MedGNN: General Medical Image Recognition Network via GNN Visual Representations

Jiayu Ye[1], An Zeng[1,*], Dan Pan[2,*], Junhao Chen[1], Guanwei Cheng[3], and for the Alzheimer's Disease Neuroimaging Initiative

[1] School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China
[2] School of Electronics and Information, Guangdong Polytechnic Normal University, Guangzhou 510665, China
[3] Alibaba International Digital Commerce Group, Hangzhou 310023, China
zengan@gdut.edu.cn, pandan@gpnu.edu.cn

Abstract. Existing medical image representations are typically processed into grid or sequence structures via Convolutional Neural Network (CNN) or Vision Transformers. However, these methods struggle to flexibly capture irregular lesion regions and reveal relationships between lesions, especially in 3D medical imaging. To address this, we transform medical images into graph structures and propose **MedGNN**, a general recognition network based on Graph Neural Network (GNN) visual representations. We first segment the image into patches and treat each patch as a node, constructing graph visual embeddings via the K-Nearest Neighbor algorithm. Then, we propose multi-scale dynamic max-relative graph convolution for feature aggregation and updating. To mitigate over-smoothing in graph models, we design a feature-enhanced feed-forward network to refine feature representations. Experiments show that MedGNN achieves strong competitive performance across various 2D and 3D medical image recognition datasets. Moreover, it visualizes lesion relationships through graphs, enabling interpretable analysis based on graph structures. Code is available at: https://github.com/IMCTGD/MedGNN.

Keywords: Graph structure · Graph neural network · Medical image recognition · Interpretability.

1 Introduction

Voxel-based models on the basis of Convolutional Neural Networks (CNNs) [14] and Vision Transformer [3] are common deep learning paradigms in the medical imaging field, such as structural magnetic resonance imaging (sMRI) [27]. CNNs conceptualize medical images as collections of voxels arranged in rectangular forms, extracting local features through convolutional operations. In contrast, Transformers model global dependencies through self-attention, capturing long-range spatial and contextual information. Recently, Mamba [5] leverages a state-space model (SSM) to dynamically capture spatio-temporal features through a selective state mechanism, forming a hierarchical representation of medical image

inputs. Unlike natural images, the irregular and non-adjacent nature of lesion regions limits grid- or sequence-based representations in modeling their structure and relationships, which are essential for recognizing degenerative brain diseases.

Graph structures, as a general data representation, offer a potential solution to these challenges [11,13]. They can flexibly model complex objects and capture the relationships between node features, providing strong interpretability. Wang et al. [23] proposed a graph matching framework combining deep learning and combinatorial optimization, encoding graph structures into high-dimensional embeddings through neural networks to achieve high-accuracy image matching and semantic alignment. Han et al. [6] introduced the Visual Graph (ViG) model for extracting graph-level features in visual tasks, representing images as graph structures. Graph-based data also finds extensive applications in medical image analysis. Li et al. [16] developed BrainGNN, a graph neural network-based framework for functional magnetic resonance imaging (fMRI) analysis, mapping regional and cross-regional functional activation patterns for classification tasks. However, current graph-based analysis methods are mainly applied to temporal medical images like fMRI, which do not directly include high-resolution brain anatomy. Applying graph structures to anatomical imaging data, such as sMRI, generated from 3D spatial sampling remains a complex challenge.

To address the above issues, we propose MedGNN, a general recognition network for medical image analysis on the basis of GNN visual representations. First, 3D medical images are segmented into equal-sized patches, which are treated as nodes and processed into a graph structure via K-Nearest Neighbor, forming GNN-based visual embeddings (GVE). Then, we introduce multi-scale dynamic max-relative graph convolution (MGC) for feature aggregation and updating. Simultaneously, we design an extended deformable convolution (EDG) for multi-scale dynamic feature optimization. Finally, we propose a feature-enhanced feed-forward network (FFFN) to alleviate over-smoothing in the graph model and further optimize the data.

Our contributions are as follows: (1) We propose MedGNN, a general recognition network for medical images on the basis of graph neural network visual representations. MedGNN applies graph neural networks to complex 3D medical image recognition tasks, offering model interpretability that differs from traditional visual representations. (2) We introduce multi-scale dynamic max-relative graph convolution and feature-enhanced feed-forward network for medical visual representation, achieving dynamic feature optimization and enhancement. (3) MedGNN achieves competitive performance on 2D/3D medical image recognition tasks across diverse benchmarks, compared to leading models.

2 Method

In this section, we extend vision GNN techniques to the medical imaging and propose MedGNN, a GNN model for general visual representation in medical images. The overall structure is illustrated in Fig.1. For clarity, we illustrate the model design via 3D medical images.

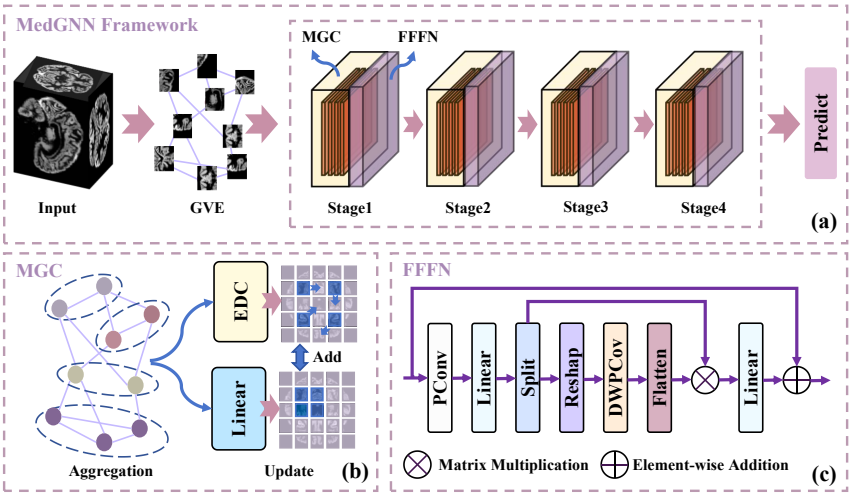

Fig. 1. Overall framework of MedGNN. (a) Overall structure, (b) MGC, (c) FFFN. The medical image input is segmented into patches, with graph-structured embeddings built via GVE. Feature aggregation and updating are then performed through MGC. Finally, the FFFN further smooths and enhances the features.

2.1 GNN-based Visual Embedding (GVE)

The function of GVE is to construct a graph-structured 3D visual embedding for medical imaging. For a medical image input $\mathbf{X} \in \mathbb{R}^{H \times W \times D \times C}$, we partition it into N patches by configuring the stem module of ConvNeXt [18]. $\mathbf{X}$ is then represented as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$, where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the voxel features of each patch. Subsequently, we treat the feature $\mathbf{x}_i$ of each patch as a node $\mathbf{v}_i$ and identify the k nearest neighbors $\mathcal{N}(v_i)$ via K-Nearest Neighbor. Meanwhile, we establish an edge e_{ij} from each node v_i to v_j, resulting in a graph representation $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E}$ denotes the set of edges. The process of graph construction is represented by $\mathcal{G} = G(X)$. For 3D medical images, calculating the distance between nodes involves extensive voxel interactions. KNN reduces computational complexity by focusing only on the k nearest neighbors of each point. Hence, we choose KNN to compute the distance relationships between nodes.

2.2 Multi-Scale Dynamic Max-Relative Graph Convolution (MGC)

Since the relationships between lesions are not simply on the basis of adjacency, we propose MGC (Multi-scale Graph Convolution) for multi-scale feature aggregation and update, enabling more comprehensive exploration of node (i.e., lesion) relationships. MGC can be divided into two steps: **Aggregate** and **update**, which involve aggregating features from neighboring nodes and updating the node representation with the aggregated feature:

$$G' = \text{Update}\left(\text{Aggregate}\left(G, W_{\text{agg}}\right), W_{\text{update}}\right), \tag{1}$$

here, Aggregate$(\cdot)$ denotes feature aggregation, Update$(\cdot)$ represents feature update, W_{agg} and W_{update} correspond to the weights of aggregation and update, respectively. Specifically, assuming $\mathcal{N}(x_i)$ is the set of neighboring nodes for a given node, the feature update process for that node can be expressed as:

$$\mathbf{x}_i' = \text{Update}\left(x_i, \text{Aggregate}\left(x_i, \mathcal{N}(x_i), W_{\text{agg}}\right), W_{\text{update}}\right). \tag{2}$$

Aggregate: We employ max-relative graph convolution [15] as the feature aggregation method due to its flexibility in adapting to irregular graph structures and its efficiency in capturing relative relationships between neighboring nodes and the central node:

$$\text{Aggregate}(\cdot) = \mathbf{x}_i'' = \left[\mathbf{x}_i, \max\left(\{\mathbf{x}_j - \mathbf{x}_i \mid j \in \mathcal{N}(x_i)\}\right)\right], \tag{3}$$

here, $\mathbf{x}_j$ represents the features of neighboring nodes of node i, and $[\cdot]$ denotes the concatenation of two vectors.

Update: To achieve multi-scale dynamic feature updates, we design an Extended Deformable Convolution (EDC) on the basis of feature linear mapping, adapting to the diverse data characteristics of medical images and recognizing lesion boundaries. This process can be expressed as:

$$\text{Update}(\cdot) = \mathbf{x}_i' = \alpha \sum_{k_i \in \mathcal{K}_l} W_{\text{L}}^{k_i} \cdot \mathbf{x}_i''(p_{k_s}) + (1-\alpha) \sum_{k_d \in \mathcal{K}_d} W_{\text{D}}^{k_d} \cdot \mathbf{x}_i''(p_0 + p_{k_d} + \Delta p_{k_d}), \tag{4}$$

here, W_{L} and W_{D} represent the weights of the linear mapping and deformable convolution [2], p_0 denotes the center point location, p_k is the sampling position, and Δp_k is the dynamic offset. Notably, we employ convolutional kernels of varying sizes k_d for multi-scale feature updates. Additionally, we introduce a weighting coefficient $\alpha \in [0, 1]$ to perform a weighted fusion of the two feature components, balancing the importance of different features.

2.3 Feature-enhanced Feed-forward Network (FFFN)

Inspired by the FFN [3] in Transformer, we design FFFN to mitigate oversmoothing in graph models and further enhance data augmentation. Specifically, FFFN consists of PConv [1] and a gating mechanism. First, PConv is used to perform a linear transformation on the input feature $\mathbf{X}' \in \mathbb{R}^{H \times W \times D \times C}$, followed by a linear projection via the weights W_1:

$$\bar{X}' = GELU\left(W_1 P\,\text{Conv}\left(X'\right)\right), \tag{5}$$

here, $GELU(\cdot)$ represents the GELU activation function. Then, $\bar{X}'$ is decomposed into $\bar{X}_1'$ and $\bar{X}_2'$ along the channel dimension, with local features extracted through depthwise convolution (DWConv):

$$\bar{X}_r' = \bar{X}_1' \otimes F\left(DW\,\text{Conv}\left(R\left(\bar{X}_2'\right)\right)\right), \tag{6}$$

here, $R(\cdot)$ and $F(\cdot)$ represent the reshape and flatten operations, respectively, while $\otimes$ denotes matrix multiplication. Finally, $\bar{X}_r'$ undergoes linear projection via the weights W_2, followed by the GELU activation function to obtain the output feature $\bar{X}_{\text{out}}'$. We add FFFN at each node, combining it with MGC.

3 Experiment and Results

The 3D medical image datasets are sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [4], Open Access Series of Imaging Studies (OASIS) [5], and Autism Neuroimaging Data Exchange I (ABIDE I) [6]. **ADNI** includes 446 AD subjects, 450 HC subjects, and 493 MCI subjects. **OASIS** includes 196 subjects are selected, including 105 AD subjects and 91 HC subjects. **ABIDE I** includes 78 autism spectrum disorder (ASD) subjects and 78 HC subjects. **All datasets use T1-weighted imaging. The goal of this study is early screening of patients, thus we use the first recorded sMRI data for each subject.** Preprocessing is performed via the open-source tool Computational Anatomy Toolbox (CAT12), with image size set to 112×128×112.

The 2D medical image datasets are sourced from SARS-CoV-2 [7] and Chest X-Ray Images (Pneumonia) [8]. The **SARS-CoV-2** dataset consists of CT scans including 1,252 pneumonia patients and 1,229 non-pneumonia patients. The size of each CT slice ranges from 119×104 to 416×512. The **Chest X-Ray Images (Pneumonia) dataset** containing 5,863 chest X-ray images of pneumonia and non-pneumonia patients.

Implementation Details. MedGNN and the comparison models in this study are implemented via the PyTorch framework and trained and tested on five 4090Ti GPUs. The learning rate is set to 1e-3, with the Adam optimizer and a weight decay of 5e-2. The weight coefficient α in MGC is 0.9, and the number of basic blocks follows the configuration 2:2:18:2. **The parameter settings for MedGNN are identical across both 2D and 3D datasets.** For more detailed parameters, please refer to the provided open-source code. The performance metrics include accuracy and F1 score. Experimental results are derived from five-fold cross-validation.

3.1 Experimental Results and Ablation Studies

Experimental Results. The experimental results are presented in Table 1 and Table 2. It is important to note that the comparison models use the same parameter and structural configurations. For 3D medical images, MedGNN achieves optimal performance across different datasets. Specifically, it performs best in the AD vs NC (ADNI) task, with an accuracy of 0.911 and an F1 score of 0.911. It also delivers competitive performance in the AD vs HC (OASIS) task, improving the accuracy by 0.052 and the F1 score by 0.059 compared to the second-best model. For 2D medical images, MedGNN achieves optimal performance in SARS-CoV-2 and Chest X-Ray Images (Pneumonia), with accuracies of 0.947 and 0.849 and F1 scores of 0.947 and 0.784, respectively. However, it does not show a significant advantage over the 3D medical image datasets.

[4] ADNI.loni.usc.edu
[5] http://www.oasis-brains.org
[6] https://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html
[7] https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset
[8] https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

Table 1. Performance comparison of MedGNN with other recognition models across various 3D medical image datasets, where $*$ indicates models derived from segmentation task models, _ denotes the second-best result.

	AD vs HC (ADNI)		**AD vs MCI** (ADNI)		**AD vs HC** (OASIS)		**ASD vs HC** (ABIDE I)	
Model	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
CNN-based								
3D ResNet [8]	0.885	0.885	0.803	0.803	<u>0.771</u>	0.760	<u>0.711</u>	<u>0.705</u>
Biceph-Net [21]	0.819	0.816	0.747	0.745	0.704	0.712	0.605	0.603
DA-MIDL [28]	0.862	0.861	0.788	0.787	0.733	0.732	0.638	0.625
Gao et al. [4]	0.870	0.870	0.793	0.793	0.742	0.736	0.655	0.647
Xing et al. [25]	0.848	0.846	0.777	0.776	0.698	0.704	0.638	0.626
Transformer-based								
Addformer [12]	0.865	0.864	0.778	0.777	0.723	0.720	0.644	0.635
VT-UNet* [19]	0.888	0.888	0.811	0.811	0.752	0.748	0.661	0.647
UNETR* [7]	0.883	0.883	0.801	0.801	0.733	0.728	0.661	0.644
M3T [9]	0.882	0.882	0.802	0.802	0.761	0.759	0.644	0.633
Other-based								
Pointnet [20]	0.852	0.850	0.780	0.779	0.733	0.729	0.633	0.626
Segmamba* [26]	0.892	0.891	0.815	0.815	0.761	0.760	0.699	0.689
TransBTS* [24]	<u>0.894</u>	<u>0.894</u>	<u>0.819</u>	<u>0.818</u>	0.766	<u>0.763</u>	0.688	0.682
GNN-based								
MedGNN (3D)	**0.911**	**0.911**	**0.837**	**0.841**	**0.823**	**0.822**	**0.727**	**0.724**
Improvement	↑1.7%	↑1.7%	↑1.8%	↑2.3%	↑5.2%	↑5.9%	↑1.6%	↑1.9%

Definition: AD (Alzheimer's disease), HC (Healthy control), MCIc (MCI patients who will convert to AD), and MCInc (MCI patients who will not convert to AD).

Ablation Studies. (1) The impact of the weight coefficient α in MGC on experimental performance is shown in Table 3. When α=0.9, the best performance is achieved in AD vs NC (ADNI), AD vs HC (OASIS), and ASD vs HC (ABIDE I). Additionally, it yields the second-best accuracy and the best F1 score in AD vs MCI (ADNI). **(2) The impact of different stage configurations on experimental performance is shown in Fig.2.** We provide four MedGNN models with varying stage configurations. Since [18] demonstrated that a higher proportion of stacked blocks in stage 3 leads to better results, we focus primarily on adjusting stage 3. The four MedGNN configurations are: 1:1:3:1 (MedGNN-T), 1:1:9:1 (MedGNN-S), 2:2:18:2 (MedGNN-B), and 3:3:27:3 (MedGNN-L). MedGNN-B and MedGNN-L consistently achieve optimal performance across different tasks and can be selected based on the balance between performance and computational requirements in practical scenarios.

Visualization Analysis. In this section, we provide a visualization of the MedGNN graph structure. Fig. 3 illustrates the graphs of two samples from stage 1 and stage 4 in both ADNI and SARS-CoV-2 datasets. The purple circle

Table 2. Performance comparison of MedGNN with various recognition models across different 2D medical image datasets.

	SARS-CoV-2		Chest X-Ray Images	
Model	**ACC**	**F1**	**ACC**	**F1**
2D ResNet [8]	0.909	0.905	0.801	0.673
ConvNext [18]	0.919	0.917	0.814	0.693
Jangam et al. [10]	0.918	0.916	0.798	0.650
ViT [3]	0.934	0.934	0.805	0.668
Swin transformer [17]	0.942	0.942	0.835	0.748
Mamba [5]	0.939	0.939	0.807	0.668
MedGNN (2D)	**0.947**	**0.947**	**0.849**	**0.784**
Improvement	↑0.5%	↑0.5%	↑1.4%	↑3.6%

Table 3. The impact of the weight coefficient α in MGC on experimental performance.

	AD vs HC (ADNI)		**AD vs MCI** (ADNI)		**AD vs HC** (OASIS)		**ASD vs HC** (ABIDE I)	
α	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
0.9	**0.911**	**0.911**	0.837	**0.841**	**0.823**	**0.822**	**0.727**	**0.724**
0.7	0.904	0.904	**0.838**	0.838	0.814	0.814	0.722	0.722
0.5	0.910	0.910	0.837	0.833	0.819	0.818	0.722	0.718
0.3	0.909	0.908	0.830	0.830	**0.823**	0.820	0.722	0.718

represents the central node, and the red connecting lines indicate the neighboring nodes. Experimental validation leads to the following conclusion: as the model depth increases, the neighbors of the central node become more semantically meaningful. In the ADNI dataset, the patch containing the node is more strongly correlated with the functional features of the brain region of the central node. According to AAL [22], we found significant graph connectivity features in regions such as the rostral lingual gyrus (rLinG) and nucleus accumbens (NAC), which show a high correlation with clinical findings. In the SARS-CoV-2 dataset, MedGNN demonstrates associations between lesions, such as ground-glass opacities, across different regions, and to some extent, establishes semantic connections within the graph structure for the same type of lesion.

4 Discussion and Conclusion

Advantages: (1) Since MedGNN does not rely on ROIs, it allows graph construction without incorporating prior domain knowledge. (2) MedGNN is a general-purpose backbone for medical imaging that can be directly applied to 2D or 3D data without requiring complex structural modifications or adaptation. (3)

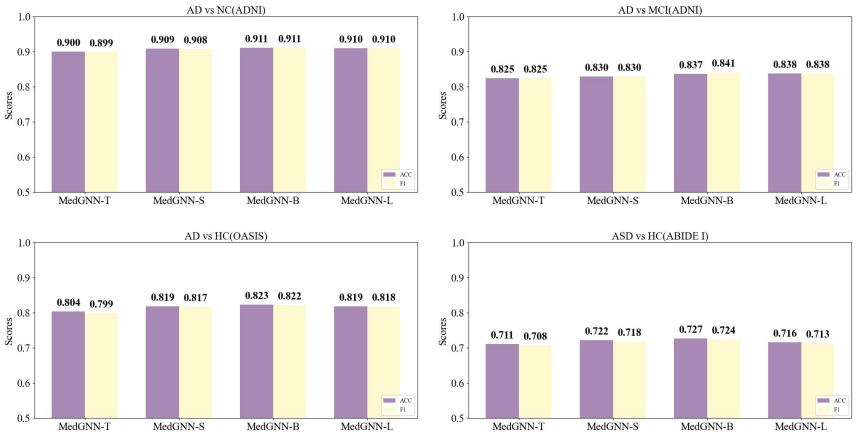

Fig. 2. The effect of different stage configurations of MedGNN on performance.

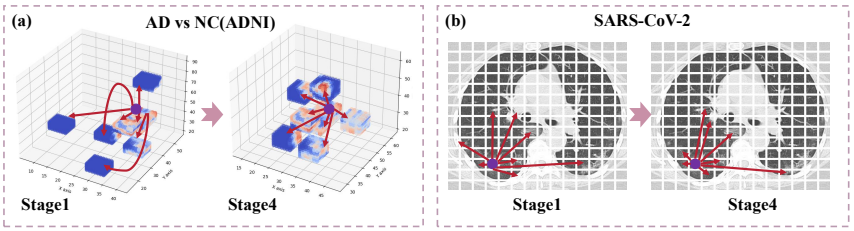

Fig. 3. An example of MedGNN's graph structure visualization. The purple circle represents the central node, and the red connecting lines indicate the neighboring nodes. For clarity, only one central node is displayed.

MedGNN operates directly on voxels, eliminating the need for handcrafted features and complicated data preprocessing commonly seen in conventional models. (4) Its multi-scale dynamic feature updating enables more effective capture and representation of critical lesion-related information.

Limitations: Although MedGNN demonstrates highly competitive performance, it still faces several issues. First, MedGNN performs poorly in medical image segmentation tasks. Graph-based models primarily rely on node-level or graph-level supervision, making it challenging to optimize for fine-grained segmentation tasks. Second, the ability to capture fine-grained brain region features remains insufficient. Despite the design of FFFN to mitigate excessive smoothing in the graph model, some information loss still occurs during the extraction of small-scale lesion/brain region features. In future work, we will optimize for these issues and expand the model to more medical image datasets and tasks.

Conclusion: In this paper, we propose MedGNN, a general medical image recognition network based on graph neural network visual representations. Unlike current mainstream visual representation models, MedGNN divides images into blocks and utilizes a graph structure for aggregation and updating, enabling

flexible feature extraction. Additionally, we introduce a feature-enhanced feed-forward network to mitigate excessive smoothing in the graph model, achieving dynamic feature optimization and enhancement. Experimental results demonstrate that MedGNN exhibits highly competitive performance on 2D and 3D medical image recognition datasets and provides interpretable analysis through the graph, revealing associations between lesions/brain regions.

Acknowledgments. This study was supported by Guangdong S&T Program (Grant Nos. 2024B1111140001), NSF of China (Grant Nos. 61976058 and 61772143), Science and Technology Planning Project of Guangdong (Grant Nos. 2021B0101220006, 2021A1515012300), in part by Alzheimer's Disease Neuroimaging Initiative (ADNI), in part by the National Institutes of Health under Grant U01 AG024904 and in part by the DOD ADNI, Department of Defense under Grant W81XWH-12-2-0012.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, J., Kao, S.h., He, H., Zhuo, W., Wen, S., Lee, C.H., Chan, S.H.G.: Run, don't walk: chasing higher flops for faster neural networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 12021–12031 (2023)
2. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 764–773 (2017)
3. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Gao, X., Cai, H., Liu, M.: A hybrid multi-scale attention convolution and aging transformer network for alzheimer's disease diagnosis. IEEE J. Biomed. Health Inform. **27**(7), 3292–3301 (2023)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
6. Han, K., Wang, Y., Guo, J., Tang, Y., Wu, E.: Vision gnn: An image is worth graph of nodes. Adv. Neural Inf. Process. Syst. **35**, 8291–8303 (2022)
7. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. pp. 574–584 (2022)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 770–778 (2016)
9. Jang, J., Hwang, D.: M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 20718–20729 (2022)
10. Jangam, E., Annavarapu, C.S.R.: A stacked ensemble for the detection of covid-19 with high recall and accuracy. Comput. Biol. Med. **135**, 104608 (2021)
11. Jing, Y., Mao, Y., Yang, Y., Zhan, Y., Song, M., Wang, X., Tao, D.: Learning graph neural networks for image style transfer. In: Eur. Conf. Comput. Vis. pp. 111–128. Springer (2022)

12. Kushol, R., Masoumzadeh, A., Huo, D., Kalra, S., Yang, Y.H.: Addformer: Alzheimer's disease detection from structural mri using fusion transformer. In: Int. Symp. Biomed. Imaging. pp. 1–5. IEEE (2022)
13. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 4558–4567 (2018)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE. **86**(11), 2278–2324 (1998)
15. Li, G., Muller, M., Thabet, A., Ghanem, B.: Deepgcns: Can gcns go as deep as cnns? In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 9267–9276 (2019)
16. Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S.: Braingnn: Interpretable brain graph neural network for fmri analysis. Med. Image Anal. **74**, 102233 (2021)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 10012–10022 (2021)
18. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 11976–11986 (2022)
19. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A robust volumetric transformer for accurate 3d tumor segmentation. In: Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 162–172. Springer (2022)
20. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 652–660 (2017)
21. Rashid, A.H., Gupta, A., Gupta, J., Tanveer, M.: Biceph-net: A robust and lightweight framework for the diagnosis of alzheimer's disease using 2d-mri scans and deep similarity learning. IEEE J. Biomed. Health Inform. **27**(3), 1205–1213 (2022)
22. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M.: Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. Neuroimage. **15**(1), 273–289 (2002)
23. Wang, R., Yan, J., Yang, X.: Learning combinatorial embedding networks for deep graph matching. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 3056–3065 (2019)
24. Wenxuan, W., Chen, C., Meng, D., Hong, Y., Sen, Z., Jiangyun, L.: Transbts: Multimodal brain tumor segmentation using transformer. In: Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 109–119 (2021)
25. Xing, X., Liang, G., Blanton, H., Rafique, M.U., Wang, C., Lin, A.L., Jacobs, N.: Dynamic image for 3d mri image alzheimer's disease classification. In: Eur. Conf. Comput. Vis. pp. 355–364. Springer (2020)
26. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 578–588. Springer (2024)
27. Ye, J., Zeng, A., Pan, D., Zhang, Y., Zhao, J., Chen, Q., Liu, Y.: Mad-former: A traceable interpretability model for alzheimer's disease recognition based on multi-patch attention. IEEE J. Biomed. Health Inform. **28**(6), 3637–3648 (2024)
28. Zhu, W., Sun, L., Huang, J., Han, L., Zhang, D.: Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural mri. IEEE Trans. Med. Imaging. **40**(9), 2354–2366 (2021)