# GSDG: Exploring a Global Semantic-Guided Dual-Stream Graph Model for Automated Volume Differential Diagnosis and Prognosis

Shouyu Chen[1(✉)], Xin Guo[2], Jianping Zhu[2], and Yin Wang[1]

[1] Tongji University, Shanghai, China
{1910667,yinw}@tongji.edu.cn
[2] Dalian University of Technology, Dalian, China
{guoxinguo,zhujp}@mail.dlut.edu.cn

**Abstract.** Three-dimensional medical images are crucial for the early screening and prognosis of numerous diseases. However, constructing an accurate computer-aided prediction model is challenging when dealing with volumes of different sizes due to numerous slices (native nodes) in a single case and variable-length slice sequence. We propose a Global Semantic-guided Dual-stream Graph model to address this issue. Our approach differs from the existing solution that aligns volumes with varying numbers of slices through downsampling. Instead, we leverage global semantic vectors to guide the grouping of native nodes, construct supernodes, and build dual-stream graphs by incorporating the sequential association of each volume's unique slices and the feature association of global semantic vectors. Specifically, we propose a shared global semantic vectors-based grouping method that aligns the number and the semantic distribution of nodes among different volumes without discarding slices. Furthermore, we construct a dual-stream graph module that enables Graph Convolutional Networks (GCN) to make clinical predictions from computer tomography (CT) volumes through the natural sequence association between native nodes and, simultaneously, the latent feature association between semantic vectors. We provide interpretability by visualizing the distribution of native nodes within each group and weakly-supervised slice localization. The results demonstrate that our method outperforms previous work in diagnostic (96.74%, +2.81%) and prognostic accuracy (84.56%, +1.86%) while being more interpretable, making it a promising approach for medical image analysis scenarios with limited fine-grained annotation.

**Keywords:** Diagnosis · Prognosis · Semantic-guided grouping · Graph convolutional networks · Lesion localization

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_45.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 462–471, 2023.
https://doi.org/10.1007/978-3-031-43904-9_45

## 1 Introduction and Related Works

Deep learning algorithms have shown success in performing computer-aided diagnosis (CAD) tasks using high-dimensional medical images, such as classification [20], detection [18], and segmentation [19]. Physicians typically review slices sequentially in CT and diagnose based on changes in lesion morphology and knowledge within the key slices. However, the variability in the number of slices between volumes challenges the CAD model in capturing the complex associations between slices and assisting medical decision-making.

The diagnosis of COVID-19 is challenging. Convolutional Neural Networks (CNNs) and their variants, such as 3D CNNs [15,21] and 2.5D CNNs [18], have shown promise, CNNs required pre-extracted regions of interest (ROI) with aligned size, fine-grained annotation, and high computational complexity. For instance, [21] employed a two-stage process where a segmentation model is trained first using segmentation masks. Then, a fixed-size 3D tensor is cropped from the lung region, transformed into a 4D tensor, and fed into the 3D classification net. Additionally, CT slices have intrinsic non-Euclidean associations, which has led to recent interest in using Transformers and Graph Neural Networks (GNNs) to handle them. For example, ViT [3], and Swin Transformer [8] are variants of Transformers that use multi-head self-attention to learn fully-connected associations between image patches. However, this architecture had primarily been applied to medical 3D patches [12,16] rather than the complete sequence. Regarding GNNs, ViG [4] organized images into patch sequences but had yet to extend the model to variable-length sequences. Another earlier work [7] used systematic sampling to align the number of slices, introducing sampling bias. We identify a common issue that existed in CNNs [15,18], Transformers [12,16], and GNNs [7] that they cannot be directly trained end-to-end on variable-length slice sequences. Therefore, this paper proposes a graph model to break through this limitation by reconstructing node knowledge at the supernode level.

One of the major challenges is achieving consistency training in GCNs while preserving the integrity of the slice information. Existing graph model [7] downsampled slices to align nodes, resulting in the loss of some critical information. This approach also treated slices at the same location after sampling from different volumes equally, assuming they have the same semantics, which contradicts semantic consistency and clinical meanings. Moreover, the complex associations between slices further complicate the modeling. In three-dimensional medical images, the slice sequence dynamics naturally encode critical knowledge about morphology changes, which is of great diagnostic value. A recent study [4] showed that using sparse connections can improve the efficiency of GNNs, at least for natural images, and lead to better performance than other architectures such as CNNs and Transformers. Existing research mainly utilized GNNs to extract associations among slices or patches, constructing topology connections using methods such as k-nearest neighbors [4,13] and cosine similarity [7], as well as a learnable adjacency matrix [13]. Such approaches have limitations in capturing

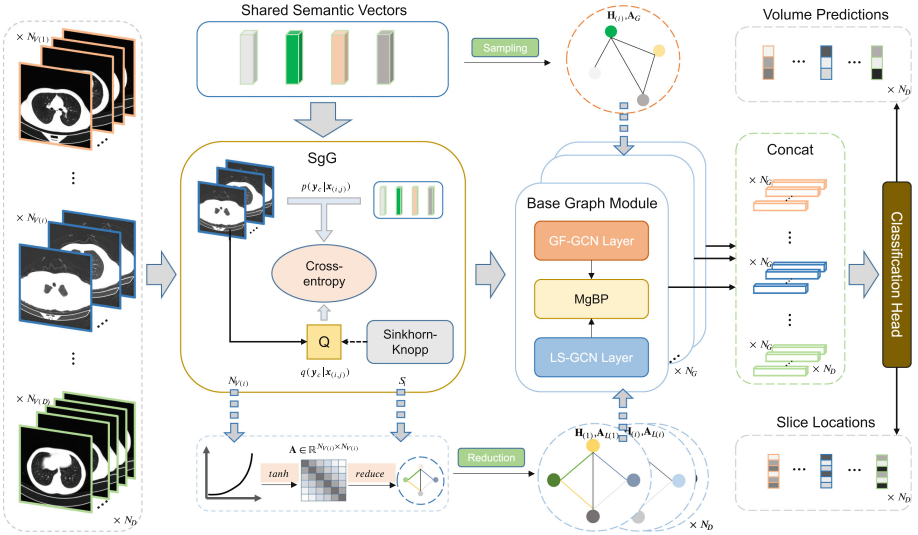


**Fig. 1.** Global Semantic-guided Dual-stream Graph (GSDG) model schematic diagram. Native CT nodes are first transformed into super-nodes guided by shared semantic vectors. Global feature and local sequence adjacency matrices are then generated to facilitate learning in the dual-stream base graph modules. The classification head produces diagnostic or prognostic probabilities and weakly-supervised slice localization upon concatenated multi-modal bilinear pooling features from all graph layers. *GF*: global feature, *LS*: local sequence. (Color figure online)

the task-specific local sequential associations between slices and the higher-level global feature associations.

Besides, various approaches have been developed to locate key slices in CT volumes under weak supervision. For instance, [7] proposed an end-to-end node masking method. In contrast, [11] used CNNs as feature extractors and selected a fixed number of substantial slices based on Shannon entropy under a multiple-instance learning framework. This method calculates the average prediction distribution of slices under random noise but cannot be trained end-to-end. This limitation is precisely the problem we aim to address. Our contributions are: (1) We propose a Global Semantic-guided Dual-stream Graph model for weakly-supervised graph classification tasks. It contains an unsupervised grouping algorithm called Semantic-guided Grouping, which aligns variable-length slice sequences using shared global semantic vectors to enable precise prediction by aligning both the node numbers and semantics. (2) We develop a dual-stream Base Graph Module incorporating the local slice sequence and global semantic knowledge by learning these two representations jointly. (3) We thoroughly evaluate the effectiveness of our proposed method through comparison and ablation experiments. Our method outperforms the weakly-supervised benchmark GCNs in terms of accuracy of diagnosis and prognosis on a publicly available

CT dataset while maintaining similar slice localization performance and offering more interpretability.

## 2 Method

### 2.1 Problem Statement

Given a dataset $\mathbf{D} = \{(\mathbf{V}_i, \mathbf{y}_i)\}_{n=1}^{N_D}$, which consists of $N_D$ volumes. Each volume is represented by a set of slice nodes $\{\mathbf{v}_{i,j}\}_{j=1}^{N_{V(i)}}$, where $N_{V(i)}$ is the cardinality and varies for each volume. The volume-level label is given by $\mathbf{y}_i$. We first extract the native descriptors of the slices $\mathbf{X}_i = \{\mathbf{x}_{i,j} | \mathbf{x}_{i,j} = \mathrm{F}_{\text{ext}}(\mathbf{v}_{i,j})\}_{j=1}^{N_{V(i)}} \in \mathcal{R}^{m_0 \times N_{V(i)}}$ using a spatial feature extractor, where the output of $\mathrm{F}_{\text{ext}}$ is an $m_0$-dimensional vector. To perform volume-level prediction under the guidance of shared semantic vectors, we introduce the **G**lobal **S**emantic-guided **D**ual-stream **G**raph (GSDG) model: $\hat{\mathbf{y}} = \mathrm{F}_{\text{GSDG}}(\mathbf{X}, \mathbf{C})$. $\mathbf{C}$ indicate semantic vectors and will be introduced in the following. Figure 1 provides a schematic diagram of our method.

### 2.2 Constructing Super-Nodes

We introduce a method for grouping native nodes $\mathbf{X}$ into super-nodes $\mathbf{H}$ which we denote as $\mathbf{H} = \mathrm{F}_{\text{Gro}}(\mathbf{X}, \mathbf{C})$. To accomplish this, we propose semantic vectors $\mathbf{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\} \in \mathcal{R}^{m_0 \times K}$ that correspond to $K$ groups and are shared across all volumes, end-to-end updated with the model. In an unsupervised setting, previous work [17] has extended cross-entropy minimization to optimal transportation. We build on this approach and draw inspiration from [1] to propose our unsupervised grouping algorithm, **S**emantic-**g**uided **G**rouping (SgG). SgG utilizes semantic vectors to guide the grouping process, ensuring that the resulting super-nodes align both the semantic and the number of nodes simultaneously on the variable-length volumes. However, minimizing cross-entropy in unsupervised classification can result in degeneration, where all slices are assigned to a single label. To address this issue, we encode the grouping label of a slice as the posterior distribution $q(\mathbf{y}_c|\mathbf{x}_{i,j})$ and re-express cross-entropy as:

$$E(p, q) = -\frac{1}{N_D} \sum_{i=1}^{N_D} \frac{1}{N_{V(i)}} \sum_{j=1}^{N_{V(i)}} \sum_{y_c=1}^{K} q(\mathbf{y}_c|\mathbf{x}_{i,j}) \log p(\mathbf{y}_c|\mathbf{x}_{i,j}) \tag{1}$$

To achieve semantic uniformity, we reformulate $p(\mathbf{y}_c|\mathbf{x}_{i,j})$ as $p(\mathbf{y}_c|\mathbf{x}_{i,j}, \mathbf{c}_{y_c})$:

$$p(\mathbf{y}_c|\mathbf{x}_{i,j}, \mathbf{c}_{y_c}) = \frac{\exp(\mathbf{x}_{i,j}^\top \mathbf{c}_{y_c}/\tau)}{\sum_{y'_c} \exp(\mathbf{x}_{i,j}^\top \mathbf{c}_{y'_c}/\tau)} \tag{2}$$

where $\tau$ is a temperature hyper-parameter. The mapping from native nodes to semantic vectors is described by Eq. 2. We represent this mapping using $\mathrm{Q} \in \mathcal{R}^{K \times N_{V(i)}}$, and optimize it to maximize the similarity between the native node

features and the semantic vectors of their corresponding groups. Therefore, the optimization objective of $\mathrm{F}_{\mathrm{Gro}}$ can be formulated as follows:

$$\min_{p,q} E(p,q) \text{ s.t. } \forall \mathbf{y}_c : q(\mathbf{y}_c|\mathbf{x}_{i,j}) \in \{0,1\} \text{ and } \sum_{j=1}^{N_{V(i)}} q(\mathbf{y}_c|\mathbf{x}_{i,j}) = \frac{N_{V(i)}}{K} \tag{3}$$

We utilize the Sinkhorn-Knopp (SK) algorithm [2] to handle the constraint term, which aims to distribute $N_{V(i)}$ native nodes uniformly into $K$ groups. The SK algorithm produces an assignment matrix $\mathbf{S} \in \mathcal{R}^{N_V \times K}$, where each row is a one-hot vector indicating the group index to which a native node belongs. Consequently, we compute $\mathbf{H} = FC(\mathbf{X}\frac{\mathrm{S}}{\|\mathrm{S}\|_0})$, where $FC$ is a linear layer and $\|\cdot\|_0 : \mathcal{R}^{N_{V(i)} \times K} \rightarrow \mathcal{R}^{1 \times K}$ computes the column-wise 0-norm of a matrix.

### 2.3 Bi-level Adjacency Matrices

We aim to train GCN on a dataset of variable-length volumes; and have already grouped the native-nodes into super-nodes, $\mathbf{H}$. Then, we could depict the adjacency relationships between nodes in $\mathbf{H}$ by semantic vectors or native nodes. Inspired by the multi-resolution model design in CNNs, we explicitly model the global and local adjacency relations: $\mathbf{A}_G, \mathbf{A}_L = \mathrm{F}_{Adj}(\mathbf{X}, \mathbf{S}, \mathbf{C})$. $\mathbf{A}_G$ represents the global semantic adjacency matrix, while $\mathbf{A}_L$ the local sequence adjacency matrix of the learned super-nodes from $\mathbf{X}$.

**Global Adjacency Matrix Based on Grouping Semantic Vectors.** The existing study [13] utilized the Gumbel reparameterization trick [6,10] to allow gradient flow through the adjacency matrix. Building upon the global semantic vectors $\mathbf{C}$ constructed in Sect. 2.2, we learn representations of the commonality between super-nodes over different volumes. Exactly, a link predictor, constructed by a 2-layer MLP, takes the concatenation of semantic vectors $\mathbf{c}_i$ and $\mathbf{c}_j$ as its input and produces the output $\theta_{i,j}$. We calculate the corresponding value, $\mathbf{A}_{G(i,j)} = sigmoid\left(\left(\log\left(\theta_{i,j}/\left(1-\theta_{i,j}\right)\right) + \left(g^1_{i,j} - g^2_{i,j}\right)\right)/s\right)$, in the global adjacency matrix, $g^1_{i,j}, g^2_{i,j} \sim \mathrm{Gumbel}(0,1)$ and $s$ is a hyper-parameter.

**Local Adjacency Matrix Based on Native Sequence Association.** To account for the associations varying with relative distance between slices within a volume, we utilize exponential smoothing to create a sequence adjacency matrix $\mathbf{A} \in \mathcal{R}^{N_{V(i)} \times N_{V(i)}}$ for each volume. The adjacency value between native nodes $i$ and $j$ is calculated by: $\mathbf{A}_{i,j} = \tanh(\sum_{d=1}^{D}(1-\mathbf{s}_d)\mathbf{s}_d^{|i-j|})$. Here, $\mathbf{s}$ represents the output of the *sigmoid* function applied to a learnable vector $\mathbf{w}_L \in \mathcal{R}^D$, and $D$ is a hyper-parameter. We combine the connectivities of native nodes belonging to the same group using the allocation matrix $\mathbf{S}$ introduced in Sect. 2.2 and obtain the reduced local adjacency matrix appliable to super-nodes: $\mathbf{A}_L = \mathbf{S}^{\mathrm{T}}\mathbf{A}\mathbf{S}$.

### 2.4 Dual-Stream Graph Classifier

We introduce a graph classification module, denoted as $\hat{\mathbf{y}} = \mathrm{F}_{\mathrm{Cls}}(\mathbf{H}, \mathbf{A}_G, \mathbf{A}_L)$, consisting of stacked **B**ase **G**raph **M**odules (BGM) and a classifier. The BGM comprises two parallel isomorphic graph convolutions, a global feature GCN layer and a local sequence GCN layer, and a **M**ulti-**g**raph **B**ilinear **P**ooling (MgBP) module.

**Base Graph Module.** The two GCN layers pass messages between super-nodes from distinct perspectives and output $\mathbf{H}_G, \mathbf{H}_L \in \mathcal{R}^{m \times K}$. Then the MgBP module extracts fine-grained graph-level representation, $\mathbf{F}$, using a low-rank multimodal bilinear module: $\mathbf{F} = \mathbf{P}\sigma\left(\mathbf{U}\mathbf{H}_G \circ \mathbf{V}\mathbf{H}_L\right) + b$, where trainable weights $\mathbf{U}, \mathbf{V} \in \mathcal{R}^{d \times m}$, $\mathbf{P} \in \mathcal{R}^{\frac{m}{N_G} \times d}$, $b \in \mathcal{R}$, $d < m$, and $N_G$ is the number of BGM layers. The resulting matrix, $\mathbf{F} \in \mathcal{R}^{\frac{m}{N_G} \times K}$, represents the output of one BGM layer.

**Classification Head.** To obtain hierarchical features, we concatenate $\mathbf{F}_i$ from each BGM layer along the feature dimension, resulting in $\mathbf{F}_{final} = \|_{i=1}^{N_G} \mathbf{F}_i \in \mathcal{R}^{m \times K}$. We then compute the mean and max along the node dimension separately, resulting in two length-$m$ vectors. These vectors are concatenated and passed through a 2-layer MLP and softmax activation for classification.

**Weakly-Supervised Informative Slice Localization.** Firstly, we obtain the predicted probability $p_{base}$ for the target class from $\mathbf{F}_{final}$ using the Classification Head. Then, we mask each super-node in turn to create $K$ sub-matrices of size $m \times (K-1)$. The Head is utilized again to calculate the new probabilities, which results in a vector $\mathbf{p} \in \mathcal{R}^K$. The groups are ranked by $\mathbf{d}_{sn} = p_{base} - \mathbf{p}$. Within a group, the distances between the native nodes and the super-node are measured using the dot product and normalized to the interval [0,1], which results in $\mathbf{d}_{rn}$. Slices' global importance within the group $i$ is $\mathbf{d}_{sn(i)}/\mathbf{d}_{rn}$. We repeat this procedure for all groups and select the top $k$ slices globally.

## 3 Experiments and Discussion

### 3.1 Dataset and Pre-processing

Our experiment used a public CT volume dataset 2019nCoVR [21], which contains complete chest CT scans from 929 COVID-19 (NCP) patients, 964 patients with common pneumonia (CP), and 849 healthy individuals (Normal). Among them, 408 patients are annotated with prognosis labels and some pneumonia patients with slice-level lesion annotations. To make a fair comparison with [7], we divided the dataset into training, validation, and testing sets using the same method with 20 random seeds. Each slice was resized to $224 \times 224$ pixels and normalized with $mean = 0.449$ and $std = 0.226$.

We chosen ResNet-50 [5] as the $F_{ext}$ corresponding to $m^0 = 2048$. Only the frozen $F_{ext}$ module was pre-trained on ImageNet, while all other modules were trained end-to-end on the 2019nCoVR dataset. The model hyper-parameters were set to $m = 256$, $K = 6$, $N_G = 2$, $d = 64$, $k = 10$, $s = 0.5$, $\tau = 1$, and $D = 8$. AdamW [9] served as the optimizer with a learning rate of $3e^{-3}$ and a batch size of 64. The model was trained with the cross-entropy loss for 20 epochs. After each feature aggregation, a Dropout [14] layer with a rate of 0.1 was added. The selection of hyper-parameter values is mainly based on experience and constraints in the formula above. The diagnostic model, which has one output node activated by the sigmoid function, was trained from scratch. In contrast, the prognosis model with three output nodes activated by softmax was initialized with the weights of the trained diagnostic model, except for the classification head. We used the same evaluation metric, precision and recall, for weakly-supervised localization as [7].

### 3.2 Differential Diagnosis, Prognosis and Weakly-Supervised Localization

Table 1 presents the diagnostic and prognostic performance of two state-of-the-art architectures and our proposed method. The clinical AI system based on 3D CNNs [21] was trained with volume-level and additional pixel-level labels. Our proposed method outperforms the state-of-the-art weakly-supervised graph model GCN-DAP [7] and 3D CNNs [21] in terms of diagnostic accuracy and AUC scores. Furthermore, GSDG also surpassed [7] in the prognostic task. These results demonstrate the superiority and effectiveness of our method in modeling full-size variable-length volumes. The left panel of Fig. 2 compares the performance of GSDG and experienced radiologists [21] in diagnostic tasks. For the NCP, GSDG outperformed the radiologists. Moreover, when identifying Normal and CP cases, GSDG achieved similarly high levels of AUCs. These results suggest that our method is advantageous over radiologists in NCP diagnosis. It is worth noting that our model required fewer training epochs than [21], as shown in the right panel of Fig. 2, which highlights the faster convergence of GSDG compared to GCN-DAP [7]. GSDG located the most informative CT

**Table 1.** Diagnostic and prognostic performance comparison. *: Results come from the original paper rather than 20 runs. *ACC*: macro accuracy, *AUC*: macro area under the receiver operating characteristic curve, *SD*: standard deviation, *CI*: confidence interval. Prefix *D* denotes the diagnosis task, and *P* stands for the prognosis task. All scores are multiplied by 100 to simplify the table.

| Method | D-ACC (SD) | D-AUC (95% CI) | P-ACC (SD) | P-AUC (95% CI) |
|---|---|---|---|---|
| 3D-CNN [21] | 92.49 (N/A)* | 98.13 (96.91–99.02)* | N/A | N/A |
| GCN-DAP [7] | 93.93 (0.41) | 99.00 (N/A) | 82.70 (3.90) | N/A (N/A) |
| GSDG (Ours) | 96.74 (0.64) | 99.65 (99.61–99.69) | 84.56 (2.35) | 90.89 (88.77–89.88) |

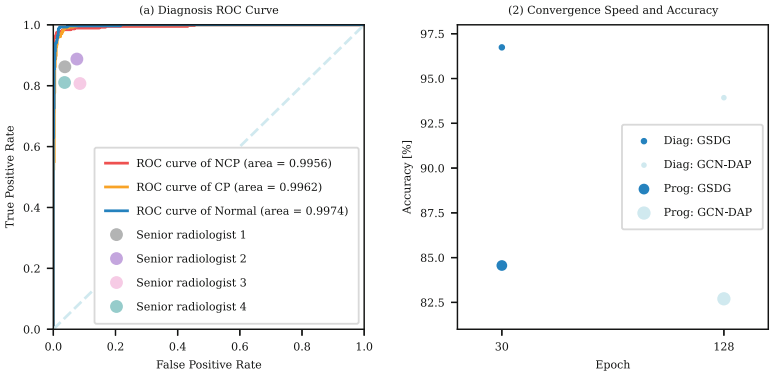


**Fig. 2.** The left panel shows the diagnostic ROC curves of GSDG and the NCP diagnostic scores of four senior radiologists with 15 to 25 years of clinical experience [21]. The right panel demonstrates convergence speed and accuracy of GSDG and [7] on diagnostic and prognostic tasks. The diameter of each circle represents its standard deviation, multiplied by 10 for ease of observation. *NCP*, *CP*, and *Normal* represent COVID-19, common pneumonia, and healthy individuals, respectively (Color figure online)

slices, achieved 51.60% (2.75%) and 89.27% (8.95%) for precision and recall, respectively, slightly worse than [7], 57.39% (3.32%), in precision, but outperformed [7], 79.89% (3.94%), in recall. During our experiments, we frequently observed that the model became unstable as the precision score increased further and the predictions degraded to a single category. This may be due to the loss of some information that only exists in the Hounsfield Unit, as the 2019nCoVR dataset uses JPG instead of DCM format to save the slices. With the emergence of possible technologies that can recover the JPGs losslessly to DCMs, training the model on lung-masked slices is expected to produce better results.

### 3.3 Visualization of Grouping and Slice Localization

We visualized the grouping of native slices in Supplementary Material S.1 and the weakly-supervised slice localization for two patients in S.2. It can be observed that slices belonging to the same group display a remarkable degree of visual resemblance. Besides, the group importance distribution of NCP and CP cases are more similar to each other than to the Normal case, which reflects patterns differences between positive and negative cases. Within the positive cases, our model exhibits a tendency to concentrate more on the lung base in the CP case, as evidenced by columns 4 and 5 of Fig. 2 (S.1), in comparison to the NCP case, where the attention is around the middle lobe, as demonstrated in columns 4, 5 and 6 of Fig. 1 (S.1). These observations may reveal group semantic differences among positive cases and provide a new perspective for clinical diagnosis. Regarding localization, our method's ability to identify lesion slices is superior to its precision performance, as indicated by Figs. 4 and 5 (S.2).

### 3.4 Ablation Study

The ablation study was conducted to evaluate the effectiveness of different node alignment methods and graph structures on the performance of the proposed model for variable-length volumes. The results are presented in Tables 1 and 2 in Supplementary Material, S.3. We compared the performance of our proposed super-node strategy to systematic sampling [7] for node alignment and found that the former outperformed the latter. We also compared global and local adjacency matrices and found that using both resulted in the best overall performance.

## 4 Conclusion

This paper proposes a novel approach for handling variable-length volume while preserving the integrity of the data by not discarding any slices, which is a departure from the previous method. Our approach first introduces a shared global semantic vectors-guided native node grouping scheme. Then we present an efficient and effective dual-stream graph module for simultaneously learning representations from global semantic vectors and sequence associations specific to each volume. Additionally, our approach offers informative slice localization and visually-consistent grouping outcomes, which enhances interpretability for clinical purposes. Moreover, the current dataset format prevents us from using existing semantic segmentation techniques to remove non-pulmonary noise. We will delve deeper into this direction to enhance localization accuracy.

**Acknowledgments.** I would like to thank my wife, Yang Feng, for her support during my doctoral studies.

## References

1. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural Inf. Process. Syst. **33**, 9912–9924 (2020)
2. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
3. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=YicbFdNTTy
4. Han, K., Wang, Y., Guo, J., Tang, Y., Wu, E.: Vision GNN: an image is worth graph of nodes. arXiv preprint arXiv:2206.00272 (2022)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with Gumbel-Softmax. arXiv preprint arXiv:1611.01144 (2016)
7. Liu, C., Cui, J., Gan, D., Yin, G.: Beyond COVID-19 diagnosis: prognosis with hierarchical graph representation learning. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 283–292. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_27

8. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=Bkg6RiCqY7
10. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: a continuous relaxation of discrete random variables. arXiv preprint arXiv:1611.00712 (2016)
11. Meng, Y., et al.: Bilateral adaptive graph convolutional network on CT based COVID-19 diagnosis with uncertainty-aware consensus-assisted multiple instance learning. Med. Image Anal. **84**, 102722 (2023)
12. Niu, C., Wang, G.: Unsupervised contrastive learning based transformer for lung nodule detection. Phys. Med. Biol. **67**(20), 204001 (2022)
13. Shang, C., Chen, J., Bi, J.: Discrete graph structure learning for forecasting multiple time series. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=WEHSlH5mOk
14. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
15. Taleb, A., et al.: 3D self-supervised methods for medical imaging. Adv. Neural Inf. Process. Syst. **33**, 18158–18172 (2020)
16. Tang, Y., et al.: Self-supervised pre-training of Swin transformers for 3D medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20730–20740, June 2022
17. Vedaldi, A., Asano, Y., Rupprecht, C.: Self-labelling via simultaneous clustering and representation learning (2020)
18. Wang, X., Han, S., Chen, Y., Gao, D., Vasconcelos, N.: Volumetric attention for 3D medical image segmentation and detection. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 175–184. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_20
19. Yeung, P.-H., Namburete, A.I.L., Xie, W.: Sli2Vol: annotate a 3D volume from a single slice with self-supervised learning. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 69–79. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_7
20. Yuan, Z., Yan, Y., Sonka, M., Yang, T.: Large-scale robust deep AUC maximization: a new surrogate loss and empirical studies on medical image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3040–3049 (2021)
21. Zhang, K., et al.: Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell **181**(6), 1423–1433 (2020)