefully tuned to achieve satisfactory performance in the MLP fusion scheme. In contrast, with the feature alignment, TRANSFORMER fusion layers are easier to train.

### IV. CONCLUSION
This work proposes a fusion framework for learning the topology-embedded images by CNN and GNN. Specifically, we incorporate GNN into the fusion model to add local geometric representations for cell-graph patches on top of CNNs which extracts a global image feature representation. The CNNs and GNNs are trained in parallel and their output features are integrated into a learnable fusion layer. We validate the framework using different combinations of CNN, GNN, and fusion modules on real H&E stained histology datasets, which surpasses the plain CNN or GNN methods by a significant margin. The experiments yield that the geometric feature and image-level feature are complementary. Finally, the constructed image-graph bimodal datasets can serve as a benchmark for future studies.

### TABLE III
**AVERAGE TEST ACC AND AUC COMPARISONS ON THREE BENCHMARKS OVER 7 REPETITIONS.**

| Model | ACC | AUC | AUC+/AUC- | ACC | AUC | AUC+/AUC- |
|-------|-----|-----|-----------|-----|-----|-----------|
| GCN   | 75.0±7.1 | 77.6±4.7 | 77.6±4.7 | 77.6±4.7 | 77.6±4.7 | 77.6±4.7 |
| GNN   | 75.0±7.1 | 77.6±4.7 | 77.6±4.7 | 77.6±4.7 | 77.6±4.7 | 77.6±4.7 |
| GIN   | 75.0±7.1 | 77.6±4.7 | 77.6±4.7 | 77.6±4.7 | 77.6±4.7 | 77.6±4.7 |

---

**Fig. 5.** The image-level test AUC of CRC-MSI is sensitive to different choices of $d_G/d_T$. The additional performance boost suggests that our design of the integrated scheme has better potential to overcome the disturbance of heterogeneous patches for patient-level overall diagnosis. The main takeaways include: 1) An individual GNN fails to achieve satisfactory performance. But as a parallel layer, GNNs can enhance the learning capability of CNN by a learnable fusion layer. 2) MLP, though simple, serves as a good fusion layer. 3) General speaking, the TRANSFORMER integrator outperforms the simple MLP scheme. However, one can not tell whether MLP or TRANSFORMER is a universally better fusion solution. 4) All the integrated models outperform the plain CNNs or GNNs. 5) For the choice of a GNN module, GCN and GIN do not present a significant advantage one over the other.

### REFERENCES

1. A. Echle et al., “Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: a systematic literature review,” *Immuninformatics*, p. 100008, 2021.
2. A. Feichtenbeiner et al., “Critical role of spatial interaction between CD8+ and Foxp3+ cells in human gastric cancer: the distance matters,” *Cancer Immunology, Immunotherapy*, vol. 63, no. 2, pp. 111–119, 2014.
3. S. Barua et al., “Spatial interaction of tumor cells and regulatory T cells correlates with survival in non-small cell lung cancer,” *Lung Cancer*, vol. 117, pp. 73–79, 2018.
4. R. Noble et al., “Spatial structure governs the mode of tumour evolution,” *Nature Ecology & Evolution*, vol. 6, no. 2, pp. 207–217, 2022.
5. C. Gunduz et al., “The cell graphs of cancer,” *Bioinformatics*, vol. 20, no. suppl_1, pp. i44–i115, 2004.
6. B. Yener, “Cell-graphs: image-driven modeling of structure-function relationship,” *Communications of the ACM*, vol. 60, no. 1, pp. 74–84, 2016.
7. Y. Wang et al., “Cell graph neural networks enable digital staging of tumour microenvironment and precisely predict patient survival in gastric cancer,” *medRxiv*, 2021.
8. M. M. Bronstein et al., “Geometric deep learning: going beyond Euclidean data,” *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.
9. Z. Wu et al., “A comprehensive survey on graph neural networks,” *IEEE Transactions on Neural Networks and Learning System*, vol. 32, no. 1, pp. 4–24, 2020.
10. Z. Zhang et al., “Deep learning on graphs: a survey,” *IEEE Transactions on Knowledge and Data Engineering*, 2020.
11. J. Liang et al., “A deep neural network combined CNN and GCN for remote sensing scene classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4325–4338, 2020.
12. F. Peng et al., “Multi-output network combining GNN and CNN for remote sensing scene classification,” *Remote Sensing*, vol. 14, no. 6, pp. 1478, 2022.
13. Y. Dong et al., “Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification,” *IEEE Transactions on Image Processing*, 2022.
14. Y. Wei et al., “Collaborative learning of images and geometrics for predicting isocitrate dehydrogenase status of glioma,” arXiv:2201.05530, 2022.
15. J. Gliner, S. S. Schoenhofe et al., “Neural message passing for quantum chemistry,” in *ICML*, 2017.
16. A. Vaiswani et al., “Attention is all you need,” in *NeurIPS*, 2017.
17. J. Huang et al., “CA2-Net: nuclei segmentation framework with a microscopy cell benchmark collection,” in *MICCAI*, 2021.
18. J. Ke et al., “Clusterseg: A crowd cluster pinpointed nucleus segmentation framework with cross-modality datasets,” *Medical Image Analysis*, vol. 85, pp. 102758, 2023.
19. Z. He et al., “Transsume: A lightweight multi-task transformer for nuclei segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 206–215.
20. P. Labin et al., “Radiomics: the bridge between medical imaging and personalized medicine,” *Nature Reviews Clinical Oncology*, vol. 14, no. 12, pp. 749–762, 2017.