Hindawi
International Journal of Intelligent Systems
Volume 2024, Article ID 1728801, 11 pages
https://doi.org/10.1155/2024/1728801

Research Article

Incorporating Adaptive Sparse Graph Convolutional Neural Networks for Segmentation of Organs at Risk in Radiotherapy

Junjie Hu, Chengrong Yu, Shengqian Zhu, and Haixian Zhang

Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, China

Correspondence should be addressed to Haixian Zhang; zhanghaixian@scu.edu.cn

Received 4 July 2023; Revised 4 March 2024; Accepted 4 April 2024; Published 26 April 2024

Academic Editor: Alexander Hošovský

Precisely segmenting the organs at risk (OARs) in computed tomography (CT) plays an important role in radiotherapy's treatment planning, aiding in the protection of critical tissues during irradiation. Renowned deep convolutional neural networks (DCNNs) and prevailing transformer-based architectures are widely utilized to accomplish the segmentation task, showcasing advantages in capturing local and contextual characteristics. Graph convolutional networks (GCNs) are another specialized model designed for processing the nongrid dataset, e.g., citation relationship. The DCNNs and GCNs are considered as two distinct models applicable to the grid and nongrid datasets, respectively. Motivated by the recently developed dynamic-channel GCN (DCGCN) that attempts to leverage the graph structure to enhance the feature extracted by the DCNNs, this paper proposes a novel architecture termed adaptive sparse GCN (ASGCN) to mitigate the inherent limitations in DCGCN from the aspect of node's representation and adjacency matrix's construction. For the node's representation, the global average pooling used in the DCGCN is replaced by the learning mechanism to accommodate the segmentation task. For the adjacency matrix, an adaptive regularization strategy is leveraged to penalize the coefficient in the adjacency matrix, resulting in a sparse one that can better exploit the relationships between nodes. Rigorous experiments on multiple OARs' segmentation tasks of the head and neck demonstrate that the proposed ASGCN can effectively improve the segmentation accuracy. Comparison between the proposed method and other prevalent architectures further confirms the superiority of the ASGCN.

1. Introduction

Segmenting the organs at risk (OARs) in computed tomography (CT) is a crucial step in radiotherapy treatment planning that helps protect the healthy tissue during irradiation [1]. OARs are defined as healthy tissues or organs near the clinical target volume (CTV) whose irradiation could potentially cause damage. For example, the heart is an OAR in the left breast cancer, and constrictor naris is an OAR in the nasopharyngeal cancer. Deep neural networks (DNNs), including discrete deep convolutional neural networks (DCNNs) [2] and continuous ordinary differential equation (ODE) based networks [3], have become ubiquitous models for the segmentation task. One of the most well-known segmentation models is U-Net [4], a simple yet powerful architecture composed of an encoder, a decoder, and shortcut connections between them. Despite its simplicity, U-Net and its variants have delivered promising performance in medical image segmentation tasks. Notably, the recently proposed nnU-Net [5], abbreviated as no new U-Net, has achieved multiple state-of-the-art records by deliberately designed rules for preprocessing, training, and postprocessing strategies, further confirming the effectiveness of the U-Net's paradigm. In addition, researchers in the field have leveraged the transformer [6] architecture to enhance the representation ability of U-Net, as seen in UNETR [7].

In addition to the medical images formed by regular 2D or 3D grids, there are numerous datasets in the non-Euclidean spaces, such as citation datasets [8], knowledge graphs [9], and protein datasets [10]. For those non-Euclidean datasets, graph-based models comprising nodes and edges are the ideal choice for effectively capturing the intrinsic characteristics. Inspired by the breakthrough achieved by the DCNNs, graph convolutional networks (GCNs) [11], specialized neural networks with embedded graph structures, have been widely employed for the analysis of non-Euclidean datasets. The GCNs attempt to learn the representation of nodes by aggregating information through the adjacency matrix in an end-to-end manner. Encouraging results on multiple graph-related tasks (e.g., node classification or graph classification) have been reported by GCNs.

The DCNNs and GCNs are considered as two distinct models for analyzing the Euclidean and non-Euclidean datasets, respectively. Nevertheless, a newly proposed method named dynamic-channel GCN (DCGCN) [12] incorporates the graph structure to model the relationships between channels of the feature maps extracted by the DCNNs. Specifically, each channel is regarded as a node, and a symmetric adjacency matrix describing the relationships between nodes is constructed using a deliberately designed activation function. The activation function is designed to constrain the symmetric channelwise relationship within the range of 0 and 1. Experimental results on the fundus' retinal vessel segmentation task demonstrate that the channel graph contributes to improving segmentation accuracy. Moreover, visualization results indicate that the adjacency matrix remains stable throughout the training phase, suggesting that the network has effectively learned the intricate channelwise relationships.

Despite the DCGCN demonstrating superior performance compared to networks lacking a graph structure, it exhibits two notable limitations. The first limitation lies in the constrained representation of each channel. In DCGCN, the global average pooling (GAP) [13] that averages the feature map along the channel is used to obtain the contextual representation of each channel. When dealing with medical images containing small OARs or feature maps with large spatial sizes, employing the average operation would introduce noise and deteriorate the network's performance. This may also account for the limited usage of DCGCN in the retinal vessel segmentation task [12], which only applies the DCGCN in the last shortcut connection in U-Net. The second limitation of DCGCN is related to the construction of the adjacency matrix within the graph. In DCGCN, the adjacency matrix is derived from the activation function, whose input is the representation of each channel. The deviation of the channel's representation would accumulate when constructing the adjacency matrix.

Faced with the aforementioned limitations in DCGCN, this paper proposes a novel graph structure termed adaptive sparse graph convolutional networks (ASGCNs). The ASGCN aims to construct a dynamic sparse graph to describe the topological connection between channels, thereby enhancing the network's feature extraction ability for OARs' segmentation tasks. The limitations of DCGCN in the aspect of node representation and adjacency matrix construction are addressed in the proposed ASGCN. For the node representation, the GAP in DCGCN is replaced by the learning mechanism that attempts to adjust each channel's representation to accommodate the segmentation task. For the construction of the adjacency matrix, motivated by the self-attention module in the transformer [6], we leverage the coefficient matrix between the key and value in the self-attention module as the adjacency matrix in the graph. Beyond the fully connected adjacency matrix, the ASGCN adaptively truncates the linkage between two nodes based on their correlation, resulting in an adaptive sparse adjacency matrix that contributes to regularizing the segmentation model. The computational principle of the proposed ASGCN can be divided into three steps, as illustrated in Figure 1. In the first step, we establish the node representation and adjacency matrix through the features generated by the neural networks to create the graph structure. The second step involves shifting the adjacency matrix from dense to sparse through an adaptive truncation operation. Finally, in the third step, the features are reconstructed by leveraging the sparse adjacency matrix. In summary, the primary objective of ASGCN is to augment the features in the segmentation network by incorporating the graph structure. The main contributions of this paper are as follows:

(i) A plug-and-play module termed ASGCN is proposed, which leverages a learning mechanism to overcome the potential noise in the node representation within the DCGCN.

(ii) Motivated by the dropout method, the adjacency matrix that describes the topological relationship between nodes is adaptively regularized during the training phase, mitigating the overfitting problem in the OARs segmentation tasks.

(iii) Experimental results of the proposed ASGCN and the control methods demonstrate the effectiveness of the graph structure in the OARs segmentation tasks.

The paper is organized as follows. Section 2 summarizes the related works of the DNN-based methods for medical image segmentation tasks and the current progress of the GCNs. Section 3 provides a detailed illustration of the proposed ASGCN, including graph construction and adaptive sparsity. Section 4 conducts experiments by first verifying the parameter sensitivity of the ASGCN, followed by comparing the ASGCN and control methods. Section 5 discusses the relationship between the proposed ASGCN and closely related approaches. Finally, Section 6 presents the conclusion of the ASGCN.

2. Related Works

In this section, we first summarize medical image segmentation using variants of DNNs, including the DCNNs-based and transformer-based models. Then, we present the architecture of GCNs and their application in medical image analysis.

2.1. Medical Image Segmentation Based on DNNs. Starting with the breakthrough brought by AlexNet [2], DCNNs have dominated the analysis of vision-related tasks, changing the paradigm from handcrafted low-level features to learning-based high-level ones driven by large-scale annotated datasets. Numerous novel architectures of DCNNs

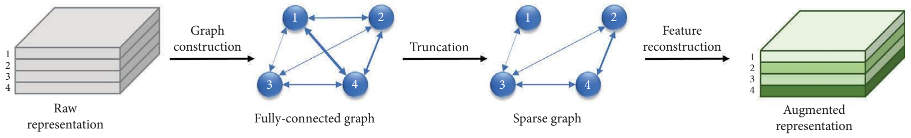

Figure 1: The computational principle of the proposed ASGCN. Each channel in the raw representation is marked by a number, and their topological relationship is described by a dynamic sparse graph.

have been proposed, including the VGG [14], Inception [15], ResNet [16], and EfficientNet [17], all of which have achieved promising accuracy in the ImageNet [18] dataset. Along with the enormous success achieved by DCNNs, it has also become the de-facto method in medical image segmentation tasks. One of the most renowned segmentation networks is U-Net [4], an efficient architecture of fully convolutional networks (FCNs) that won the ISBI cell tracking challenge in 2015. Besides the 2D biomedical image addressed by the original U-Net, there are multiple 3D volumetric datasets, such as CT and magnetic resonance imaging (MRI), where models that consider the depth information are preferred. Based on this concern, the 3D U-Net [19, 20] that models the volumetric characteristics of the target is proposed. It achieves superior performance compared with its 2D counterparts. In addition, the nnU-Net [5] has achieved multiple state-of-the-art records in various biomedical segmentation competitions, further indicating the efficiency of the U-Net architecture. Besides the U-Net architecture, a patch-based learning framework [21] has been proposed to segment the spine in the CT slices. Experimental results on multiple public datasets demonstrate that the patch-based method delivers precise segmentation results.

Among the various architectures of DNNs, perhaps one of the most exciting developments is the attention mechanism [22], which was first proposed to augment the recurrent neural networks (RNNs) to learn the long-range relationship in machine translation tasks. Besides the natural language processing tasks, the attention mechanism has also been widely applied in the medical image segmentation field. Oktay et al. [23] proposed the attention U-Net, which leverages the gate block [24] to modulate the channelwise information produced by the encoder of U-Net. Experimental results show that the attention U-Net is superior to the vanilla U-Net on the pancreas segmentation task. Following the channelwise and spatialwise attention mechanism [25, 26] used in the natural image, CS2-Net [27] extends the attention method to curvilinear structures (e.g., blood vessels) in the 3D medical image. The CS2-Net achieves a higher segmentation accuracy compared with variants of U-Net on multiple retinal vessel datasets. The idea behind the attention mechanism is to highlight the significant features while suppressing the irrelevant ones. Reverse attention [28], a novel attention module that works oppositely compared with the commonly used one, has also been applied in the polyp segmentation [29] and achieved encouraging segmentation accuracy. Wang et al. [30] proposed the mask attention module to precisely segment the lung regions. The mask attention module attempts to focus on lung regions while suppressing the lesion-related artifacts.

While the attention module helps to enhance DCNNs' representation ability, it is not until the introduction of the self-attention module [6] that the attention-based models became the mainstream for both natural image and medical image analysis tasks. Compared with the traditional convolutional operator or recurrent connection, the self-attention module exhibits superiority in modeling long-range dependencies with the assistance of the large-scale annotated dataset. The self-attention first overwhelms the RNNs in the natural language processing field. It later shows superiority over the DCNNs in the ImageNet dataset with the proposition of vision transformer (ViT) [31]. However, one significant limitation of self-attention for the image-related tasks lies in the increased requirement for the annotated dataset, which poses challenges for tasks with a limited dataset, e.g., medical image segmentation. This motivates researchers to propose hybrid networks that blend the convolution and self-attention modules. For example, the UNETR [7] and Swin UNETR [32] are segmentation architectures for volumetric medical images that use the transformer-based encoders and convolutional-based decoders. Recently, Zhang et al. [33] proposed a Swin-based [34] two-stage method for the segmentation of the spine [33]. The method first localizes the spine by using Swin-YOLOX, and then accomplishes the segmentation by using Swin-UNet.

In addition to their applications in medical imaging, variants of DNNs have been used in remote sensing, a field characterized by vast quantities of multimodal images. For example, Hong et al. introduced HighDAN [35], a high-resolution domain adaptation network aimed at addressing the challenge of cross-city semantic segmentation. To tackle the task of land use and land cover classification, a novel architecture combining the transformer and CNNs has been developed [36]. Furthermore, a cross-attention mechanism has been designed to enhance the spatial resolution of hyperspectral images by leveraging multispectral counterparts [37]. In the realm of semisupervised learning, strategies have been employed to harness the vast amount of unlabeled images for hyperspectral image classification [38]. In addition, SpectralGPT [39], a foundational model tailored for remote sensing image analysis, demonstrates promising performance across various downstream tasks following its pretraining on a dataset of one million spectral images.

2.2. Graph Convolutional Networks.

GCNs are graph-based neural networks that attempt to model non-Euclidean structures, e.g., social networks. The GCNs can be categorized into two classes, i.e., the spectral and spatial approaches. For the spectral class, Bruna et al. [40] defined the convolution for the graph in the Fourier domain by utilizing the eigenvectors of its graph Laplacian. Nevertheless, the eigendecomposition of the graph Laplacian is computationally expensive, which limits the GCNs' application for large-scale graphs. Defferrard et al. [41] later proposed to use the Chebyshev polynomials to approximate the localized convolutional filters, eliminating the eigendecomposition of the Laplacian. For the spatial class, Kipf and Welling [11] introduced the GCNs by considering the first-order neighborhood of each node, significantly outperforming the commonly used methods and becoming the de-facto approach for the analysis of graph datasets. Later, multiple variants of GCNs have been proposed, including the GraphSAGE [42] and graph attention networks (GAT) [43].

Many researchers have also explored the application of GCNs in medical image analysis tasks. For instance, the multimodal graph learning (MMGL) framework [44] is proposed to utilize GCNs for disease prediction using a multimodal dataset, which includes images, biomarkers, demographics, and more. ImageGCN [45] builds a multi-relational image-level GCN to identify diseases in chest X-rays. Each node in the ImageGCN represents features from an X-ray image, where the edge between the two nodes indicates their relationships. The recently proposed DCGCN [12] is a specialized graph-based model designed to augment the feature in the DCNNs-based segmentation network for retinal vessels. Despite the accuracy improvements reported by DCGCN, the constrained representation of each node may introduce noise. This inherent limitation inspires us to propose the ASGCN, which is detailed in the next section.

3. Methodology

In this section, we illustrate the details of the proposed ASGCN, including the graph construction and the adaptive sparse adjacency matrix, which are shown in the blue and green blocks in Figure 2, respectively. Finally, we summarize the training algorithm of the proposed ASGCN. We now delve into the graph construction of ASGCN.

3.1. Graph Construction. Let $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ denote the graph with the node set $\mathscr{V}$ and edge set $\mathscr{E}$. For the node set, we suppose there are C nodes in graph $\mathscr{G}$, and each node is denoted as $v_i \in \mathscr{V}$. Let $X \in \mathbb{R}^{C\times K}$ represent the matrix of nodes' raw features, with each row denoting the feature of a node having dimension K. For the edge set, the edge between $(v_i, v_j) \in \mathscr{E}$ is described by the adjacency matrix $A \in \mathbb{R}^{C\times C}$, where each element $A_{i,j}$ represents the connectivity between the i-th and j-th nodes. Alongside the adjacency matrix A, the diagonal degree matrix D , with its elements defined as $D_{ii} = \sum_j A_{ij}$, represents the degree of node i.

In the GCNs proposed by the authors in [11], the propagation of information can be formulated as

$$\begin{cases} H^{l+1} = \sigma\left(\widehat{A}H^l W^l\right), \\ \widehat{A} = \widehat{D}^{-1/2}(A+I)\widehat{D}^{-1/2}, \end{cases} \tag{1}$$

where H^{l+1} represents the transformed representation of nodes in layer $l+1$ and $\widehat{A} \in \mathbb{R}^{N\times N}$ denotes the renormalized adjacency matrix. The first line in equation (1) represents the parameterized transformation between H^l and H^{l+1}, where σ is the nonlinear activation function and W^l is the learnable parameters. The second line in equation (1) denotes the renormalization of the adjacency matrix A whose node is added with self-connection I. The $\widehat{D}$ is the diagonal degree matrix of $A+I$. The renormalization assures that the message passing between the two nodes considers the degree of the two connected nodes simultaneously.

As demonstrated in equation (1), the graph's adjacency matrix $\widehat{A}$ and the node's representation H^l simultaneously impact the representation of nodes in layer $l+1$. Inspired by these two components in the GCN, this paper proposes a graph structure to enhance the features extracted by DCNNs. It is known that the channel of the feature map in DCNNs is an essential factor that describes the semantic patterns of the input. When considering the 3D feature map in the shape of $[C, D, H, W]$ as a graph, each channel can be regarded as a node whose dimension is $D \times H \times W$. Then, the interchannel relationship and intrachannel representation correspond to the adjacency matrix and node's representation, respectively. However, directly applying equation (1) to model the graph of the feature map in DCNNs is difficult for the disparate structures between the non-Euclidean graph described by equation (1) and the feature map's Euclidean graph. The following paragraphs illustrate the strategies used in constructing adjacency matrix $\widehat{A}$, followed by the computation of the nodes' representation H^l for the graph of the feature map.

First is the construction of the adjacency matrix. In conventional GCNs, the adjacency matrix A describes the connections between nodes and its renormed form $\widehat{A}$ is used in the layerwise forward computation. The simplest form of A in the conventional undirected graph is in the binary format, in which 1 and 0 represent two nodes having or not connected, respectively. The binarized adjacency matrix A can be regarded as a *hard* mode, while its renormed form $\widehat{A}$ is in the *soft* mode. This paper aims to design a soft adjacency matrix to describe the channelwise relationship of the feature map extracted by the DCNNs. Motivated by the self-attention approach [6], we leverage the scaled-dot product paradigm to obtain the relationship between channels, which can be formulated as follows:

$$A^l = \frac{F_q\left(H^l; W_q^l\right) F_k^T\left(H^l; W_k^l\right)}{\sqrt{K}}, \tag{2}$$

where the F_q and F_k denote the transformation embedded with the learnable parameters W_q^l and W_k^l, respectively. The divided K represents the dimension of each node for numerical stability. The original scaled-dot product described in [6] is implemented by the fully connected linear transformation designed for the word token in natural language processing. Nevertheless, the H^l obtained from 3D DCNNs is a tensor with dimensions of $[C, D, H, W]$. Due to the high dimensionality of the feature in 3D DCNNs, directly applying the fully connected transformation to each node may

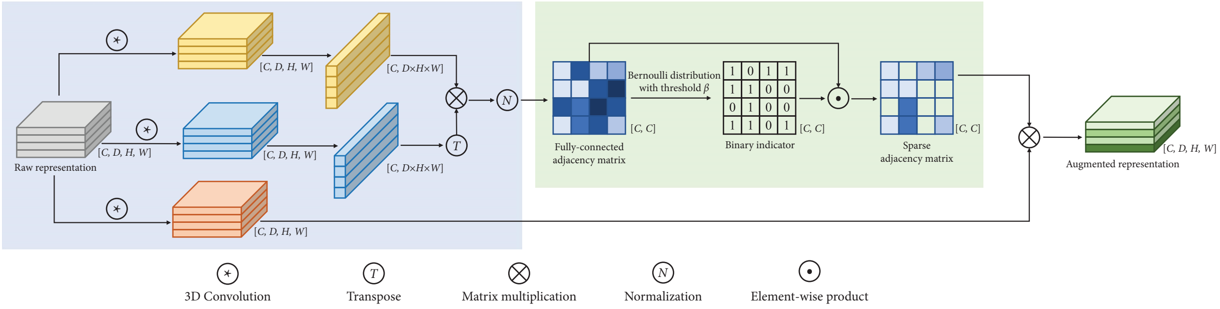

Figure 2: The framework of the proposed ASGCNs.

introduce a severe overfitting problem. Therefore, a transformation that can adapt to the characteristics of the feature map is preferred. Based on this consideration, the F_q and F_k in equation (2) are replaced by the 3D convolution to efficiently extract the features. Note that, the adjacency matrix A^l in equation (2) is appended with the superscript l, indicating that the adjacency matrix is adaptively adjusted along with the layer, rather than the fixed one in equation (1).

After obtaining the adjacency matrix A^l, the next required operation is normalization, which aims to control the scale of each node by considering its connected nodes. Let the normalization be abbreviated as Norm. Then, the resulting normalized adjacency matrix can be denoted as $\widehat{A}^l = \text{Norm}(A^l)$. In ASGCN, two commonly used normalization operations are considered, including the softmax and the degree-based normalization. We consider the element $\widehat{A}^l_{i,j}$ in $\widehat{A}^l \in \mathbb{R}^{C\times C}$. The softmax normalization can be formulated as $\widehat{A}^l_{i,j} = e^{A^l_{i,j}}/\sum_{k=1}^{C} e^{A^l_{i,k}}$, which ensures that the sum of each row is equal to 1. The degree-based normalization can be described as $\widehat{A}^l_{i,j} = A^l_{i,j}/\sqrt{d_i d_j}$, where d_i and d_j denote the degree of the i-th and j-th nodes in $A + I$, respectively. Take the d_i as an example, its value is determined as $d_i = \sum_j (A^l + I)_{ij}$. Both the two normalization approaches own appealing properties. The softmax function aims to normalize the adjacency coefficient as probability, ensuring that each element is squashed to a value between 0 and 1 while their sum equals 1. The degree-based method accomplishes normalization by considering the degree of the two connected nodes. To investigate their effectiveness in the medical image segmentation tasks, we rigorously compared the two normalization approaches in Section 4.2.

The preceding paragraphs illustrate the construction of the normalized adjacency matrix $\widehat{A}^l$ in ASGCN, analogous to the second line in equation (1). The $\widehat{A}^l$ is then utilized to aggregate features from H^l to H^{l+1}, as shown in the first line of equation (1). In the proposed ASGCN, the feature H^l undergoes initial processing through another 3D convolution, which can be considered analogous to the W^l in the first line of equation (1). This processed feature is subsequently multiplied by the adjacency matrix, resulting in the transformed representation H^{l+1}. However, rather than directly applying the $\widehat{A}^l$ to aggregate nodes' representation, another property of interest is sparsity, which contributes to the network's generalization ability. The following subsection illustrates the strategies employed to introduce sparsity to $\widehat{A}^l$.

3.2. Adaptive Sparse Adjacency Matrix. Sparsity is a critical factor that helps to alleviate the overfitting problem, especially for tasks with a limited annotated dataset size, e.g., medical image segmentation. Endow sparsity to $\widehat{A}^l$ could result in a compact representation of features, thus contributing to the networks' generalization ability. A straightforward way to implement the sparsity is to leverage the idea of dropout [46] to randomly reset the coefficient in $\widehat{A}^l$ to 0. The vanilla dropout randomly drops neurons following the Bernoulli distribution parameterized by θ, i.e., each neuron has the probability θ to be kept or $1 - \theta$ to be dropped. When the neuron is dropped, its activation value is reset to 0, which implies that it does not contribute to the prediction result. Many studies have adopted dropout as a regularizer in training DNNs to mitigate the overfitting problem.

While dropout is an effective regularization method, our recent study [47] shows that the dropout is sensitive to its applied layer and the value of θ. Previous experimental results have shown that applying vanilla dropout to multiple shallow layers can increase the risk of underfitting. This phenomenon is attributed to the stagnation of information flow during forward computation. This inherent limitation has motivated us to propose a variant known as surrogate dropout [47], an effective method that does not suffer from the abovementioned constraint yet surpasses the vanilla dropout on both natural and medical image analysis tasks. In the proposed ASGCN, surrogate dropout is employed to adaptively regularize the normalized adjacency matrix $\widehat{A}^l$, resulting in a sparse matrix $\widetilde{A}^l$. Specifically, the element $\widehat{A}^l_{i,j}$ is categorized into two classes based on the hyperparameter β. $\widehat{A}^l_{i,j} \geq \beta$ denotes that the i-th and j-th nodes are correlated, while the opposite suggests that the two nodes are uncorrelated. Following the strategies used in the surrogate dropout, we adaptively regularize the correlated nodes while keeping the rest uncorrelated, which can be described as follows:

$$b_{i,j}^{l} = \begin{cases} \text{Bernoulli}\left(\widehat{A}_{i,j}^{l}\right), & \text{if } \widehat{A}_{i,j}^{l} \geq \beta, \\ 1, & \text{if } \widehat{A}_{i,j}^{l} < \beta. \end{cases} \tag{3}$$

The $b_{i,j}^{l}$ is a binary variable that indicates the kept or dropped status of the $\widehat{A}_{i,j}^{l}$. Equation (3) indicates that the correlated nodes are moderately penalized, preventing the segmentation results from overly depending on them. Moreover, adaptive regularization effectiveness is achieved since the $\widehat{A}_{i,j}^{l}$ that parameterizes the Bernoulli distribution is continuously updated along with the training phase. Afterwards, the adaptively regularized sparse adjacency matrix $\widetilde{A}_{i,j}^{l}$ can be obtained through the pointwise multiplication between $\widehat{A}_{i,j}^{l}$ and $b_{i,j}^{l}$ as

$$\widetilde{A}_{i,j}^{l} = \widehat{A}_{i,j}^{l} \odot b_{i,j}^{l}. \tag{4}$$

In the vanilla dropout, each neuron follows a fixed Bernoulli distribution with a probability denoted as θ. The fixed θ may impede information propagation when extensive layers are applied with dropout. In the proposed ASGCN, the kept probability is determined by $\widehat{A}_{i,j}^{l}$, which can be adaptively adjusted during the training. This self-regulated property of the ASGCN can greatly expand the application of dropout since the Bernoulli distribution for each coefficient can be independently adjusted according to the task's requirement. The ASGCN changes the θ in the vanilla dropout that represents the kept probability of each neuron to β that categorizes $\widehat{A}_{i,j}^{l}$.

During the training phase, each coefficient $\widehat{A}_{i,j}^{l}$ is adjusted according to the rules defined in equations (3) and (4). When in the test phase, the ideal inference mode is to iterate the kept and dropped status for all the $\widehat{A}_{i,j}^{l}$ that is larger than β and average each model's prediction as the final result. One distinct limitation of this inference mode is the prohibitive time cost caused by the exponential possible combinations. Based on this consideration, the expectation of the Bernoulli distribution is used during the test phase. The value of $\widetilde{A}_{i,j}^{l}$ in the test phase is determined as follows:

$$\widetilde{A}_{i,j}^{l} = \begin{cases} \widehat{A}_{i,j}^{l}, & \text{if } \widehat{A}_{i,j}^{l} \geq \beta, \\ 1, & \text{if } \widehat{A}_{i,j}^{l} < \beta. \end{cases} \tag{5}$$

3.3. Algorithm of ASGCN. Algorithm 1 summarizes the computation process of the proposed ASGCN method. Given the feature map H^{l} extracted by the DCNNs, the channelwise adjacency matrix A^{l} is obtained in line 1, where the symbol $*$ denotes the convolution operator. The matrix is later normalized to $\widehat{A}^{l}$ in line 2, where the possible normalization methods include the softmax and degree-based approaches. Then, its sparse version $\widetilde{A}^{l}$ is computed from lines 3 to 7. Finally, the augmented feature $\widehat{H}^{l}$ can be obtained through the matrix multiplication shown in line 8. Similar to the self-attention module in the transformer, there are three learnable parameter matrices in the proposed method. However, the proposed method is only applied to the shortcut connections in the 3D U-Net. Thus, the incremental computational cost incurred by the proposed method is marginal.

4. Experiments

This section empirically verifies the effectiveness of the proposed method through multiple OARs' segmentation tasks. The used datasets and experimental setups are presented first. Then, we evaluated the ASGCN through three aspects, including its parameter sensitivity, application to segment OARs with varied sizes, and visualization of the predictions.

4.1. Dataset and Experimental Setup. The CT dataset for the head and neck cancers [1] is used in the experiments, which can be downloaded from the following website: https://github.com/uci-cbcl/UaNet. The dataset is proposed for radiotherapy treatment planning of nasopharyngeal carcinoma, where a large number of OARs are considered during the treatment. Six OARs are used in the experiments, including the constrictor naris, eyes, lens, optic nerves, temporal lobes, and thyroids. The reason for choosing the OARs lies in the variety, which contains tissues ranging from small (lens) to large (temporal lobes). For OARs consisting of both left and right components, such as the left and right lens, the two parts are merged into one class to eliminate the impact of positions. The dataset is split into training, validation, and test, with three partitions containing roughly 90, 20, and 20 cases, respectively.

The proposed ASGCN is a plug-and-play module that can be incorporated into modern segmentation networks to enhance its representation ability. The broadly used 3D U-Net [19] architecture is employed in the experiments. The 3D U-Net comprises three components, including an encoder with four stages, a decoder with three stages, and shortcut connections between the encoder and decoder. During the experiments, the ASGCN is applied to all the shortcut connections. Cross entropy is used as the loss function, which is optimized by the Adam [48] with the default learning rate of 0.0001. Note that, this paper is not intended to achieve the state-of-the-art records but to leverage the graph structure to augment the representation ability of the network.

The two most widely used segmentation metrics are considered in the experiments, including the dice similarity coefficient (DSC) and 95th percentile Hausdorff distance (HD). The DSC is defined as $\text{DSC} = 2|V_p \cap V_g|/|V_p| + |V_g|$, where the V_p and V_g denote the volume of prediction and ground truth, respectively. For the definition of HD, let C_p and C_g represent the contours of the prediction and ground truth mask, respectively. Then, the HD can be described as $\max\{d(C_p, C_g), d(C_g, C_p)\}$, where $d(C_p, C_g)$ is the distance between the two sets that is defined as

```
Input: Threshold $\beta$; convolutional kernel $W_q^l, W_k^l, W_v^l$; and input feature $H^l$
Output: Augmented feature $\widehat{H}^l$
$A^l = (H^l * W_q^l) \cdot (H^l * W_k^l)^T / \sqrt{K}$
$\widehat{A}^l = \mathrm{Norm}(A^l)$
if $\widehat{A}_{i,j}^l \geq \beta$ then
    $b_{i,j}^l = \mathrm{Bernoulli}(\widehat{A}_{i,j}^l)$
else
    $b_{i,j}^l = 1$
$\widetilde{A}^l = \widehat{A}^l \odot b^l$
$\widehat{H}^l = \widetilde{A}^l \cdot (H^l * W_v^l)$
```

Algorithm 1: Training phase of the proposed ASGCN.

$d(C_p, C_g) = \max_{a \in C_p} \min_{b \in C_g} \|a - b\|$. Since the HD is sensitive to outliers, the 95th percentile of HD between the prediction and ground truth is used.

4.2. Study of Parameter Sensitivity. This subsection first studies the parameter sensitivity of the proposed ASGCN. Two normalization methods, including the softmax and degree-based ones, are empirically compared across a range of β values from 0.1 to 0.9. The DSC values of the two methods for the segmentation of thyroids are shown in Figure 3. For the softmax-based normalization, the ASGCN consistently surpasses the baseline, with the highest DSC achieved when the value of β is 0.3. Note that, the value of β indicates the strength of regularization, where the very high (0.9) or low (0.1) values represent weak and strong regularization, respectively. Though the value of β refers to extreme values, the DSC of the ASGCN is still higher than the baseline, implying the broad applicability of the proposed method. For the degree-based normalization, it can be seen that the DSC of ASGCN fluctuates with the increment of β. The DSC of the degree-based normalization is even lower than the baseline when the value of β is 0.5. We hypothesize that the reason for the fluctuation of the degree-based normalization lies in the differences between the Euclidean and non-Euclidean datasets. The degree-based approach is frequently used for the non-Euclidean dataset by considering the degree of each two connected nodes. However, the graph of the feature map modeled by the ASGCN is in the Euclidean space, where directly leveraging the degree defined in the non-Euclidean space cannot precisely measure the strength of connectivity between nodes. Based on the results shown in Figure 3, the softmax normalization is used in the following experiments with β set to 0.3.

4.3. U-Net with ASGCN. We compare the vanilla 3D U-Net and the one applied with ASGCN on the six OARs. Experimental results are shown in Table 1. It can be found that the ASGCN consistently achieves better metrics than the vanilla 3D U-Net for all the OARs. Regarding the DSC score, the highest increment can be observed for the optic nerves, where the ASGCN increases the DSC from 0.6077 to 0.6418.

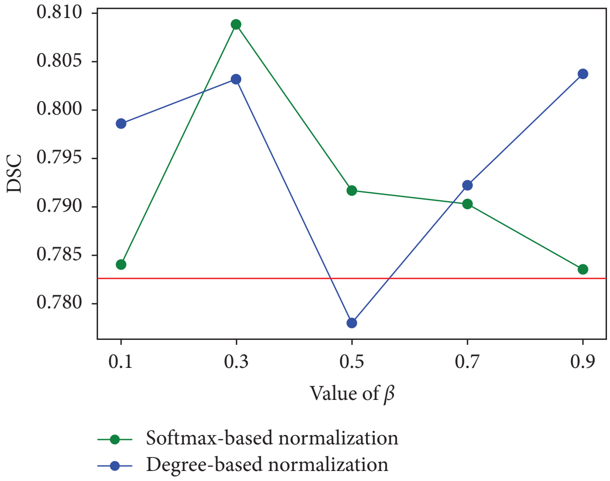

Figure 3: Comparison of softmax and degree-based normalization in the ASGCN for the segmentation of thyroid.

For the HD score, the maximum decrease can be found at the lens, where the ASGCN decreases the HD from 4.6414 to 2.4956. Note that, both the lens and optic nerves are OARs with small sizes, which are sensitive to the features extracted by the network. The promotion brought by the ASGCN manifests that the graph structure helps the network segment small targets. Besides the small OARs, the proposed ASGCN also contributes to enhancing the network's performance for the OARs with large sizes. For example, the ASGCN increases the DSC of the temporal lobes from 0.8595 to 0.8690 while decreasing the HD from 3.0174 to 2.1383. The experimental results presented in Table 1 demonstrate the effectiveness of the ASGCN in enhancing the network's segmentation performance for OARs of different sizes.

4.4. Comparison of Related Approaches. Besides the comparison between the vanilla 3D U-Net and the network applied with the ASGCN, we also quantitatively compare the ASGCN with four closely related methods, namely, the Swin UNETR [49], DCGCN [12], squeeze-and-excitation (SE) [50], and attention gate (AG) [23]. Both compared methods aim to enhance the representation ability of the DCNNs. Experimental results are shown in Table 2. For the

Table 1: Comparison between the vanilla 3D U-Net and the network applied with the proposed ASGCN.

ROI	DSC (↑) of 3D U-Net	DSC (↑) of ASGCN	HD (↓) of 3D U-Net	HD (↓) of ASGCN
Constrictor naris	0.7565	**0.7722**	2.1913	**1.7650**
Eyes	0.8791	**0.8951**	2.0778	**1.1163**
Lens	0.6160	**0.6495**	4.6414	**2.4956**
Optic nerves	0.6077	**0.6418**	2.5521	**2.3080**
Temporal lobes	0.8595	**0.8690**	3.0174	**2.1383**
Thyroids	0.7826	**0.8089**	2.9000	**2.0280**

The upward and downward arrows indicate that the higher and lower are better, respectively. The bold value denotes better performance.

Table 2: Comparison between the network applied with the ASGCN and the related approaches.

ROI	DSC of ASGCN	DSC of Swin UNETR	DSC of DCGCN	DSC of SE	DSC of AG	HD of ASGCN	HD of Swin UNETR	HD of DCGCN	HD of SE	HD of AG
Constrictor naris	0.7722	0.7428	0.7738	0.7702	**0.7774**	1.7650	2.6254	1.7373	1.7634	**1.6910**
Eyes	0.8951	0.8855	0.8890	**0.8968**	0.8864	**1.1163**	1.1728	1.3087	1.1728	1.1584
Lens	0.6495	**0.6554**	0.6119	0.6454	0.6300	2.4956	**1.4060**	4.4368	4.3967	5.2654
Optic nerves	**0.6418**	0.6294	0.6326	0.6220	0.6372	**2.3080**	2.7100	2.6552	2.8887	2.5706
Temporal lobes	**0.8690**	0.7939	0.8577	0.8594	0.8579	**2.1383**	5.8505	3.3170	2.8090	3.6000
Thyroids	**0.8089**	0.7686	0.7990	0.7965	0.7886	**2.0280**	2.9508	4.8513	2.3406	2.3329

Higher DSC and lower HD represent superior performance. The bold value denotes better performance.

constrictor naris, the AG achieves the highest DSC of 0.7774, while also obtaining the lowest HD among the methods. For the eyes, the SE obtains the highest DSC of 0.8968, while the ASGCN achieves a comparable score of 0.8951. The lowest HD of the eyes can be observed in the ASGCN, which is 1.1163. For the lens, which is a very small OAR, the Swin UNETR shows the best performance among the five models. For the rest three OARs, the ASGCN consistently achieves the highest DSC and the lowest HD simultaneously. Regarding the DSC, the ASGCN shows remarkable advantages in the segmentation of thyroids, surpassing the rest of the methods by a large margin. For the HD, the ASGCN exhibits superiority in the segmentation of the temporal lobes with an HD value of 2.1383, which is significantly lower than that of other methods.

4.5. Visualization of Predictions. We further qualitatively compare the predictions of all the methods with the ground truth in this subsection. The comparisons are shown in Figure 4. For the image of the constrictor naris shown in the first row, it can be observed that both the ASGCN, Swin UNETR, AG, DCGCN, and SE have precisely identified the target, except for the vanilla 3D U-Net, whose prediction is discrete. Similar results can be observed in the eyes and lens, where the prediction of vanilla 3D U-Net is less smooth compared to that of the ASGCN. For the optic nerves shown in the fourth row, it can be found that the prediction of the ASGCN is closest to the ground truth among the compared methods. In contrast, the predictions of other methods are either discrete (vanilla 3D U-Net and DCGCN) or incomplete (Swin UNETR and SE). Regarding the temporal lobes, most methods have recognized the majority region of the target (except Swin UNETR), which can be attributed to the clear boundaries between the soft and bone tissues. For the thyroids shown in the last row, only the ASGCN has identified the top regions, while the prediction of the rest of the methods is incomplete. The visualization results shown in Figure 4 demonstrate that ASGCN exhibits the best performance among the compared methods, which can be attributed to the embedded sparse graph in the segmentation network.

5. Discussion

The proposed ASGCN aims to leverage the graph structure to enhance the features extracted by the DCNNs. By examining the computational principle of the ASGCN, we can observe that the GCN [11] is closely related to the transformer [6] architecture. Specifically, by regarding the adjacency matrix in the GCN to the product between the query and the key in the transformer, we can instantly bridge the gap between the GCN and transformer. Nevertheless, there are noteworthy differences between the proposed ASGCN and the transformer. The transformer utilizes the fully connected layer to obtain the representation of query, key, and value. The fully connected layer is a powerful generic connectionism given the large-scale annotated datasets. However, its effectiveness would be significantly reduced when dealing with limited dataset sizes, e.g., the segmentation task of OARs. Thus, the proposed ASGCN replaces the fully connected layers with the 3D convolution to take advantage of the efficient parameter-sharing property inherent in the convolution.

The ASGCN addresses the limitations of node representation in the DCGCN. The GAP that attempts to obtain the node representation in the DCGCN is replaced by the learning mechanism introduced in the ASGCN. The latter approach proves to be more robust in segmenting OARs with varied appearances and sizes. In addition, the adjacency

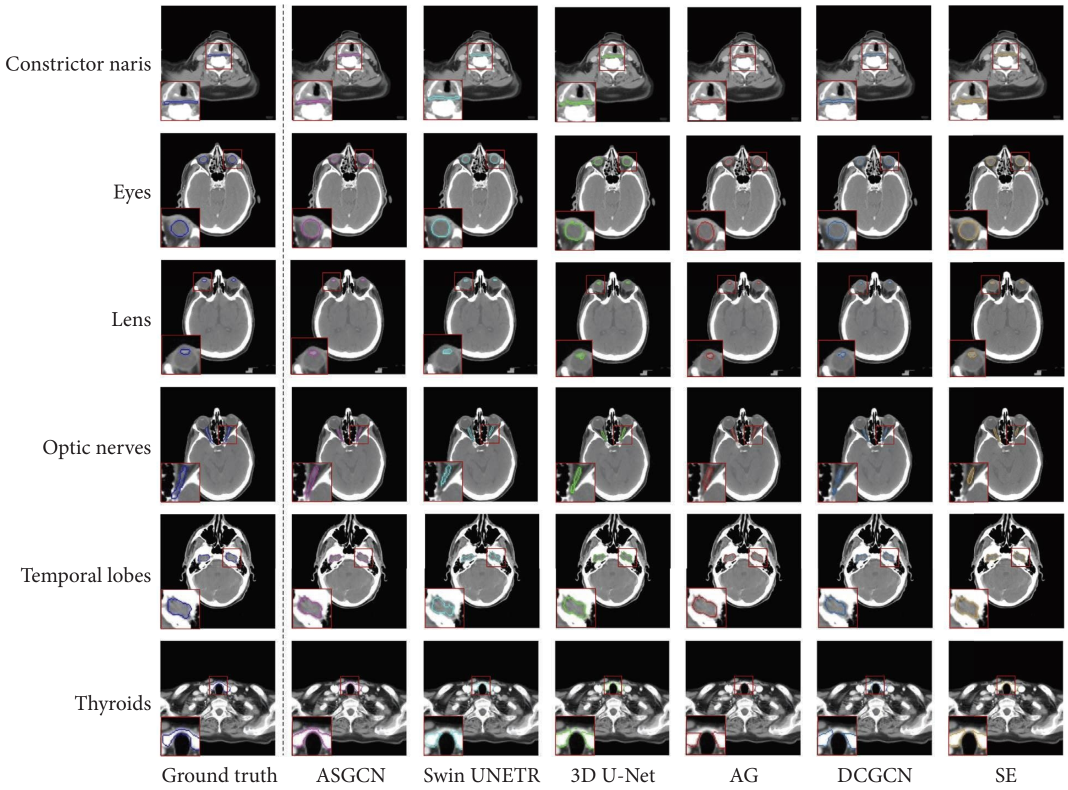

FIGURE 4: Visualization of the ground truth and the predictions from different methods for segmenting the six OARs.

matrix in the ASGCN exhibits sparsity compared to the DCGCN. In the proposed ASGCN, the adjacency matrix is adaptively regularized throughout the training process, preventing the network from being overdependent on significant features and alleviating the overfitting problem.

6. Conclusions

This paper proposes a graph-based module named ASGCN to increase the representation ability of DCNNs. By considering each channel of the features extracted by the DCNNs as a node, we construct the adjacency matrix that describes the relationship between nodes in the Euclidean space. The adjacency matrix is adaptively regularized, displaying insensitivity towards hyperparameters. Experiments are carried out on segmenting six OARs in the head and neck. Results demonstrate the superiority of the ASGCN over the compared methods.

In terms of the limitations of the proposed ASGCN, the property of sparsity can be further improved. Currently, the idea of dropout is leveraged to explicitly introduce the sparsity into the adjacency matrix in the ASGCN. However, other approaches, such as the deliberately designed regularization term in the loss function, can be used to implicitly assign sparsity. Moreover, besides the task-specific segmentation model, the prompt-guided universal segmentation model, either the task prompt or the location prompt, is prevalent in medical image segmentation tasks. In future works, we plan to integrate the graph structure into the design of the prompt to construct the universal task-agnostic medical image segmentation model.

Data Availability

The data used to support the findings of the study can be downloaded from the following website https://github.com/uci-cbcl/UaNet.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China with Grant no. 62106162, the China Postdoctoral Science Foundation with Grant no. 2021M692269, and the Sichuan University Postdoctoral Science Foundation with Grant no. 2022SCU12080.

References

[1] H. Tang, X. Chen, Y. Liu et al., “Clinically applicable deep learning framework for organs at risk delineation in ct images,” *Nature Machine Intelligence*, vol. 1, no. 10, pp. 480–491, 2019.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[3] Z. Yi, “nmode: neural memory ordinary differential equation,” *Artificial Intelligence Review*, vol. 56, no. 12, pp. 14403–14438, 2023.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, Berlin, Germany, 2015.

[5] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[6] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," 2017, https://arxiv.org/abs/1706.03762.

[7] A. Hatamizadeh, Y. Tang, V. Nath et al., "Unetr: transformers for 3d medical image segmentation," 2022, https://arxiv.org/abs/2103.10504.

[8] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, "Collective classification in network data," *AI Magazine*, vol. 29, no. 3, pp. 93–106, 2008.

[9] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell, "Toward an architecture for never-ending language learning," in *Proceedings of the Twenty-Fourth AAAI conference on artificial intelligence*, Atlanta, Georgia, July 2010.

[10] M. Zitnik and J. Leskovec, "Predicting multicellular function through multi-layer tissue networks," *Bioinformatics*, vol. 33, no. 14, pp. i190–i198, 2017.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017, https://arxiv.org/abs/1609.02907.

[12] Y. Li, Y. Zhang, W. Cui, B. Lei, X. Kuang, and T. Zhang, "Dual encoder-based dynamic-channel graph convolutional network with edge enhancement for retinal vessel segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 8, pp. 1975–1989, 2022.

[13] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, https://arxiv.org/abs/1312.4400.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[15] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International conference on machine learning, PMLR*, pp. 448–456, Lille, France, July 2015.

[16] 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*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[17] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *Proceedings of the International conference on machine learning, PMLR*, pp. 6105–6114, Lille, France, December 2019.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the 2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Miami, FL, USA, June 2009.

[19] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, Springer, Berlin, Germany, 2016.

[20] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, IEEE, Stanford, CA, USA, October 2016.

[21] S. F. Qadri, H. Lin, L. Shen et al., "Ct-based automatic spine segmentation using patch-based deep learning," *International Journal of Intelligent Systems*, vol. 2023, Article ID 2345835, 14 pages, 2023.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2015, https://arxiv.org/abs/1409.0473.

[23] O. Oktay, J. Schlemper, L. L. Folgoc et al., "Attention u-net: learning where to look for the pancreas," 2018, https://arxiv.org/abs/1804.03999.

[24] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, https://arxiv.org/abs/1505.00387.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, Springer, Berlin, Germany, 2018.

[26] J. Fu, J. Liu, H. Tian et al., "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3146–3154, Long Beach, CA, USA, June 2019.

[27] L. Mou, H. Fu, Y. Zhao et al. "Cs2-net: curvilinear structure segmentation network for medical images," *Medical Image Analysis*, vol. 15, Article ID 101874, 2020.

[28] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," 2018, https://arxiv.org/abs/1807.09940.

[29] D.-P. Fan, G.-P. Ji, T. Zhou et al., "Pranet: parallel reverse attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 263–273, Springer, Berlin, Germany, 2020.

[30] Y. Wang, L. Zhong, W. Huang, X. He et al., "An edge-assisted computing and mask attention based network for lung region segmentation," *International Journal of Intelligent Systems*, vol. 2023, Article ID 8589867, 13 pages, 2023.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: transformers for image recognition at scale," 2020, https://arxiv.org/abs/2010.11929.

[32] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images," 2022, https://arxiv.org/abs/2201.01266.

[33] Y. Zhang, X. Ji, W. Liu et al., "A spine segmentation method under an arbitrary field of view based on 3d swin transformer," *International Journal of Intelligent Systems*, vol. 2023,16 pages, 2023, 8686471.

[34] Z. Liu, Y. Lin, Y. Cao et al., "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, Montreal, Canada, September 2021.

[35] D. Hong, B. Zhang, H. Li et al., "Cross-city matters: a multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sensing of Environment*, vol. 299, Article ID 113856, 2023.

[36] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (exvit) for land use and land cover classification: a multimodal deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[37] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised

hyperspectral super-resolution," in *Computer Vision–ECCV 2020: 16th European Conference*, pp. 208–224, Springer, Glasgow, UK, 2020.

[38] J. Yao, X. Cao, D. Hong et al., "Semi-active convolutional neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[39] D. Hong, B. Zhang, X. Li et al., "Spectralgpt: spectral remote sensing foundation model," 2023, https://arxiv.org/abs/2311.07113.

[40] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and deep locally connected networks on graphs," 2014, https://arxiv.org/abs/1312.6203.

[41] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," 2017, https://arxiv.org/abs/1606.09375.

[42] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," 2018, https://arxiv.org/abs/1706.02216.

[43] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2018, https://arxiv.org/abs/1710.10903.

[44] S. Zheng, Z. Zhu, Z. Liu et al., "Multi-modal graph learning for disease prediction," *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2207–2216, 2022.

[45] C. Mao, L. Yao, and Y. Luo, "Imagegcn: multi-relational image graph convolutional networks for disease identification with chest x-rays," 2019, https://arxiv.org/abs/1904.00325.

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

[47] J. Hu, Y. Chen, L. Zhang, and Z. Yi, "Surrogate dropout: learning optimal drop rate through proxy," *Knowledge-Based Systems*, vol. 206, Article ID 106340, 2020.

[48] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2015, https://arxiv.org/abs/1412.6980.

[49] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images," *International MICCAI Brainlesion Workshop*, pp. 272–284, Springer, Berlin, Germany, 2021.

[50] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.