Unified Multi-Modal Learning for Any Modality Combinations in Alzheimer's Disease Diagnosis

Yidan Feng[1], Bingchen Gao[1], Sen Deng[1], Anqi Qiu[2], and Jing Qin[1](✉)

[1] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China
harry.qin@polyu.edu.hk
[2] Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, China

Abstract. Our method solves unified multi-modal learning in an diverse and imbalanced setting, which are the key features of medical modalities compared to the extensively-studied ones. Different from existing works that assumed fixed or maximum number of modalities for multi-modal learning, our model not only manages any missing scenarios but is also capable of handling new modalities and unseen combinations. We argue that, the key towards this any combination model is the proper design of alignment, which should guarantee both modality invariance across diverse inputs and effective modeling of complementarities within the unified metric space. Instead of exact cross-modal alignment, we propose to decouple these two functions into representation-level and task-level alignment, which we empirically show are both indispensable in this task. Moreover, we introduce tunable modality-agnostic Transformer to unify the representation learning process, which significantly reduces modality-specific parameters and enhances the scalability of our model. The experiments have shown that the proposed method enables one single model handling all possible combinations of the six seen modalities and two new modalities in Alzheimer's Disease diagnosis, with superior performance on longer combinations.

Keywords: Multi-Modal Learning · Alzheimer's Disease · Missing Modality

1 Introduction

Advancements in diagnostic technologies have empowered the integration of diverse yet complementary data to better support clinical decisions particularly in complex diseases. This trend has sparked a growing interest in multi-modal deep learning for objective and quantitative computer-aided diagnosis [1]. Compared to the extensive studies in modeling language, vision, and audio [9], medical multi-modal learning for diagnosis possesses unique characteristics: it involves a more diverse set of modalities derived from various biosensors and laboratory tests. Unlike previous assumptions [21,12,2,20], these modalities are not exactly aligned in semantics, and exhibit varying missing patterns in practical scenarios.

The diagnosis of Alzheimer's Disease (AD) is a typical task involving diverse and imbalanced medical modalities. Various clinical rating scales are employed to assess observable behavioral and cognitive symptoms, complemented by neuroimaging techniques to provide a more comprehensive understanding of the structural and functional changes in the brain. Existing studies have investigated multi-modal learning on specific modality combinations for AD diagnosis, including T1-MRI & FDG-PET [5,14], T1-MRI & specific tabular data [15,11,8], Amyloid-PET, T1-MRI & FDG-PET [22,23], etc. These works assume a limited and fixed set of modalities, requiring training and inference on completely matched multi-modal data. However, this assumption will not always hold in practice due to the imbalanced modality distribution. As the assumed number of modalities increases, the modality-complete set will contract, leaving insufficient data for training. Meanwhile, the overall set of possible combinations will expand significantly, and may not necessarily match the specific combination encountered during inference. Additionally, patients may possess AD-related data outside the assumed modalities. Therefore, we propose to enhance medical multi-modal learning with flexibility on processing **any modality combinations** (short for AnyMod), with the following desired properties: 1) fully leveraging the available training data (may or may not matched) from different modalities; 2) easy adaptation to new modalities; 3) capability of inference on any (even unseen) combinations.

To achieve this goal, the first challenge is the architectural design for accommodating any modality combinations. In recent literature, Transformer [16,7] has been increasingly applied for this purpose. Many studies [18,12] have suggested parallel processing of multi-modal tokens by sharing self-attention layers, but computing attention of increasing modalities leads to quadratically growing cost. An alternative approach involves serial computation of different modalities through cross-attention [10], which is however, not permutation-invariant to input modalities. Moreover, these approaches necessitate modality-specific backbone models [12], FFNs [18], or separate Transformers [10], whose parameter sizes constitute a significant proportion within the overall model. This diminishes the scaling capability of their models and is prone to training instability or overfitting, especially with limited data in the medical context.

Another critical challenge is modality alignment. In multi-modal learning, the role of alignment is two-fold: 1) modeling the relationships among different modalities in a common metric space, thereby assisting the fusion module in learning inter-modal interactions [12,18]; 2) enhancing modality invariance for robustness against varying inputs (missing modalities) [24,17,19]. Cross-modal alignment, which encourages features from co-occurring modalities to exactly match each other, is widely used for both purposes [12,18,19]. In essence, this alignment method confines the modeling of relationships among diverse modalities to a single mode (see Fig. 2), thus simultaneously fulfilling the two functions of alignment. However, this approach is inadequate for addressing the specificities within the medical context. Unlike common modalities (i.e., vision, language, and audio), there lacks underlying semantics to connect medical modalities. Instead, the relationship among medical modalities is task-specific, often tied to a particular disease of interest. Therefore, directly enforcing cross-modal alignment will hinder the exploration of distinctive yet complementary information from diverse modalities for differential diagnosis.

To address these issues, we propose to decouple multi-modal alignment into representation-level and task-level: before fusion, the relation modeling of co-occurring modalities is expanded to N modes (see Fig. 2) in the unified metric space, which allows for exploration of complementary features from different modalities; after fusion, different samples and combinations within the same class are aligned to ensure robustness against varying input combinations. Our architecture is based on multi-modal Transformer, but with two improvements: 1) to model the projection from raw inputs to the unified metric space, we integrate tunable modality-agnostic Transformer to significantly reduce modality-specific parameters, which mitigates training instabilty and overfitting. Meanwhile, the trained projector can be easily adapted to new modalities with few modifications; 2) after projection, the multimodal feature tokens are clustered within the fixed number of task factors, ensuring stable computational costs for any length of combination. Experimental results have shown that, for this new setting of learning any modality combinations, our solution enables a single unified model to achieve growing advantages to models trained separately on each combination as the number of modalities increase. Moreover, the adapted model for new modalities can effortlessly handle unseen combinations without further training, which indicates the promising scaling capability of the proposed model.

2 Methods

2.1 Problem Statement

Our objective is to learn from the diverse combinations of modalities present in the training data to make predictions for Alzheimer's Disease classification on any combination of seen modalities while ensuring adaptability to unseen modalities with minimal computational cost. We denote a set of K modalities that will be seen in the training set as $M = \{m_1, m_2, ..., m_K\}$, and the set of unseen modalities as $M_u = \{m_{u1}, ...\}$. Then all the combinations of seen modalities can be expressed as the non-empty $2^M = \{X \mid X \subseteq M, X \neq \emptyset\}$, and the seen combinations in the training process can be expressed as $C = \left\{X_i \mid X_i \in 2^M, \cup X_i = M\right\}$. We consider two types of unseen scenarios at inference: 1) unseen combination of seen modalities $C_u = 2^M \backslash C$, which can be directly inferred after the main training process; 2) unseen combination involving unseen modalities $C'_u = \{X \mid X \subseteq M \cup M_u, X \cap M_u \neq \emptyset\}$. We refer to $c \in C_*$ as a 'combination', and when $|c| = 1$ it is considered a special case of the combination, equivalent to a single modality.

We define a function f that could accept the sample of any modality combination, which can be written as $f\left(\mathcal{X}_c; \theta, \cup\theta_{m_i}\right), m_i \in c$, where $\mathcal{X}_c = \{x_{m_i}\}$ is the sample of combination $c \in C_*$. The function is modelled using a deep neural network with learnable parameters θ and $\cup\theta_{m_i}$, where θ denotes the common

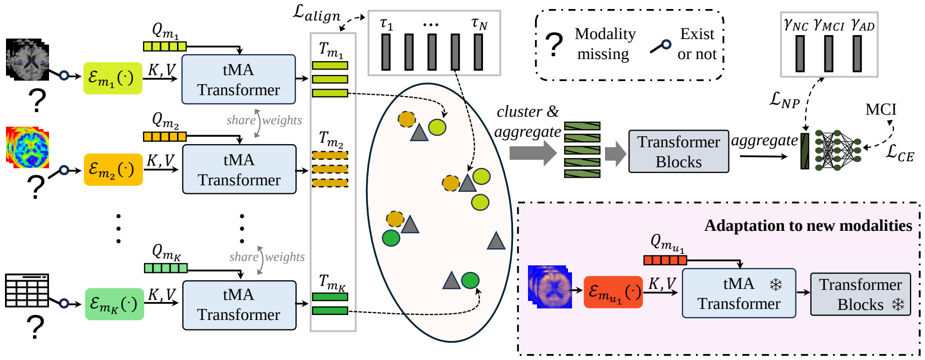

Fig. 1. Pipeline of the proposed method, which consists of projection to map raw data to the common metric space, and fusion to aggregate features for the final task.

parameters shared across all cases, and θ_{m_i} denotes the modality-specific parameters. Through our architectural design, the number of modality-specific parameters is significantly smaller than the modality-agnostic processing module, that is, $|\theta_{m_i}| \ll |\theta|$. After training on C, the network can process test samples $X_{c'}, c' \in 2^M$. For samples that involves unseen modalities, it requires additional training only on samples of single unseen modalities $m'_i \in M_u$ to obtain $\theta_{m'_i}$, which then can be directly applied to infer on any unseen combinations on C'_u.

2.2 Architectural Design

The AnyMod architecture is designed with two Transformer-based modules dedicated to projection and fusion (see Fig. 1). The size of θ_{m_i} is significantly reduced by introducing modality-specific processing solely for 1) initial feature consolidation and 2) distance modeling in the projection phase. For 1), we employ different embedding layers $\mathcal{E}_{m_i}$ tailored for different format of modalities. In handling 3D volumes, we utilize the embedding layer to compress redundant information through only two 3D ResNet blocks. The resulting 3D features are directly flattened into initial tokens. Conversely, for tabular data, the embedding layer serves to expand each attribute value to the feature dimension. We follow [4] to use linear layer for continuous values and look-up table for categorical values. For 2), we introduce tunable modality-agnostic Transformer, denoted as $\mathcal{G}(E_{m_i}, Q_{m_i}; \theta_G)$, derived from [7]. The architecture of $\mathcal{G}$ is composed of alternating cross-attention layers and Transformer blocks. Each cross-attention layer incorporates a modality-specific query Q_{m_i} that is trainable, enabling tailored modification of distance modeling for diverse modalities. It's noteworthy that, aside from Q_{m_i}, the parameters θ_G are shared uniformly across all modalities. The tunable modality-agnostic Transformer is independently applied to each modality, producing a set of multi-modal feature tokens $t^i_j \in T_{m_i}$, which denotes the j^{th} feature of modality m_i. After projection, the multi-modal feature tokens

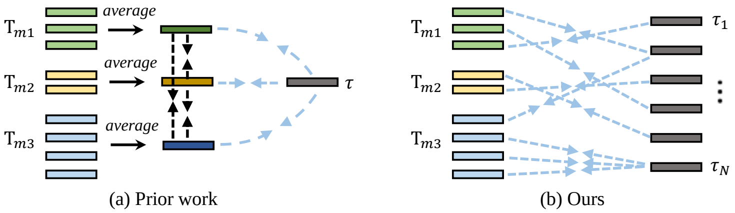

Fig. 2. Expanding the modes in multi-modal representation learning.

are distilled into fixed length N, and then processed by a fusion transformer to generate class embeddings for the final classification. This framework is not constrained by the maximum assumed number of modalities; both the Transformers for projection and fusion process all modalities in a unified manner.

2.3 Task-Oriented Fusion

Our approach to unified multi-modal fusion emphasizes the invariant aspects on the task side. First, we assume there exist a certain set of N task-related factors, and set implicit anchors $\{\tau_1, \tau_2, ..., \tau_N\}$ for these factors. Then our model learns to align each feature token t_j^i to one of the implicit anchor by

$$\mathcal{L}_{\text{align}} = -\sum_{ij} \log \left(\frac{\max \left\{ e^{\left(t_j^i\right)^\top \tau_1}, \cdots, e^{\left(t_j^i\right)^\top \tau_N} \right\}}{\sum_{n=1}^{N} e^{\left(t_j^i\right)^\top \tau_n}} \right). \tag{1}$$

Fig. 2 illustrates the difference of our alignment method from previous approaches [21,12,2]. For aligning co-occurring modalities, it's common practice to maximize the similarity of features across pairs of modalities (black arrows), which is equivalent to using only one anchor (blue arrows). However, this approach encounters two issues: 1) the feature tokens are directly compressed to obtain a single feature for each modality, overlooking potential differences among intra-modal features; 2) directly minimizing the inter-modal distances between different modalities during the feature extraction stage contradicts the goal of preserving complementary features among different modalities. Differently, our method introduces multiple implicit task anchors, and thus allows modelling of both intra- and inter-modal similarities and differences.

For fusion, each feature token is then clustered to a task-related factor $\upsilon_j^i = \underset{n}{\operatorname{argmax}}((t_j^i)^\top \tau_n)$, and the features in each cluster n are then aggregated by

$$t_n' = \sum_{\upsilon_i^j = n}^{i,j} \omega_i^j t_i^j, \omega_i^j = \frac{e^{\left(t_j^i\right)^\top \tau_n}}{\sum_{\upsilon_{j'}^{i'}=n}^{i',j'} e^{\left(t_{j'}^{i'}\right)^\top \tau_n}}. \tag{2}$$

Thus, the set of aggregated features is constrained to a fixed length $|\{t_n'\}| = N$, and uniformly processed by the transformer blocks. The fusion transformer's outputs are subsequently averaged to yield the class embedding μ_y, where y denotes the class label of the input sample. For task-level alignment, we establish a set of explicit task anchors γ_i for each class i and employ the N-pair Loss [13]. This loss function drives the embedding towards the respective class token while pulling away from tokens belonging to other classes:

$$\mathcal{L}_{\mathrm{NP}}\left(\mu_y, \gamma_y, \{\gamma_i\}_{i \neq y}\right) = \log\left[1 + \sum_{i \neq y} e^{\left((\mu_y)^\top \gamma_i - (\mu_y)^\top \gamma_y\right)}\right] \tag{3}$$

For samples of different combinations of the same class y, their embeddings are brought closer together through the intermediate class anchor γ_y. This approach eliminates the necessity to sample various combinations for calculating feature distances in each training step, contributing to enhanced training stability.

3 Experiments

3.1 Data Processing and Implementation Details

Following [11], our task is set as 3-way classification of Normal Cognition (NC), Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD). We extracted eight modalities: T1w MRI (T1), T2w MRI (T2), FDG-PET (F), Amyloid-PET (A), MMSE (Mm), MoCA (Mo), NeuroBat (Ne) and NPI-Q (Np), from Alzheimer's Disease Neuroimaging Initiative (ADNI) database [6]. The set of unseen modalities is set as $M_u = \{\text{A,Ne}\}$. We utilize all available time points from different modalities, and they are matched to form a sample of multi-modal combination if their examination time are within a 6-month window. The training and testing data are split in an 8:2 ratio for each combination at patient-level, ensuring there is no data leakage. We use cross-entropy loss $\mathcal{L}_{CE}$ for classification, and the overall loss function is a weighted combination of $\mathcal{L}_{CE}$, $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{NP}}$. For optimization, AdamW was applied with an initial learning rate of $3e-4$. Cosine Annealing scheme with linear warm-up is adopted for scheduling. In our model, the task anchors are all learnable parameters of $\mathbb{R}^{128}$. For evaluation, we use weighted F1-score and Accuracy (ACC). To update the model for new modalities, all the available data in the training set is leveraged for supervised training.

3.2 Ablation Studies

Experiments for ablation studies were conducted on the set {T1, F, Mo, Np}. Specific module is evaluated by removing it from the complete model.

Ablation of Architectural Design. The proposed architectural design involves 1) tunable modality-agnostic (MA) Transformer for sharing main projection parameters, and 2) clustering to fixed length in the fusion module. For 1), we first prohibit the tunable property of MA transformer by sharing Q_{m_i}, and then remove all parameter sharing by using pure modality-specific backbone

Table 1. Ablation of the projection architecture. MSP for modality-specific projection.

	T1		F		T1&F		T1&F&Mo		T1&Mo&F&Np		Mean		PARAMS
	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	
MSP	0.296	0.408	**0.633**	**0.622**	**0.660**	**0.661**	**0.687**	**0.686**	$\underline{0.732}$	**0.735**	$\underline{0.589}$	$\underline{0.585}$	74.83 M
share Q_{m_i}	$\underline{0.465}$	$\underline{0.461}$	0.433	0.518	0.434	0.525	$\underline{0.680}$	0.674	0.727	0.725	0.588	0.582	11.08 M
Ours	**0.512**	**0.515**	$\underline{0.603}$	$\underline{0.595}$	$\underline{0.606}$	$\underline{0.618}$	0.675	$\underline{0.677}$	**0.734**	$\underline{0.732}$	**0.595**	**0.590**	11.09 M

Table 2. Ablation of the fusion module. c for clustering, $\mathcal{L}_a$ for $\mathcal{L}_{align}$. 'Mean' represents the average performance of all tested modalities.

		T1&F		T1&Mo		F&Mo		T1&F&Mo		Mo&F&Np		T1&Mo&F&Np		Mean	
		F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
w/ $\mathcal{L}_a$	w/ c	$\underline{0.606}$	**0.618**	**0.657**	**0.654**	**0.680**	$\underline{0.674}$	**0.675**	**0.677**	$\underline{0.722}$	**0.721**	**0.734**	**0.732**	**0.595**	**0.590**
	w/o c	**0.612**	$\underline{0.610}$	0.609	0.606	0.676	0.668	**0.675**	$\underline{0.671}$	**0.723**	$\underline{0.718}$	$\underline{0.732}$	$\underline{0.728}$	0.590	0.585
w/o $\mathcal{L}_a$	w/ c	0.600	0.592	0.620	0.615	0.657	0.654	0.653	0.635	0.674	0.671	0.661	0.660	0.582	0.576
	w/o c	0.581	0.579	0.606	0.598	0.678	**0.694**	0.631	0.631	0.671	0.665	0.669	0.666	0.586	0.579
w/o $\mathcal{L}_{NP}$		0.578	0.577	$\underline{0.623}$	$\underline{0.628}$	$\underline{0.679}$	0.668	0.631	0.635	0.720	0.712	0.665	0.666	**0.595**	$\underline{0.589}$

models for projection (MSP). As shown in Tab. 1, our original model achieved best average performance and close result to MSP in different combinations using only 15% parameters. MSP suffers from slow convergence (3 × ours), highly oscillating outcomes, and collapse on specific modalities (T1). For 2), the effect of clustering is demonstrated in Tab. 2, indicating clustering will reduce computational cost without sacrificing performance.

Decoupled Alignment and Modality Imbalance. The critical challenge brought by training various combinations of diverse and heterogeneous modalities is the exacerbated modality imbalance, where different modalities converge and overfit at different rates [3]. Our experiments have revealed that the performance of learning any combinations is closely related to early overfitting, and both the proposed $\mathcal{L}_{align}$ and $\mathcal{L}_{NP}$ are indispensable for preventing this early overfitting (See Fig. 3 (a)). As shown in Tab. 2, representation-level alignment has brought significant improvement especially on long combinations. However, $\mathcal{L}_{align}$ cannot maintain effectiveness without task-level alignment $\mathcal{L}_{NP}$. As indicated in Fig. 3 (a), $\mathcal{L}_{NP}$ benefits robustness against varying combinations, without it, the validation losses are more dispersed, while only with $\mathcal{L}_{NP}$ will narrow the performance gap but at an overfitted point.

3.3 Comparative Studies

Since we proposed a new setting, there are no existing work for direct comparison. Instead, we compare with flexible architectures for solving missing modalities in general multi-modal learning (Everything [12]) and AD classification (CasAD [10]), where Everything is a typical cross-modal alignment method based on parallel Transformer and CasAD uses cascaded Transformer for flexible fusion. For separate models, we adopted 3D ResNet and late fusion for 3D volumes, FT-Transformer [4] for tabular, and parallel Transformer for the fusion of tabular &

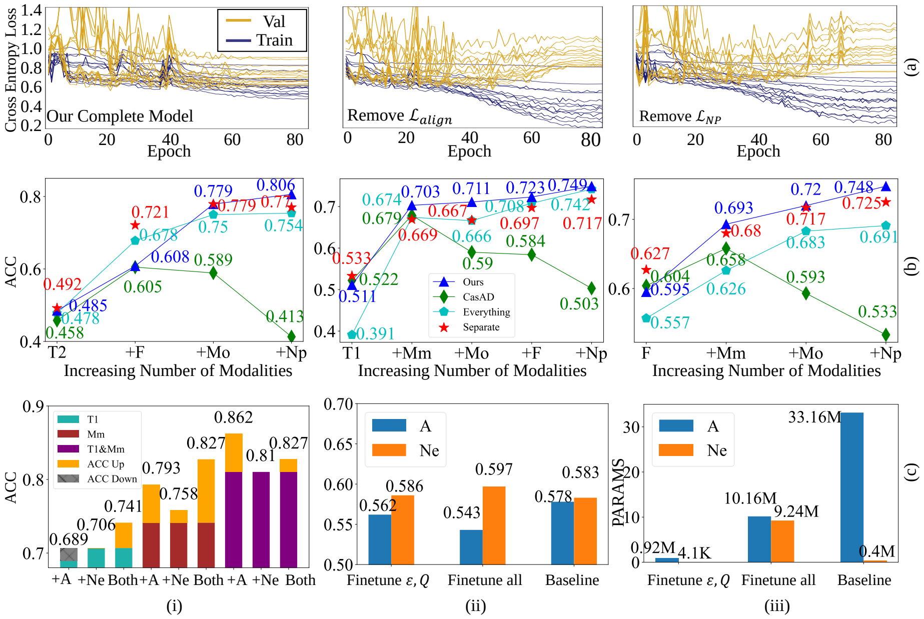

Fig. 3. (a) Loss curves with each line from a single combination; (b) results of comparative studies; (c) results on new modalities and unseen combinations.

3D volumes following [8]. The results are shown in Fig. 3 (b), our model shows growing advantages to separately trained models as the number of modalities increases, while Everything model is tending to flatten the top performances.

3.4 Adaptation to New Modalities

After trained on all the combinations of the six seen modalities, we fix the two Transformer in our model, and append few parameters $(\mathcal{E}_{m_u}, Q_{m_u})$ for the new modalities $m_u \in \{A, Ne\}$. The 2nd figure in Fig. 3 (c) compares the performance of updating only $(\mathcal{E}_{m_u}, Q_{m_u})$, updating the whole model and the baseline model solely trained for new modalities, while the 3rd figure compares the parameter cost of the three strategies. The results show that comparable performance can be achieved at significantly reduced cost. The 1st figure shows our performance on unseen combinations without any further training. We marked the performance gain and drop by adding A or Ne on three base combinations. The results show performance gain in orange for most cases, with all positive results by adding both new modalities, but not always gain from adding modalities.

4 Conclusion

This work solves unified multi-modal learning for the challenging setting of diverse and imbalanced medical modalities, which involves a new task to address any missing scenarios, as well as new modalities and unseen combinations. This is achieved by decoupled alignment to ensure both feature exploration and modality-invariance, facilitated with unified architectural design for both fusion and representation learning. Experimental results demonstrate that the proposed unified model exhibits a growing advantage to separately trained models as the number of modalities increases, as well as the easy adaptation to new modalities.

Acknowledgments. The study was funded by the General Research Fund of Hong Kong Research Grants Council (No. 15218521); the grant under Theme-based Research Scheme of Hong Kong Research Grants Council (No. T45-401/22-N).

Disclosure of Interests. Author Feng Yidan has received research grants from the General Research Fund of Hong Kong Research Grants Council (No. 15218521) and Theme-based Research Scheme of Hong Kong Research Grants Council (No. T45-401/22-N).

References

1. Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L.A., Wilson, K.T., Landman, B., Huo, Y.: Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering (2023)
2. Duan, J., Chen, L., Tran, S., Yang, J., Xu, Y., Zeng, B., Chilimbi, T.: Multi-modal alignment using representation codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15651–15660 (2022)
3. 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)
4. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems **34**, 18932–18943 (2021)
5. Huang, Y., Xu, J., Zhou, Y., Tong, T., Zhuang, X., , t.A.D.N.I.A.: Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network. Frontiers in Neuroscience **13** (2019). https://doi.org/10.3389/fnins.2019.00509, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2019.00509
6. 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)
7. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. pp. 4651–4664. PMLR (2021)
8. 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)

9. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: Multimodal prompting with missing modalities for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14943–14952 (2023)
10. Liu, L., Liu, S., Zhang, L., To, X.V., Nasrallah, F., Chandra, S.S.: Cascaded multimodal mixing transformers for alzheimer's disease classification with incomplete data. NeuroImage **277**, 120267 (2023)
11. 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)
12. Shvetsova, N., Chen, B., Rouditchenko, A., Thomas, S., Kingsbury, B., Feris, R.S., Harwath, D., Glass, J., Kuehne, H.: Everything at once-multi-modal fusion transformer for video retrieval. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 20020–20029 (2022)
13. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems **29** (2016)
14. Song, J., Zheng, J., Li, P., Lu, X., Zhu, G., Shen, P.: An effective multimodal image fusion method using mri and pet for alzheimer's disease diagnosis. Frontiers in digital health **3**, 637386 (2021)
15. Tu, Y., Lin, S., Qiao, J., Zhuang, Y., Zhang, P.: Alzheimer's disease diagnosis via multimodal feature fusion. Computers in Biology and Medicine **148**, 105901 (2022)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
17. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15878–15887 (2023)
18. Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., Wang, X., Zhou, C.: One-peace: Exploring one general representation model toward unlimited modalities. arXiv preprint arXiv:2305.11172 (2023)
19. Wenfang Yao, Kejing Yin, W.K.C.J.L.J.Q.: Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
20. Woo, S., Lee, S., Park, Y., Nugroho, M.A., Kim, C.: Towards good practices for missing modality robust action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2776–2784 (2023)
21. Yao, J., Zhu, X., Zhu, F., Huang, J.: Deep correlational learning for survival prediction from multi-modality data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 406–414. Springer (2017)
22. Zhou, R., Zhou, H., Chen, B.Y., Shen, L., Zhang, Y., He, L.: Attentive deep canonical correlation analysis for diagnosing alzheimer's disease using multimodal imaging genetics. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 681–691. Springer (2023)
23. Zhou, T., Thung, K.H., Zhu, X., Shen, D.: Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Human brain mapping **40**(3), 1001–1016 (2019)
24. Zuo, H., Liu, R., Zhao, J., Gao, G., Li, H.: Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)