Anatomical Graph-based Multilevel Distillation for Robust Alzheimer's Disease Diagnosis with Missing Modalities

Fei Liu[1], Huabin Wang[2(✉)], Mohamed Hisham Jaward[3], Shiuan-Ni Liang[1,4], Huey Fang Ong[5], and Jiayuan Cheng[6]

[1] Department of Electrical and Robotics Engineering, School of Engineering, Monash University Malaysia, Bandar Sunway, Malaysia
[2] School of Computer Science and Technology, Anhui University, Hefei, China
[3] School of Technology, Business, and Arts, University of Suffolk, Ipswich, UK
[4] Medical Engineering and Technology Hub, School of Engineering, Monash University Malaysia, Bandar Sunway, Malaysia
[5] School of Information Technology, Monash University Malaysia, Bandar Sunway, Malaysia
[6] School of Information Engineering, Anhui Finance and Trade Vocational College, Hefei, China

fei.liu1@monash.edu; wanghuabin@ahu.edu.cn

Abstract. The multimodal model has shown superior potential for accurate Alzheimer's disease (AD) diagnosis; however, its reliance on complete modalities limits its use in a clinical setting. This study proposes a novel **A**natomical **G**raph-based **M**ultilevel **D**istillation (AGMD) framework that effectively transfers multimodal knowledge using layered modeling. Specifically, we develop a hierarchical distillation framework with three dedicated branches to explicitly capture the features of AD from multiple levels (local structural details, regional connectivity patterns, and global semantic information) to achieve complete knowledge transfer. Moreover, we introduce anatomical constraints to model the brain adjacent connection patterns to help better learn the relationships between key ROIs, particularly in disease-relevant regions, e.g., the hippocampus. The prediction entropy as regularization is introduced to refine instance-level knowledge, comprehensively alleviating the negative impact of the teacher's noisy information. Extensive experiments on the ADNI dataset demonstrate that AGMD achieves the best classification accuracy, with an improvement of 3.7% over the state-of-the-art methods, while significantly reducing the performance gap between teacher and student models. The code is available at https://github.com/LiuFei-AHU/AGMD.

Keywords: Alzheimer's Disease Diagnosis · Multilevel Distillation · Graph Distillation · Uncertainty Evaluation · MRI

1 Introduction

Alzheimer's disease (AD), the most prevalent neurodegenerative disorder affecting more than 50 million people worldwide [8], is characterized by progressive cognitive decline and irreversible brain atrophy. Early diagnosis and progression prediction (e.g., distinguishing progressive mild cognitive impairment (pMCI) from stable MCI (sMCI)) are critical for timely intervention [22]. Although multimodal neuroimaging combining Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) achieves superior diagnostic performance [15,16], insufficient data is a critical challenge in clinical practice: the frequent unavailability of PET due to cost constraints or safety concerns [17]. Models relying solely on MRI suffer severe performance degradation because they fail to capture disease-related metabolic patterns.

Knowledge distillation (KD) offers a potential solution by transferring knowledge from a multimodal teacher (e.g., trained with MRI and PET) to a student model using only MRI [7]. Existing KD methods focus on aligning logits or features between models. Logit distillation [9] is simple but transfers insufficient information. Thus, it has limited application, particularly for complex tasks such as the early diagnosis of AD. Feature distillation [14,5,10,13,24] improves flexibility by distilling multiscale intermediate or global features but struggles to filter task-irrelevant noise (e.g., non-disease-related anatomical variations). Attention-based distillation [25,20,23,12] partially addresses this by emphasizing discriminative regions. In addition, adversarial and contrastive distillation strategies [21,1] were also proposed to improve the robustness of the student model. Although existing distillation methods achieve promising results in natural images, their direct application to AD diagnosis remains suboptimal due to two issues: (1) **Ignoring hierarchical pathology**: AD manifests multilevel interactions, such as local tissue atrophy (e.g., hippocampus), interrupted interregional connectivity, and global dysfunction [2]. Single-level knowledge modeling [14,5,25,20,10,13,23,24] fails to capture the hierarchical pathology of AD, as evidenced by recent neuropathological study [19]. (2) **Anatomically implausible representations**: Graph-based methods [11,4,26] capture brain connectivity but rely on simple graph construction (e.g., correlation-based edges), neglecting anatomical priors (e.g., white matter tracts between ROIs). Without explicit anatomical constraints, it is challenging to construct a complete and rational pathological distribution network, and the learned features may lack biological interpretability.

To address these limitations, we propose a novel **A**natomical **G**raph-based **M**ultilevel **D**istillation (AGMD) framework, specifically designed for the diagnosis of AD. Our main contributions include the following: (1) **Hierarchical pathology modeling**: Our AGMD explicitly encodes AD pathology at three levels: local (3D CNN for ROI atrophy), regional (graph convolution for inter-ROIs connectivity), and global (transformer for the semantic context of the brain) inspired by neuropathological studies [19]. (2) **Anatomically constrained graph learning**: To the best of our knowledge, we are the first to construct anatomically guided graphs using structural connectivity derived from

AD-aware brain regions with their adjacent neighbors, ensuring that the edges reflect biologically plausible pathways between disease-relevant ROIs (e.g., hippocampus). Furthermore, we introduce an uncertainty-aware gating mechanism to dynamically adjust instance-level distillation based on the teacher model's prediction entropy estimates, prioritizing reliable teacher predictions.

Experiments on the ADNI dataset demonstrate that our AGMD achieves state-of-the-art (SOTA) accuracy (75.9% on AD/pMCI/sMCI/NC classification), outperforming previous distillation methods by 3.7% and significantly reducing the teacher-student performance gap with ablation studies validating the necessity of each component.

2 Methodology

2.1 Proposed Framework

We propose a novel **A**natomical **G**raph-based **M**ultilevel **D**istillation (AGMD) framework to transfer multimodal knowledge from a teacher model (trained on MRI and PET) to a student model (using MRI only). Fig. 1 illustrates the overall framework. In particular, our AGMD includes three main components: 1) **Cross-Modal Attentive Fusion Transformer (CMT)** for teacher model's shallow features fusion, 2) **Anatomical-guided graph learning** for brain connectivity modeling, and 3) **Uncertainty-Aware Gating (UAG)** for dynamic knowledge refinement.

The teacher model takes MRI ($\mathbf{X}_{\text{MRI}} \in \mathbb{R}^{H\times W\times D}$) and PET ($\mathbf{X}_{\text{PET}} \in \mathbb{R}^{H\times W\times D}$) as inputs, while the student model takes only MRI. Both models use 3D convolutional encoders to extract multiscale features: $\{F^l_{\text{MRI}}, F^l_{\text{PET}}\}^K_{l=1}$, where $F^l \in \mathbb{R}^{C_l\times H_l\times W_l\times D_l}$ is the feature map at layer l, K is the total number of layers, while C_l is the channel number. For the teacher model, each layer of features of $\mathbf{X}_{\text{MRI}}$ and $\mathbf{X}_{\text{PET}}$ are fused as F^l_{fused} in the **CMT**, then the last two layers' features $(F^{K-1}_{\text{fused}}, F^K_{\text{fused}})$ are input into the **Anatomical-guided Graph Learning** to construct the brain graphs to model brain connectivity. Finally, the classifier will output the predicted disease label $\hat{y_T}$ on the fused brain graphs. For the student model, it has the same workflow but without the **CMT** module. We evaluate the teacher model's prediction entropy in **UAG** to refine the knowledge when training the student model.

2.2 Cross-Modal Attentive Fusion Transformer (CMT)

For each layer of MRI and PET features of the teacher model, we first calculate their joint channel attention: $C_{\text{Attn}} = \text{Attention}(F^l_{\text{MRI}} \oplus F^l_{\text{PET}})$, then $F^l_{\text{MRI}} = C_{\text{Attn}} \odot F^l_{\text{MRI}}$. Next, cross-attention is utilized to capture associations across MRI and PET. Specifically, for layer l, the MRI and PET features are fused via:

$$F^l_{\text{fused}} = \text{Conv}(\text{Attn}^l \otimes V^l + F^l_{\text{MRI}}), \quad \text{Attn}^l = \text{Softmax}\left(\frac{Q^l(K^l)^\top}{\sqrt{C_l}}\right), \tag{1}$$

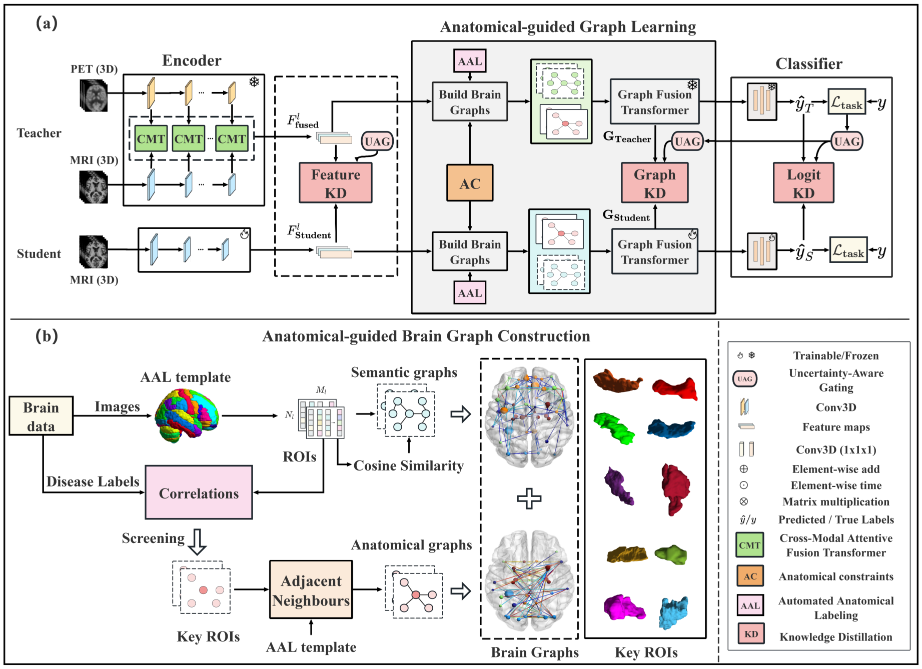


Fig. 1: (a) The architecture of the proposed AGMD: local structural details extracted by a 3D CNN **Encoder** with a **Cross-Modal Attentive Fusion Transformer (CMT)**, brain connectivity patterns, and global semantic context through the **Anatomical-guided Graph Learning**, while **Uncertainty-Aware Gating (UAG)** dynamically refines instance-level knowledge. (b) The semantic and anatomical graphs are constructed based on the brain template and anatomical constraints.

where $Q^l = \text{Conv}(F^l_{\text{MRI}}), K^l = \text{Conv}(F^l_{\text{PET}}), V^l = \text{Conv}(F^l_{\text{PET}})$ are query, key, and value matrices for cross-modal interaction, Attn^l is the attention weights highlighting PET-informed regions, and $F^l_{\text{fused}}$is the fused features retaining MRI structural details and focusing on relevant functional areas.

We then transfer the learned cross-modal associations to the student model, alleviating the impact of missing modality and improving robustness. To achieve this, we align student features F^l_{Student} with F^l_{fused} via:

$$\mathcal{L}_{\text{shallow}} = \sum_{l=1}^{K} \alpha_l \cdot \|F^l_{\text{fused}} - F^l_{\text{Student}}\|_2^2, \tag{2}$$

where α_l is a layer-wise weight used to constrain the learning of the student model.

2.3 Anatomical-guided Graph Learning

Semantic Graph Construction. Following [3], the ROI features are extracted from the encoder guided by the Automated Anatomical Labeling (AAL) template [18] for the last two layers of the encoder. The AAL template is downsampled to match the dimension of $F^l \in \mathbb{R}^{H^l \times W^l \times D^l}$. Then, we construct the brain semantic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes $\mathcal{V}$ and edges $\mathcal{E}$ represent brain ROIs and regional connectivity, respectively. Cosine similarity is computed between node features, and edges are established using KNN (in our experiments, the $k = \sqrt{N_\mathcal{V}}$, $N_\mathcal{V}$ is the number of nodes):

$$\text{Cosine}(v_i, v_j) = \frac{\mathbf{h}_i^l \cdot \mathbf{h}_j^l}{\|\mathbf{h}_i^l\| \|\mathbf{h}_j^l\|}, \tag{3}$$

$$\mathbf{A}_{\text{semantic}}^l(i,j) = \begin{cases} 1 & \text{if } v_j \in \text{top-}k \text{ neighbors of } v_i, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\text{Cosine}(v_i, v_j)$ calculates the feature similarity between nodes $(v_i, v_j) \in \mathcal{V}$, $\mathbf{h}_i^l \in \mathbb{R}^{C_l}$ is the feature of i^{th} node, and $\mathbf{A}_{\text{semantic}}^l(i,j) = 1$ represents there is an edge between nodes (v_i, v_j).

Anatomical Graph Construction. To create a graph structure that better reflects the characteristics of the brain network, we introduce anatomical constraints that can effectively learn the relationships between key ROIs. According to the constructed semantic graph, we select the top M disease-sensitive nodes (e.g., hippocampus) based on the contributions to AD, where M is the number of key ROIs and is set to 10 in our experiment. Then, we screen neighbors for disease-sensitive nodes to construct the spatially adjacent graph. Specifically, for each node v_i, we identify adjacent ROIs based on the AAL template and the adjacency of brain regions, and then edges between these adjacent ROIs are added to the dynamic graph:

$$\mathbf{A}_{\text{anatomical}}^l(i,j) = \begin{cases} 1 & \text{if adjacent}(v_i, v_j) = True, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Graph Fusion and Learning. The semantic and anatomical graphs are fused into a complete brain graph to encode semantic and anatomical connections in each layer: $\mathbf{A}_{\text{fused}}^l = \mathbf{A}_{\text{semantic}}^l + \mathbf{A}_{\text{anatomical}}^l$. Then, we have brain graphs in multiple scales, better reflecting the complex connectivity patterns of the brain network, which can be learned using GCNs:

$$\mathbf{H}^{l+1} = \text{ReLU}\left(\mathbf{A}_{\text{fused}}^l \mathbf{H}^l \mathbf{W}^l\right), \tag{6}$$

where $\mathbf{H}^l$ is the node feature matrix at layer l, and $\mathbf{W}^l \in \mathbb{R}^{C_l \times C_{l+1}}$ is a learnable weight matrix. For each layer, the teacher and student's graph representations are aligned via:

$$\mathcal{L}_{\text{graph}} = \sum_{l} \beta_l \cdot \|\mathbf{H}^l_{\text{Teacher}} - \mathbf{H}^l_{\text{Student}}\|^2_F, \tag{7}$$

where β_l is the layer-wise weight used to constrain the learning of the student model.

After that, these multiscale graphs are fused via a Graph Transformer to model cross-graph interactions, using multi-head attention to capture global dependencies: $\mathbf{G} = \text{GraphTransformer}\left(\mathbf{H}^{l-1}, \mathbf{H}^l\right)$. Then, we use $\mathcal{L}_{\text{global}}$ to minimize the cosine distance of global brain networks between the teacher and student models:

$$\mathcal{L}_{\text{global}} = 1 - \frac{\mathbf{G}_{\text{Teacher}} \cdot \mathbf{G}_{\text{Student}}}{\|\mathbf{G}_{\text{Teacher}}\|\|\mathbf{G}_{\text{Student}}\|}. \tag{8}$$

2.4 Uncertainty-Aware Gating (UAG)

The gating mechanism can adaptively prioritize reliable knowledge transfer during distillation. We dynamically adjust distillation weights to refine knowledge transfer according to the teacher's uncertainty. Specifically, we compute the entropy-based confidence weight w from the teacher's Softmax probabilities p:

$$w = 1 - \frac{-\sum_{c=1}^{N_{\text{class}}} p_c \log p_c}{\log N_{\text{class}}}, \tag{9}$$

where N_{class} is the number of classification groups.

Then, the confidence weight is applied to the distillation losses. The total loss combines shallow, regional, and global distillation losses with uncertainty weights:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + w \cdot \sum_{l} \mathcal{L}_l, \tag{10}$$

where $\mathcal{L}_{\text{task}}$ is the cross-entropy loss and $l \in \{\text{shallow, graph, global}\}$.

3 Experiments and Results

Dataset and Preprocessing. We collected multi-center 3D neuroimaging data (MRI and PET) from the ADNI1-2 databases, comprising 269 subjects from four groups: AD, pMCI, sMCI, and NC (89, 38, 64, and 78, respectively). We chose 1.5T / 3T T1-weighted MRI and 18F-FDG PET for this study. Following the common practice outlined in [6], the MRI images were processed using a standard pipeline, which mainly included tissue segmentation, normalization to the MNI152 space, and smoothing. The paired MRI and PET images are spatially aligned using the SPM12 tools, and the final data is resampled to $113 \times 113 \times 137$.

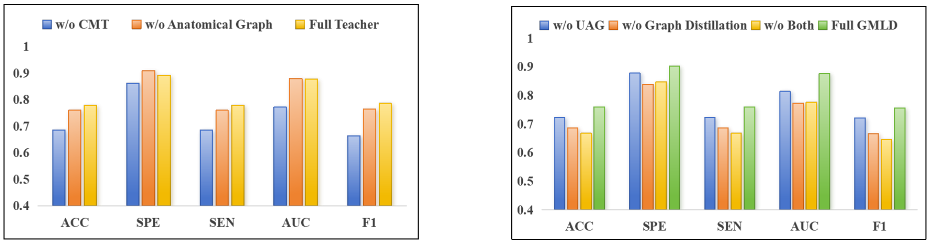
Experimental Setup. We evaluated our model on the AD/pMCI/sMCI/NC classification task. Evaluation metrics include accuracy (ACC), specificity (SPE), sensitivity (SEN), AUC, and weighted-F1 to address class imbalance. The data is divided into three sets: training, validation, and testing (7:1:2). We train the model on a 3090 NVIDIA GPU for 150 epochs with Adam optimizer, and the learning rate is set to 1e-5. We compared the results of our model with those of various state-of-the-art methods to verify its superior performance (see Table 1 for details). Among them, "Baseline" represents the traditional convolutional model using MRI and PET, "Graph Transformer" is a graph learning-based method [4] using MRI and PET, "Teacher (Ours)" represents the multimodal teacher model, and "AGMD (Ours w/o KD)" is the model using the same architecture as the student model without distillation. We further compared with existing distillation methods, such as traditional logit distillation (KD) [9], three feature distillation methods: shallow feature, global feature and cross-layer distillation (FD [12], GFD [14], and CFD [5,24]), one attention distillation (AFD) [23], one adversarial contrastive distillation (ACD) [21], and a graph distillation (GD) [11]. We ran all the methods under the same conditions for a fair comparison. In addition, we conducted ablation experiments to prove the validity of each module of our AGMD, as shown in Fig. 2.

Table 1: Performance comparison on ADNI dataset

Method	**Modality**	**ACC↑**	**SPE↑**	**SEN↑**	**AUC↑**	**F1↑**
Baseline	MRI+PET	0.6111	0.7937	0.6111	0.7008	0.5044
Graph Transformer [4]	MRI+PET	0.7407	0.8794	0.7407	0.8406	0.7217
Teacher (Ours)	MRI+PET	0.7778	0.8917	0.7778	0.8770	0.7851
Logit [9]	MRI	0.6481	0.8610	0.6481	0.7720	0.6595
Shallow Features[12]	MRI	0.6481	0.8700	0.6481	0.7120	0.6571
Global Features [14]	MRI	0.6296	0.8180	0.6296	0.7447	0.6011
Attentive Features [23]	MRI	0.6481	0.8179	0.6481	0.7081	0.6350
Cross-layer Features [5,24]	MRI	0.6296	0.8451	0.6296	0.7956	0.6239
Adversarial Contrastive [21]	MRI	0.7222	0.8805	0.7222	0.8610	0.7242
Brain Subgraphs [11]	MRI	0.7037	0.8886	0.7037	0.8244	0.7225
AGMD (Ours w/o KD)	MRI	0.6481	0.8146	0.6481	0.7376	0.5293
AGMD (Ours)	MRI	**0.7593**	**0.9014**	**0.7593**	**0.8760**	**0.7560**

3.1 The performance valuation on the AD classification task

Table 1 compares the performance of the teacher model, the student model, and various distillation methods. The performance upper bound is given by our teacher model, which achieved the highest ACC (0.7778) and weighted F1 (0.7851), validating its ability to extract intricate pathological patterns related to AD. The student model without distillation obtains lower performance (ACC = 0.6481, F1 = 0.5293), indicating that it fails to capture complete disease-specific features using only MRI. In addition, the baseline model achieves the worst performance, demonstrating its low ability to capture complex brain connections with only CNNs. The proposed AGMD achieves ACC (0.7593) and F1 (0.756) that are close to the teacher model's performance, and significantly outperform other distillation approaches (e.g., logit distillation [9]: F1 = 0.6595). Compared with traditional methods, the graph models [11,4] and the adversarial and contrastive distillation method [21] obtain better classification results, indicating their better ability for AD diagnosis, particularly the graph-based models [11,4] show superior performance than other methods because of their ability in learning brain connections. However, our AGMD achieves the highest SPE (0.9014) and AUC (0.876), demonstrating its ability to distinguish challenging classes while minimizing false positives, which is critical for early screening of AD.



(a) Teacher ablation

(b) Student ablation

Fig. 2: Ablation study of AGMD components.

3.2 Ablation Study

Fig. 2 validates the necessity of each component. From the ablation experiment of the teacher model (Fig. 2a), its classification result (blue) decreased significantly without the **CMT** module, demonstrating that cross-modal fusion can effectively combine the advantages of two modalities. We evaluated the impact of the graph learning module, showing that the teacher model's classification performance (orange) decreased slightly when removing the anatomical graph. This suggests that only the semantic graph may not effectively learn brain connectivity; our AGMD, however, achieves the best performance (yellow) since it can capture key regional relationships. For the student model (Fig. 2b), the main observations are as follows: 1) our AGMD (green) achieves the best performance with complete distillation modules, suggesting it effectively learns disease features; 2) Removing graph distillation (orange) causes a 9.7% ACC drop (0.6853 vs. 0.7593), demonstrating that it cannot effectively learn the complex

AD-pathology distribution pattern without regional connectivity modeling. 3) Disabling the **Uncertainty-Aware Gating (UAG)** module (blue) reduces the overall performance, explaining the necessity to refine the teacher's knowledge in the distillation process. 4) Only using feature and logit distillation (yellow) obtains the worst classification results, suggesting that graph distillation can transfer complex knowledge about brain connections.

4 Conclusion

This paper proposes a novel **A**natomical **G**raph-based **M**ultilevel **D**istillation (AGMD) framework for diagnosing AD with missing modalities. AGMD achieves **SOTA** results on AD/pMCI/sMCI/NC classification task, with an accuracy of 0.7593 and a specificity of 0.9014, by integrating brain network topology modeling and uncertainty-aware knowledge transfer. Extensive studies confirm the necessity of multilevel and graph distillation with anatomical constraints, which provides complete and rational pathology distribution information. The proposed framework provides a solution for cross-modal knowledge transfer in disease diagnosis with missing modality, with potential applications in other medical vision tasks requiring robust unimodal inference. In the future, we plan to explore the relationship between brain connectivity patterns and disease progression, thus contributing to a deeper understanding of the development of Alzheimer's disease.

Acknowledgments. Data used in the preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

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

References

1. Bai, T., Zhao, J., Wen, B.: Guided adversarial contrastive distillation for robust students. IEEE Transactions on Information Forensics and Security (2023)
2. Braak, H., Braak, E.: Neuropathological stageing of alzheimer-related changes. Acta neuropathologica **82**(4), 239–259 (1991)
3. Cai, H., Gao, Y., Liu, M.: Graph transformer geometric learning of brain networks using multimodal mr images for brain age estimation. IEEE Transactions on Medical Imaging **42**(2), 456–466 (2022)
4. Cai, H., Gao, Y., Liu, M.: Graph transformer geometric learning of brain networks using multimodal mr images for brain age estimation. IEEE Transactions on Medical Imaging **42**(2), 456–466 (2023)
5. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5008–5017 (2021)

6. Feng, Y., Chen, W., Gu, X., Xu, X., Zhang, M.: Multi-modal semi-supervised evidential recycle framework for alzheimer's disease classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 130–140. Springer (2023)
7. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
8. Gustavsson, A., Norton, N., Fast, T., Frölich, L., Georges, J., Holzapfel, D., Kirabali, T., Krolak-Salmon, P., Rossini, P.M., Ferretti, M.T., et al.: Global estimates on the number of persons across the alzheimer's disease continuum. Alzheimer's & Dementia **19**(2), 658–670 (2023)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
10. Hu, Y., Huang, Y., Zhang, K.: Multi-scale information distillation network for efficient image super-resolution. Knowledge-Based Systems **275**, 110718 (2023)
11. Luo, X., Wu, J., Yang, J., Chen, H., Li, Z., Peng, H., Zhou, C.: Knowledge distillation guided interpretable brain subgraph neural networks for brain disorder exploration. IEEE Transactions on Neural Networks and Learning Systems **36**(2), 3559–3572 (2025)
12. Pham, C., Nguyen, V.A., Le, T., Phung, D., Carneiro, G., Do, T.T.: Frequency attention for knowledge distillation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2277–2286 (2024)
13. Qi, L., Kuen, J., Gu, J., Lin, Z., Wang, Y., Chen, Y., Li, Y., Jia, J.: Multi-scale aligned distillation for low-resolution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14443–14453 (2021)
14. Qin, D., Bu, J.J., Liu, Z., Shen, X., Zhou, S., Gu, J.J., Wang, Z.H., Wu, L., Dai, H.F.: Efficient medical image segmentation based on knowledge distillation. IEEE Transactions on Medical Imaging **40**(12), 3820–3831 (2021)
15. Qiu, S., Miller, M.I., Joshi, P.S., Lee, J.C., Xue, C., Ni, Y., Wang, Y., De Anda-Duran, I., Hwang, P.H., Cramer, J.A., et al.: Multimodal deep learning for alzheimer's disease dementia assessment. Nature communications **13**(1), 3404 (2022)
16. Qiu, Z., Yang, P., Xiao, C., Wang, S., Xiao, X., Qin, J., Liu, C.M., Wang, T., Lei, B.: 3d multimodal fusion network with disease-induced joint learning for early alzheimer's disease diagnosis. IEEE Transactions on Medical Imaging **43**(9), 3161–3175 (2024)
17. Ross, E.L., Weinberg, M.S., Arnold, S.E.: Cost-effectiveness of aducanumab and donanemab for early alzheimer disease in the us. JAMA neurology **79**(5), 478–487 (2022)
18. 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)
19. Vogel, J.W., Iturria-Medina, Y., Strandberg, O.T., Smith, R., Levitis, E., Evans, A.C., Hansson, O.: Spread of pathological tau proteins through communicating neurons in human alzheimer's disease. Nature communications **11**(1), 2612 (2020)
20. Wang, T., Zhu, Y., Zhao, C., Zhao, X., Wang, J., Tang, M.: Attention-guided knowledge distillation for efficient single-stage detector. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2021)
21. Xu, Q., Chen, Z., Ragab, M., Wang, C., Wu, M., Li, X.: Contrastive adversarial knowledge distillation for deep model compression in time-series regression tasks. Neurocomputing **485**, 242–251 (2022)

22. Yiannopoulou, K.G., Papageorgiou, S.G.: Current and future treatments in alzheimer disease: an update. Journal of central nervous system disease **12**, 1179573520907397 (2020)
23. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (2017)
24. Zhai, Z., Liang, J., Cheng, B., Zhao, L., Qian, J.: Strengthening attention: knowledge distillation via cross-layer feature fusion for image classification. International Journal of Multimedia Information Retrieval **13**(2), 23 (2024)
25. Zhang, L., Ma, K.: Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In: International Conference on Learning Representations (2021)
26. Zhou, H., He, L., Chen, B.Y., Shen, L., Zhang, Y.: Multi-modal diagnosis of alzheimer's disease using interpretable graph convolutional networks. IEEE Transactions on Medical Imaging (2024)