UniCross: Balanced Multimodal Learning for Alzheimer's Disease Diagnosis by Uni-modal Separation and Metadata-guided Cross-modal Interaction

Lisong Yin[1*], Chuyang Ye[2*], Tiantian Liu[1**], Jinglong Wu[1], and Tianyi Yan[1**]

[1] School of Medical Technology, Beijing Institute of Technology, Beijing, China
tiantian2bit@bit.edu.cn
yantianyi@bit.edu.cn
[2] School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China

Abstract. Early and accurate diagnosis of Alzheimer's disease (AD) is crucial for effective treatment and patient care. In clinical practice, physicians can achieve precise diagnoses through the integration of multimodal image information, and it is desired to develop automated diagnosis approaches based on the multimodal information. However, existing multimodal deep learning methods face a critical paradox: although models excel at leveraging joint features to improve task performance, they often neglect the optimization of independent representation capabilities for uni-modal. This shortcoming, known as **Modality Laziness**, stems from imbalanced modality contributions within conventional joint training frameworks, where models predominantly rely on dominant modalities and neglect to learn weaker ones. To address this challenge, we propose UniCross, a novel balanced multimodal learning paradigm. Specifically, UniCross employs separate learning pathways with specialized training objectives for each modality to ensure comprehensive uni-modal feature learning. In addition, we design a Metadata Weighted Contrastive Loss (MWCL) to facilitate effective cross-modal information interaction. The MWCL leverages patient metadata (e.g., age, gender, and years of education) to adaptively calibrate both cross-modal and intra-modal feature distances between individuals. We validated our approach through extensive experiments on the ADNI dataset, using structural MRI and FDG-PET modalities for AD diagnosis and mild cognitive impairment (MCI) conversion prediction tasks. The results demonstrate that UniCross not only achieves state-of-the-art overall performance, but also significantly improves the diagnosis performance when only a single modality is available. Our code is available at `https://github.com/Alita-song/UniCross`.

* L. Yin and C. Ye: Co-first authors, equal contribution.
** Corresponding authors: tiantian2bit@bit.edu.cn, yantianyi@bit.edu.cn

Keywords: Alzheimer's Disease · Balanced Multimodal Learning · Contrastive Learning.

1 Introduction

Alzheimer's disease (AD), the predominant cause of dementia, is a neurodegenerative disorder characterized by progressive cognitive decline [2]. It poses severe challenges to healthcare systems worldwide, particularly in developing countries and regions. Given the current lack of effective clinical treatments, early diagnosis and intervention for AD have become increasingly crucial [18]. Mild Cognitive Impairment (MCI) is a potential prodrome of AD, and it can be categorized as progressive MCI (pMCI) and stable MCI (sMCI) based on whether it progresses to AD within 36 months [21]. Identifying individuals with pMCI is essential for early intervention and treatment planning, as these patients are at a higher risk of converting to AD compared to sMCI patients, who maintain relatively stable cognitive function.

Given the complexity of pathological mechanisms and heterogeneity of clinical manifestations in Alzheimer's disease, recent research has increasingly focused on multimodal diagnostic approaches [27]. Among the various modalities, structural magnetic resonance imaging (sMRI) and positron emission tomography (PET) are the most commonly used [24]. sMRI provides detailed anatomical information to assess structural brain changes, while PET detects early functional changes by measuring cerebral glucose metabolism. Some studies have achieved impressive performance improvements by fusing information from these two modalities [22,17], which is reasonable since signals from different modalities often provide complementary information. However, recent studies [10,28] have shown that while benefiting from cross-modal interactions, these methods fail to adequately learn uni-modal features, potentially leading to suboptimal overall performance. This phenomenon is referred to as Modality Laziness [3]. Specifically, during multimodal representation learning, models tend to rely heavily on dominant modalities while neglecting weaker ones, resulting in insufficient learning of uni-modal features.

Several works have attempted to address the modality laziness problem by balanced training strategies [30] or gradient modulation methods [19,5]. However, these methods still follow the widely used joint training framework, which sets uniform learning objectives for all modalities, inherently leading to insufficient uni-modal feature learning. To explore possible solutions beyond the conventional joint training framework, some recent approaches [31,9] have reformulated it into a novel multimodal alternating learning paradigm. However, this reformulation sacrifices effective cross-modal information interaction.

Inspired by recent advances [4,31] in balanced multimodal representation learning, we propose UniCross, a novel multimodal learning paradigm that ensures sufficient uni-modal learning while maintaining effective cross-modal interactions for early diagnosis of AD and prediction of MCI conversion. Specifically, to tackle modality laziness, we design separate learning pathways with

specialized training objectives for each modality. Moreover, as naive separate training can lead to difficult feature fusion and lack of cross-modal interaction due to the heterogeneity across modalities [3], we introduce a shared head and a multimodal contrastive loss to facilitate effective cross-modal information interaction. In particular, considering that the symptoms of Alzheimer's disease may present demographic differences, we design a Metadata Weighted Contrastive Loss (MWCL) that leverages metadata (age, gender, and years of education) to adaptively calibrate both intra-modal and cross-modal feature distances in the representation space. Finally, we freeze the parameters of the modality-specific encoder and retrain a concatenation-based feature fusion module to obtain final predictions. Extensive experiments on the ADNI dataset [11] demonstrate that our method not only achieves state-of-the-art performance in both AD diagnosis and MCI conversion prediction tasks, but also improves the performance when only a single modality is available.

2 Methodology

2.1 Overall Design of UniCross

The overall design of UniCross is shown in Fig. 1. It adopts a two-stage training strategy to achieve balanced multimodal learning. The encoder training stage focuses on learning comprehensive uni-modal representations while maintaining effective cross-modal interactions, and the fine-tuning stage performs multimodal fusion for final prediction. We use a 3D patch embedding module that employs one 3D convolution with a kernel size and stride equal to the patch size to partition the input image into non-overlapping patches and convert them into embedding vectors for subsequent processing. To effectively capture long-range global dependencies in high-dimensional 3D medical images, we employ Vision Transformer (ViT-B) as our modality-specific encoders [16]. For the metadata, we encode it and then use a linear transformation as the meta encoder. During fine-tuning, we employ a simple yet effective concatenation-based fusion module to aggregate features from both image modalities.

2.2 Encoder Training

Uni-modal separation For a given dataset, there are three modalities $\mathcal{M} = \{\text{sMRI}, \text{PET}, \text{metadata}\}$. The i-th sample in a training batch of N samples can be represented as $\{\mathbf{s}_i, \mathbf{p}_i, \mathbf{c}_i, y_i\}$, where $\mathbf{s}_i$, $\mathbf{p}_i$, and $\mathbf{c}_i$ are the sMRI image, PET image, and metadata, respectively, and $y_i \in \{0, 1\}$ is the binary disease label.

To alleviate modality laziness, unlike existing methods, we do not perform multimodal fusion during the encoder training stage. Instead, we propose separate learning pathways with specialized training objectives for each modality. After patch embedding, $\mathbf{s}_i$ and $\mathbf{p}_i$ are fed into modality-specific encoders E_s and E_p to obtain features f_i^s and f_i^p respectively. For each modality, we employ separate classifiers ϕ_s and ϕ_p to predict the labels. The uni-modal training objective

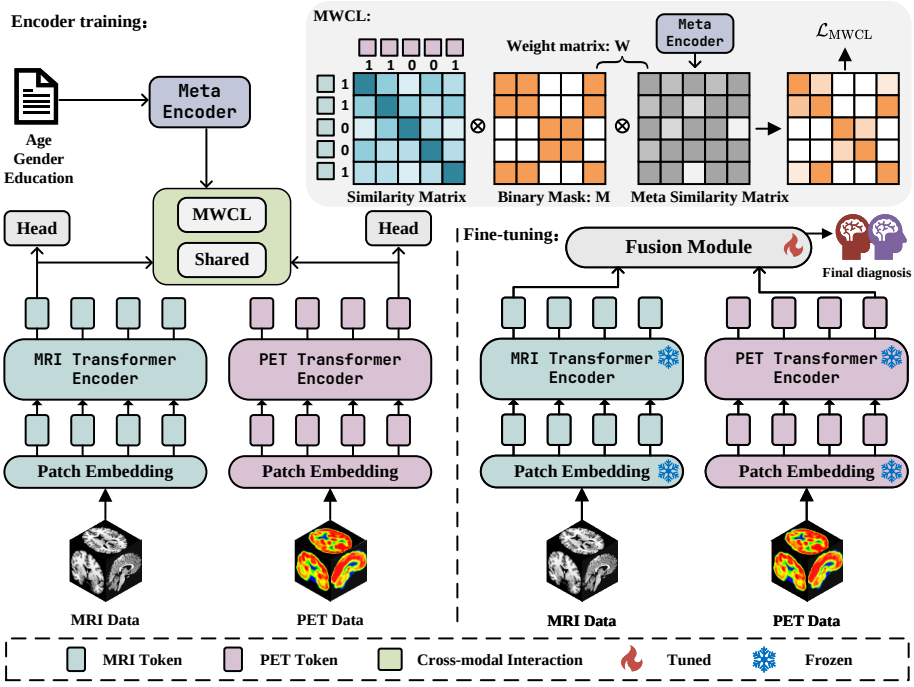

Fig. 1. The proposed UniCross framework, including the encoder training stage and fine-tuning stage. The top right corner shows the specific implementation of the MWCL. Note that we only show the cross-modal MWCL.

is defined as:

$$\mathcal{L}_{\text{uni}} = \sum_{i=1}^{N} [\mathcal{H}(y_i, \phi_{\text{s}}(f_i^{\text{s}})) + \mathcal{H}(y_i, \phi_{\text{p}}(f_i^{\text{p}}))], \tag{1}$$

where $\mathcal{H}$ is the softmax loss.

Moreover, due to the heterogeneity between modalities, independent training processes may lead to discrepancies in the representation spaces. To bridge this gap, we introduce a shared head (classifier) ϕ_{sp} that processes features from both modalities with the following training objective:

$$\mathcal{L}_{\text{sp}} = \sum_{i=1}^{N} [\mathcal{H}(y_i, \phi_{\text{sp}}(f_i^{\text{s}})) + \mathcal{H}(y_i, \phi_{\text{sp}}(f_i^{\text{p}}))]. \tag{2}$$

MWCL To further facilitate effective cross-modal interaction, we propose the MWCL, which is a contrastive loss that aligns sMRI and PET images in the feature space with the calibration of patient metadata. First, since samples from the same category are expected to be semantically similar, we extend the conventional multimodal contrastive loss to supervised contrastive learning, leveraging

class labels to define positive and negative sample pairs. Moreover, considering that AD symptoms are inherently correlated with demographic characteristics, even samples within the same category may exhibit varying distances in the feature space. Thus, we further incorporate metadata about demographic characteristics, including age, gender, and years of education in the MWCL. We assume that patients with similar demographic characteristics tend to exhibit similar pathological patterns and should therefore maintain closer proximity in the feature space. Based on the above motivations, the MWCL is computed as

$$\mathcal{L}_{\mathrm{MWCL}} = -\frac{1}{4N} \sum_{v,v' \in \{\mathrm{s,p}\}} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \log \frac{\exp(f_i^v \cdot f_j^{v'} / \tau)}{\sum_{k=1}^{N} \exp(f_i^v \cdot f_k^{v'} / \tau)} \tag{3}$$

where v and v' represent modality types, τ is a temperature parameter, and w_{ij} represents the normalized adaptive weight determined by disease labels and metadata for the sample pair (i, j). The weight is computed as

$$w_{ij} = \frac{\tilde{w}_{ij}}{\sum_{j=1}^{N} \tilde{w}_{ij}}, \tag{4}$$

where

$$\tilde{w}_{ij} = m_{ij} \cdot \psi(f_i^c, f_j^c) \quad \text{with} \quad m_{ij} = \begin{cases} 1, & \text{if } y_i = y_j \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

Here, m_{ij} is a binary mask that equals 1 if samples i and j belong to the same class, and 0 otherwise; f_i^{c} and f_j^{c} are meta-features obtained from the meta encoder E_{c}; $\psi(\cdot, \cdot)$ is a cosine similarity function that measures the relevance between samples based on meta-features. Note that Eq. (3) performs contrastive learning not only between sMRI and PET images but also within the single modality of sMRI or PET images (when v and v' are identically s or p). This design further enhances the representation of each modality.

The MWCL adaptively calibrates both intra-modal and cross-modal distances in the feature space. By incorporating meta-information into the contrastive learning framework, the MWCL encourages the model to learn representations capable of distinguishing between meta-related and disease-related brain changes.

The final training objective $\mathcal{L}$ of UniCross combines $\mathcal{L}_{\mathrm{uni}}$, $\mathcal{L}_{\mathrm{sp}}$, and $\mathcal{L}_{\mathrm{MWCL}}$:

$$\mathcal{L} = \mathcal{L}_{\mathrm{uni}} + \mathcal{L}_{\mathrm{sp}} + \mathcal{L}_{\mathrm{MWCL}} \tag{6}$$

2.3 Fine-tuning

During fine-tuning, we freeze all parameters of the trained encoders to preserve the learned features. Then, the outputs of the encoders are fused by concatenation and the fused result is fed into a fully connected layer for final prediction.

2.4 Implementation Details

In the encoder training stage, following [22], we use a Cosine Annealing with Warm Restarts strategy, where the initial period was set to 10, the period multiplication factor was set to 3, and the minimum learning rate was set to 1×10^{-5}. The model is trained for 40 epochs with a batch size of 8 and a contrastive learning temperature parameter τ of 0.07. In the fine-tuning stage, we freeze the encoders and apply the Adam optimizer with 10 training epochs and a learning rate of 0.001.

3 Experiments and Results

3.1 Data Description and Experimental Settings

Paired sMRI and PET images were collected from the ADNI [11] dataset. In total 1,044 subjects were selected in the experiment, including 284 AD patients, 385 normal controls (NC), 183 pMCI subjects, and 192 sMCI subjects. sMRI was preprocessed with FreeSurfer [6] for motion correction, intensity normalization, and skull stripping. sMRI and PET images were coregistered to the Colin27 [8] template with FSL FLIRT [12]. In addition, we applied extra smoothing to the PET images [22]. All images were resampled to $128 \times 128 \times 128$ with a resolution of $1 \times 1 \times 1$ mm^3. Data augmentation [26] was performed on the training set, including random 3D rotation, random zoom, and random shift. The metadata included the age, gender, and years of education, where categorical variables (gender) were converted with one-hot encoding and continuous variables (age and years of education) were standardized with z-score normalization.

All experiments were conducted with five-fold cross-validation. Note that for sMCI/pMCI classification, we followed [13] and used the model pre-trained on the AD early diagnosis task for initialization. Three metrics were used to evaluate the performance: accuracy (ACC), F1-score (F1), and area under the receiver operating characteristic curve (AUC).

3.2 Comparison with the State-of-the-art Methods

We compared our method with several state-of-the-art multimodal fusion approaches on both AD/CN classification and MCI conversion prediction tasks. The compared methods include early and middle fusion strategies: 1) MFNET [1], which uses a Multi-Fiber architecture and multiplexer modules to facilitate multimodal information interaction; 2) MDL-NET [22], which employs a multi-fusion joint learning module to improve global, local and latent feature representation; 3) SSFTT [25], which combines convolutional neural networks and transformers to capture multimodal features and high-level semantic features; 4) DiaMond [17], which utilizes self-attention, bi-attention, and a RegBN mechanism for effective multimodal fusion. Our method was also compared with recent balanced multimodal learning methods, including: 5) OGM-GE [19], which employs real-time gradient modulation to adaptively control the optimization process of each

Table 1. Comparison with State-of-the-art methods on AD/CN and sMCI/pMCI classification tasks.

Method	AD/CN			sMCI/pMCI		
	ACC	F1	AUC	ACC	F1	AUC
MFNet [1]	88.94±2.32	86.45±3.20	93.69±1.71	71.93±6.07	67.04±8.05	82.57±4.19
MDL-Net [22]	89.68±2.98	87.20±4.25	96.17±1.10	71.90±7.07	65.21±13.36	81.78±2.23
SSFTT [25]	92.97±0.76	91.72±0.81	96.12±1.01	78.87±1.38	76.24±2.82	85.05±2.61
DiaMond [17]	88.49±1.76	86.55±1.97	93.82±2.08	73.79±3.01	69.78±4.35	79.93±3.23
OGM-GE [19]	92.23±1.73	90.98±1.63	95.91±0.52	75.92±3.18	72.37±4.81	83.95±3.88
DI-MML [4]	90.58±1.45	88.51±2.11	95.79±0.84	78.86±6.10	75.54±8.54	82.70±5.69
UniCross	**93.57±1.39**	**92.30±1.24**	**97.04±1.24**	**79.67±3.36**	**78.84±4.09**	**85.22±3.08**

Table 2. An ablation study on the AD/CN classification task.

Method	ACC	F1	AUC
Shared (w/o MWCL)	92.53±1.88	90.99±2.16	96.56±1.25
MWCL (w/o shared head)	91.18±2.23	89.45±2.67	95.57±1.83
CLIP [23]+Shared	90.43±0.98	88.27±1.23	95.07±1.33
SupCon [14]+Shared	90.88±0.57	89.23±0.65	96.11±1.04
DeCUR [29]+Shared	92.38±1.72	91.02±1.96	96.50±1.33
UniCross	**93.57±1.39**	**92.30±1.24**	**97.04±1.24**

modality; 6) DI-MML [4], which also adopts separate training pathways while using dimension-decoupled unidirectional contrastive (DUC) loss for cross-modal information transfer.

The performance of each method is shown in Table 1. For AD/CN classification, our method achieves the best performance across all metrics, with an accuracy of 93.57%, F1-score of 92.30%, and AUC of 97.04%. The MCI conversion prediction task is generally more challenging due to the subtle differences between stable and progressive MCI. Even in this challenging scenario, our method demonstrates superior performance with a 79.67% accuracy, 78.84% F1-score, and 85.22% AUC. These results demonstrate that our UniCross framework not only excels in standard AD diagnosis but also shows promising capability in the more challenging task of early conversion prediction.

3.3 Ablation Study

Next, we explore the importance of the shared head and MWCL in UniCross with an ablation study on the AD/CN classification task. The results are shown in Table 2. First, we removed the shared head (replaced by separate heads) or MWCL, both which lead to lower classification performance. In addition, we compared the MWCL with other comparative losses, including: 1) CLIP [23] loss, which maximizes the similarity between paired cross-modal inputs while minimizing unpaired samples from different modalities; 2) SupConloss [14], which pulls together samples from the same class while pushing apart samples from different classes; 3) DeCUR [29], which uses multimodal redundancy reduction to learn

Table 3. Accuracy of linear probing on encoders for various multi-modal late-fusion methods and uni-modal training on the AD/CN classification task.

Method	sMRI	PET	Multi
Uni1	80.72±1.47	-	-
Uni2	-	90.14±1.72	-
Concat	78.77±1.84	88.64±1.08	91.63±1.37
Sum	78.48±0.52	90.73±1.32	90.73±1.86
Film [20]	82.37±2.50	84.76±2.41	86.85±1.20
Gated [15]	74.29±3.18	90.88±1.18	90.43±1.28
CrossAttention [7]	81.61±2.13	88.34±1.67	91.78±3.23
UniCross	**83.71±0.58**	**92.07±0.11**	**93.57±1.39**

common and unique representations of the modality. These contrastive losses were integrated with the proposed shared head, and their results are worse than that of UniCross, which further confirms the benefit of the proposed MWCL.

3.4 Effectiveness in Addressing Modality Laziness

Finally, following [31], to validate the effectiveness of UniCross in addressing modality laziness, we performed linear probing on encoders trained by different multimodal late fusion approaches for AD/CN classification. We compared UniCross with various late fusion methods, which can be categorized into two groups: (1) traditional multimodal fusion methods, including summation (Sum), concatenation (Concat), and multimodal CrossAttention [7]; (2) modulation-based fusion methods, including FiLM [20] and Gated [15].

The accuracy of each method is shown in Table 3, where the results of uni-modal baselines (Uni1 for sMRI and Uni2 for PET images) are given for reference. Concat, Sum, and Gated show degraded performance on sMRI (78.77%, 78.48%, and 74.29%, respectively), indicating the presence of modality laziness where the model relies heavily on the stronger modality (PET) while compromising the weaker one (sMRI). Although Film and CrossAttention facilitate more effective cross-modal interactions, they only slightly improve the sMRI performance while significantly degrading the PET performance (to 84.76% and 88.34%, respectively). In contrast, our UniCross framework not only achieves superior overall performance (93.57%) but also improves the performance for each individual modality, with an increase of 2.99% for sMRI and 1.93% for PET images compared to the uni-modal baselines. This demonstrates that our method effectively alleviates modality laziness.

4 Conclusion

We have proposed UniCross, a balanced multimodal learning framework that effectively addresses modality laziness. Through separate learning pathways with specialized training objectives, our approach ensures comprehensive uni-modal

feature learning while maintaining effective cross-modal interactions by proposing the MWCL. The MWCL leverages patient metadata to adaptively calibrate feature distances in the representation space. Extensive experiments on the ADNI dataset demonstrate that UniCross not only achieves state-of-the-art performance in both AD diagnosis and MCI conversion prediction tasks but also significantly improves the performance given a single modality.

Acknowledgments. T. Liu is supported by the National Natural Science Foundation of China (grant number 62406025). C. Ye is supported by the Beijing Municipal Natural Science Foundation (7242273). J. Wu is supported by the National Natural Science Foundation of China (grant number 62373056) and the Shenzhen Basic Research Program (JCYJ20241202124804007). T. Yan is supported by the National Natural Science Foundation of China (grant number 62336002), the STI 2030-Major Projects (grant number 2022ZD0208500), and the Beijing Nova Program (grant number 20230484465).

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

References

1. Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: Multi-fiber networks for video recognition. In: Proceedings of the european conference on computer vision (ECCV). pp. 352–367 (2018)
2. DeTure, M.A., Dickson, D.W.: The neuropathological diagnosis of alzheimer's disease. Molecular neurodegeneration **14**(1), 32 (2019)
3. Du, C., Teng, J., Li, T., Liu, Y., Yuan, T., Wang, Y., Yuan, Y., Zhao, H.: On uni-modal feature learning in supervised multi-modal learning. In: International Conference on Machine Learning. pp. 8632–8656. PMLR (2023)
4. Fan, Y., Xu, W., Wang, H., Liu, J., Guo, S.: Detached and interactive multimodal learning. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 5470–5478 (2024)
5. Fan, Y., Xu, W., Wang, H., Wang, J., Guo, S.: Pmr: Prototypical modal rebalance for multimodal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20029–20038 (2023)
6. Fischl, B.: Freesurfer. Neuroimage **62**(2), 774–781 (2012)
7. Golovanevsky, M., Eickhoff, C., Singh, R.: Multimodal attention-based deep learning for alzheimer's disease diagnosis. Journal of the American Medical Informatics Association **29**(12), 2014–2022 (2022)
8. Holmes, C.J., Hoge, R., Collins, L., Woods, R., Toga, A.W., Evans, A.C.: Enhancement of mr images using registration for signal averaging. Journal of computer assisted tomography **22**(2), 324–333 (1998)
9. Hua, C., Xu, Q., Bao, S., Yang, Z., Huang, Q.: Reconboost: Boosting can achieve modality reconcilement. arXiv preprint arXiv:2405.09321 (2024)
10. Huang, Y., Lin, J., Zhou, C., Yang, H., Huang, L.: Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In: International conference on machine learning. pp. 9226–9259. PMLR (2022)
11. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al.: The alzheimer's

disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
12. Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M.: Fsl. Neuroimage **62**(2), 782–790 (2012)
13. Kang, L., Gong, H., Wan, X., Li, H.: Visual-attribute prompt learning for progressive mild cognitive impairment prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 547–557. Springer (2023)
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
15. Kiela, D., Grave, E., Joulin, A., Mikolov, T.: Efficient large-scale multi-modal classification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
16. Kunanbayev, K., Shen, V., Kim, D.S.: Training vit with limited data for alzheimer's disease classification: An empirical study. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 334–343. Springer (2024)
17. Li, Y., Ghahremani, M., Wally, Y., Wachinger, C.: Diamond: Dementia diagnosis with multi-modal vision transformers using mri and pet. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 107–116 (2025). https://doi.org/10.1109/WACV61041.2025.00021
18. Ngandu, T., Lehtisalo, J., Solomon, A., Levälahti, E., Ahtiluoto, S., Antikainen, R., Bäckman, L., Hänninen, T., Jula, A., Laatikainen, T., et al.: A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (finger): a randomised controlled trial. The Lancet **385**(9984), 2255–2263 (2015)
19. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8238–8247 (2022)
20. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
21. Petersen, R.C., Negash, S.: Mild cognitive impairment: an overview. CNS spectrums **13**(1), 45–53 (2008)
22. 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). https://doi.org/10.1109/TMI.2024.3386937
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
24. Shanmugavadivel, K., Sathishkumar, V., Cho, J., Subramanian, M.: Advancements in computer-assisted diagnosis of alzheimer's disease: A comprehensive survey of neuroimaging methods and ai techniques for early detection. Ageing Research Reviews **91**, 102072 (2023)
25. Sun, L., Zhao, G., Zheng, Y., Wu, Z.: Spectral–spatial feature tokenization transformer for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2022)

26. Turrisi, R., Verri, A., Barla, A.: The effect of data augmentation and 3d-cnn depth on alzheimer's disease detection. arXiv preprint arXiv:2309.07192 (2023)
27. Walhovd, K., Fjell, A., Brewer, J., McEvoy, L., Fennema-Notestine, C., Hagler, D., Jennings, R., Karow, D., Dale, A., Initiative, A.D.N., et al.: Combining mr imaging, positron-emission tomography, and csf biomarkers in the diagnosis and prognosis of alzheimer disease. American Journal of Neuroradiology **31**(2), 347–354 (2010)
28. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12695–12705 (2020)
29. Wang, Y., Albrecht, C.M., Braham, N.A.A., Liu, C., Xiong, Z., Zhu, X.X.: Decoupling common and unique representations for multimodal self-supervised learning. In: European Conference on Computer Vision. pp. 286–303. Springer (2024)
30. Yao, Y., Mihalcea, R.: Modality-specific learning rates for effective multimodal additive late-fusion. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 1824–1834 (2022)
31. Zhang, X., Yoon, J., Bansal, M., Yao, H.: Multimodal representation learning by alternating unimodal adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27456–27466 (2024)