Semantic-Aware Multimodal Fusion with Cross-Scale Features for Alzheimer's Disease Detection

Rui Gu[1], **Lei Jin**[1], **Chunfeng Yang (chunfeng.yang@seu.edu.cn)**[1], **Xiaojia Wang**[2], **Yuting He**[3] **& Yang Chen**[1]
[1]School of Computer Science and Engineering, Southeast University, Nanjing, 211189, China
[2]School of Internet of Things and Artificial Intelligence, Wuxi Vocational College of Science and Technology, Wuxi, 214028, China
[3]Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH 44106, USA

Abstract

With the increasing prevalence of Alzheimer's disease (AD), automatic detection has emerged as a practical auxiliary diagnostic tool. Communication-related signals, being non-invasive, have been widely used in unimodal and multimodal approaches for AD detection with promising results. However, existing methods face two limitations: (1) relying heavily on final-layer encoder representations prevents comprehensive characterization of modality-specific features, and (2) treating all modalities equally introduces noise and reduces discriminative capability. Therefore, we propose SCMF-Net, a multimodal fusion framework. Specifically, it incorporates a cross-scale feature alignment module to exploit complementary representations across encoder layers, producing comprehensive modality-specific representations. In addition, SCMF-Net integrates a semantic-aware attention module that uses textual semantics to guide cross-modal fusion, thereby promoting semantic alignment. Experiments on the ADReSS and ADReSSo benchmark datasets demonstrate the effectiveness of SCMF-Net, achieving accuracies of 91.67% and 90.14%, respectively, with ablation studies validating the contribution of each component.

Keywords:

Alzheimer's disease; multimodal fusion; cross-scale features; semantic-aware attention

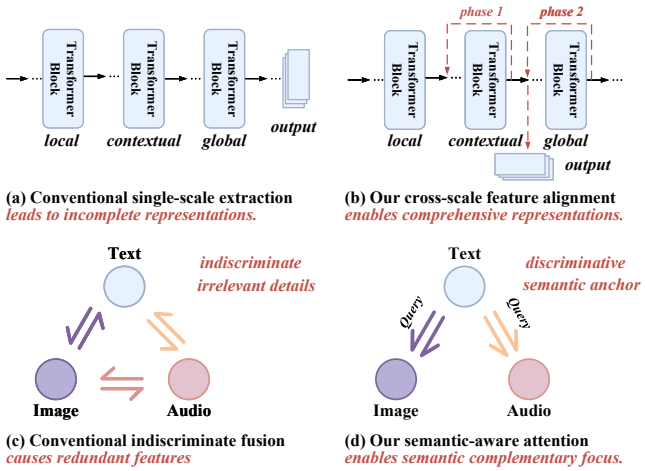

Figure 1: **Comparison between conventional methods and our method**: (a–b) Our method integrates cross-scale features, enabling more comprehensive representations beyond final-layer modeling. (c–d) Our semantic-aware attention leverages text to guide fusion, yielding more discriminative multimodal representations.

Introduction

Alzheimer's Disease (AD) represents one of the most critical public health challenges worldwide, as it is the leading cause of dementia. Its prevalence is projected to reach 115 million by 2050 (Wortmann, n.d.). Given its irreversible cognitive decline and the absence of curative treatments, early detection is of paramount importance (Livingston et al., 2017). However, conventional diagnostic approaches are often costly and invasive (Khan & Alkon, 2015; Tang & Tian, 2025). In contrast, speech- and language-based automatic detection provides a cost-effective and non-invasive alternative by capturing early-stage linguistic impairments across multiple levels (Szatloczki et al., 2015).

Both unimodal and multimodal approaches have been extensively explored for automatic AD detection. Early studies primarily relied on unimodal approaches based on audio or text, modeling cognitive and linguistic impairments through acoustic, prosodic, and linguistic features, and later through deep representations such as wav2vec 2.0 and BERT (Baevski et al., 2020; Eyben et al., 2015; Schuller et al., 2010). However, unimodal methods provide only a partial view of AD, motivating increasing interest in multimodal fusion. This paradigm is consistent with the inherently nature of human cognition, in which language, speech, and vision jointly reflect cognitive state. (Drijvers & Holler, 2023; Holler & Levinson, 2019). Early multimodal strategies, including early and late fusion (Pappagari et al., 2021; Pompili et al., 2020; M. S. S. Syed et al., 2020), are limited in modeling fine-grained cross-modal interactions, which has led to the development of intermediate fusion mechanisms based on attention and Transformers (Ilias & Askounis, 2023; Ying et al., 2023; Zhang et al., 2024). More recently, visual information has been incorporated into the Cookie Theft task to further enhance multimodal representations for AD detection (Bouazizi et al., 2023; Wang et al., 2024; Zhu et al., 2023). Nevertheless, current research exhibits two notable limitations.

First, a key limitation in current multimodal studies is the insufficient exploitation of cross-scale representations from different encoder layers, which may result in an incomplete characterization within each modality, as illustrated in Fig. 1(a) (Bang et al., 2024; Pappagari et al., 2021). **Specifically**, Speech and language impairments in AD affect acoustic

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

and linguistic structures in a stage-dependent manner (Emery, 2000; Fraser et al., 2015), suggesting that informative cues emerge at multiple representational scales rather than a single abstraction level. Notably, encoder-focused studies suggest that different encoder layers capture information at different scales, yielding representations with rich multi-scale information (Jawahar et al., 2019; Pasad et al., 2021). **Motivated by these observations**, we propose a Cross-scale Feature Alignment (CFA) module, as illustrated in Fig. 1(b), which aligns multi-layer representations through two successive phases of cross-attention. The first phase aligns local- and contextual-scale features, followed by alignment with global-scale features to integrate complementary cross-scale information.

Second, another limitation in existing methods is treating all modalities equally during fusion, which can degrade the model performance when unreliable modalities dominate the fusion process, as illustrated in Fig. 1(c) (Ilias & Askounis, 2023; Ying et al., 2023). **In practice**, textual representations exhibit more structured characteristics and contain richer semantic information, as text tokens explicitly encode lexical units and their contextual dependencies, whereas acoustic and visual modalities lack such structured organization and are more prone to learning irrelevant or noisy details (Fraser et al., 2014; Liu et al., 2024). **Therefore**, we propose a semantic-aware attention (SAA) module, as shown in Fig. 1(d), which leverages textual representations to guide multimodal fusion via cross-attention. By emphasizing task-relevant information and suppressing irrelevant features, SAA yields more discriminative multimodal representations for AD detection.

Based on the above analysis, we propose SCMF-Net, which jointly exploits cross-scale alignment and semantic-aware fusion to enhance clinically relevant representations and improve the reliability of AD detection. The main contributions are as follows:

1. We propose SCMF-Net, a semantic-aware cross-scale multimodal fusion framework for AD detection, which enables effective intra-modal alignment and inter-modal fusion.
2. We design a Cross-Scale Feature Alignment (CFA) module that integrates complementary representations extracted from different encoder layers, enabling more comprehensive modeling within each modality.
3. We propose a Semantic-Aware Attention (SAA) module that employs semantically complementary focus between audio and image modalities, thereby enhancing model's discriminative capability by reducing task-irrelevant redundancy.
4. Experiments on the ADReSS and ADReSSo datasets show that the proposed framework achieves accuracies of 91.67% and 90.14%, respectively.

Methodology

We propose SCMF-Net, a semantic-aware cross-scale multimodal fusion framework for automated AD detection. The proposed method follows a unified pipeline that progressively aligns multimodal features across different scales and enhances cross-modal interactions under semantic guidance.

Overall Framework

As illustrated in Fig. 2, our framework integrates text, audio, and image features through a Cross-Scale Feature Alignment (CFA) module and a Semantic-Aware Attention (SAA) module, enabling effective intra-modal alignment and inter-modal fusion. Specifically, we adopt the BERT-base[1], Wav2Vec2-base[2], and ViT-base[3] as backbone encoders. To reduce overfitting, we freeze the first nine layers and fine-tune only the top three. Given all encoders consist of 12 stacked Transformer layers and produces a set of layer-wise representations denoted as $\mathbf{H}^m = \{\mathbf{H}_1^m, \mathbf{H}_2^m, \ldots, \mathbf{H}_{12}^m\}$, where $m \in \{t, a, i\}$ indicates the modality and $\mathbf{H}_\ell^m$ represents the output of the ℓ-th Transformer layer.

Within each modality, features extracted from different encoder layers are first fused through CFA module to exploit complementary information across depths. We denote the output of the CFA module as $\mathbf{F}_{\text{out}}^m$. Given the selected layer-wise representations $\{\mathbf{H}_\ell^m\}_{\ell \in \mathcal{S}}$, CFA is formulated as $\mathbf{F}_{\text{out}}^m = \text{CFA}\big(\{\mathbf{H}_\ell^m\}_{\ell \in \mathcal{S}}\big)$, where $\mathcal{S}$ corresponds to the layers capturing local, contextual, and global representations, respectively. The modality-specific representations are then refined by SAA module to facilitate inter-modal interaction. We denote the output of the SAA module as $\mathbf{f}_{\text{concat}}$. Given the CFA-enhanced features $\{\mathbf{F}_{\text{out}}^m\}$, SAA is defined as $\mathbf{f}_{\text{concat}} = \text{SAA}\big(\{\mathbf{F}_{\text{out}}^m\}\big)$. Finally, the fused multimodal representation is fed into a multilayer perceptron (MLP) classifier to produce the final prediction. Overall, the proposed framework is formulated as:

$$\hat{y} = \text{MLP}\Big(\text{SAA}\big(\{\text{CFA}(\{\mathbf{H}_\ell^m\}_{\ell \in \mathcal{S}})\}_{m \in \{t,a,i\}}\big)\Big), \quad (1)$$

The model is trained end-to-end using the cross-entropy loss $\mathcal{L}_{\text{CE}}$, which is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \quad (2)$$

where C is the number of classes, y_c denote the ground-truth label. Details of the CFA and SAA modules are provided in the subsequent subsections.

CFA module: Cross-Scale Feature Alignment

To exploit complementary information from different encoder layers, we design a CFA module to align cross-scale representations, as shown in Fig. 3. The module aggregates features from multiple layers using cross-attention and residual connections. Specifically, the 3rd, 7th, and 12th Transformer layers are selected to represent local-, contextual-, and global-scale features, respectively. Before alignment, a BiLSTM is

[1]https://huggingface.co/bert-base-uncased
[2]https://huggingface.co/facebook/wav2vec2-base
[3]https://huggingface.co/google/vit-base-patch16-224

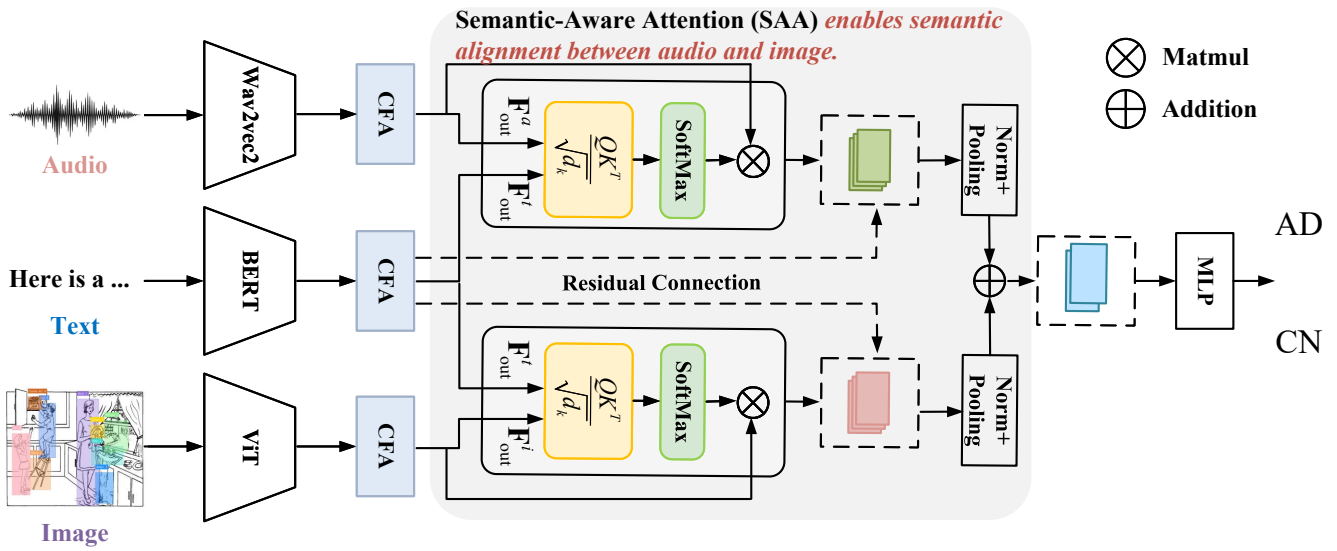

Figure 2: Overview of our SCMF-Net framework. Three modalities are processed in parallel, followed by cross-scale feature alignment and semantic-aware attention for multimodal interaction, before final prediction via an MLP classifier.

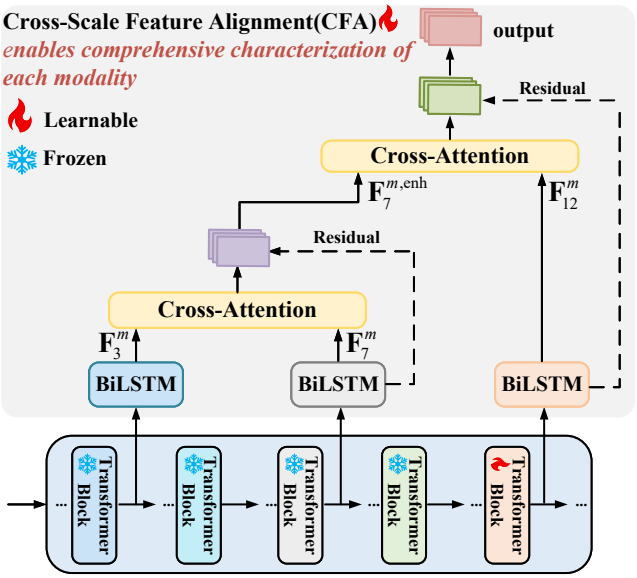

Figure 3: Structure of the CFA module. cross-scale features extracted from the 3rd, 7th, and 12th Transformer layers are fused via cross-attention and residual connections.

employed to capture contextual dependencies and reduce feature redundancy. The refined representation of modality m at layer j is denoted as $\mathbf{F}_j^m$:

$$\mathbf{F}_j^m = \text{BiLSTM}\left(\mathbf{H}_j^m\right), \quad j \in \{3, 7, 12\}, \tag{3}$$

where $\mathbf{H}_j^m \in \mathbb{R}^{B\times L_m\times 768}$, B is the batch size, L_m is the sequence length of modality m, 768 denotes the feature dimension of the encoder output, and $\mathbf{F}_j^m \in \mathbb{R}^{B\times L_m\times 128}$, where 128 denotes the concatenated hidden dimension of the BiLSTM.

First-Phase Alignment The local-scale features $\mathbf{F}_3^m$ capture fine-grained structural details, while the contextual-scale features $\mathbf{F}_7^m$ encode richer contextual semantics. To inject fine-grained local-scale structural cues into contextual-scale representations, we use cross-attention mechanism in which $\mathbf{F}_7^m$ queries $\mathbf{F}_3^m$, allowing contextual-scale features to selectively aggregate informative local-scale details. The resulting representation is denoted as $\tilde{\mathbf{F}}_7^m$:

$$\text{CrossAttn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{4}$$

$$\tilde{\mathbf{F}}_7^m = \text{CrossAttn}\left(\mathbf{F}_7^m, \mathbf{F}_3^m, \mathbf{F}_3^m\right), \tag{5}$$

where $\tilde{\mathbf{F}}_7^m \in \mathbb{R}^{B\times L_m\times 128}$.

To integrate the injected local-scale structural cues with the original contextual-scale representation, the enhanced feature, denoted as $\mathbf{F}_7^{m,\text{enh}}$, is obtained via a residual fusion strategy, $\mathbf{F}_7^{m,\text{enh}} = \mathbf{F}_7^m + \tilde{\mathbf{F}}_7^m$, with $\mathbf{F}_7^{m,\text{enh}} \in \mathbb{R}^{B\times L_m\times 128}$.

Second-Phase Alignment While the global-scale feature $\mathbf{F}_{12}^m$ captures high-level semantic information, it may lack explicit access to contextual-scale cues. To refine global-scale semantic representations with enriched contextual information, we adopt $\mathbf{F}_{12}^m$ as the query to attend to the enhanced contextual-scale representation $\mathbf{F}_7^{m,\text{enh}}$. The resulting cross-scale enhanced global representation is denoted as $\tilde{\mathbf{F}}_{12}^m$:

$$\tilde{\mathbf{F}}_{12}^m = \text{CrossAttn}\left(\mathbf{F}_{12}^m, \mathbf{F}_7^{m,\text{enh}}, \mathbf{F}_7^{m,\text{enh}}\right), \tag{6}$$

where $\tilde{\mathbf{F}}_{12}^m \in \mathbb{R}^{B\times L_m\times 128}$.

The final output representation, denoted as $\mathbf{F}_{\text{out}}^m$, is obtained following the same residual fusion strategy, $\mathbf{F}_{\text{out}}^m = \mathbf{F}_{12}^m + \tilde{\mathbf{F}}_{12}^m$, with shape $\mathbf{F}_{\text{out}}^m \in \mathbb{R}^{B\times L_m\times 128}$.

SAA mdoule: Semantic-Aware Attention

Given the structured architecture and strong semantic expressiveness of textual representations, we introduce a SAA module. In this module, text features act as queries to provide guidance for cross-modal fusion, while audio and image features serve as keys and values. The resulting semantic-aware representation from text to modality $u \in \{a, i\}$ is denoted as $\mathbf{A}^{t\to u}$ and computed as:

$$\mathbf{A}^{t\to u} = \text{CrossAttn}\left(\mathbf{F}^{t}_{\text{out}}, \mathbf{F}^{u}_{\text{out}}, \mathbf{F}^{u}_{\text{out}}\right), \tag{7}$$

where $\mathbf{A}^{t\to u} \in \mathbb{R}^{B\times L_u\times 128}$, and L_u is the sequence length of modality u.

Then, the modality-specific feature sequence, denoted as $\mathbf{F}'_u$, is obtained by residual fusion followed by Layer Normalization, $\mathbf{F}'_u = \text{LayerNorm}(\mathbf{F}^{t}_{\text{out}} + \mathbf{A}^{t\to u})$, with shape $\mathbf{F}'_u \in \mathbb{R}^{B\times L_u\times 128}$. A global modality-specific representation, denoted as $\mathbf{f}'_u$, is computed by temporal average pooling over the enhanced sequence, $\mathbf{f}'_u = \frac{1}{L_u}\sum_{i=1}^{L_u}\mathbf{F}'_{u,i}$, with $\mathbf{f}'_u \in \mathbb{R}^{B\times 128}$; the unified multimodal representation, denoted as $\mathbf{f}_{\text{concat}}$, is obtained by concatenating the global audio and image representations, $\mathbf{f}_{\text{concat}} = [\mathbf{f}'_a, \mathbf{f}'_i]$, with shape $\mathbf{f}_{\text{concat}} \in \mathbb{R}^{B\times 256}$.

Experiment

Experimental Setting

Dataset To ensure reliable and comparable experiments, we conduct evaluations on two widely recognized benchmarks for AD detection. Specifically, the **ADReSS dataset** (Luz et al., 2020) comprises 78 non-AD and 78 AD participants with speech recordings and transcripts from the Cookie Theft picture description task, and the **ADReSSo dataset** (Luz et al., 2021) includes 119 non-AD and 118 AD participants collected using the same task. Since transcripts are not provided in ADReSSo, text data are generated using the Whisper-base model[4].

Data Preprocessing To handle variable-length transcripts, all text sequences are truncated or padded to 512 tokens, with padding tokens masked during training. In addition, semantically equivalent words are mapped to canonical labels to reduce lexical variability and obtain more consistent textual representations. Subsequently, textual transcripts are encoded into image representations by mapping word occurrence frequencies to the color intensities of predefined image regions. The resulting images are resized to 224×224 pixels and divided into non-overlapping 16×16 patches for subsequent feature extraction. Similarly, to handle variable-length audio recordings, all audio samples are truncated or zero-padded to 60 s, with padding regions masked during training.

Implementation and Metrics We employ 5-fold cross-validation for hyperparameter tuning. The optimal configuration is used to retrain models on the entire training set with five random seeds, and results on the test set are averaged across these runs. Given balanced datasets, accuracy is used as the primary evaluation metric, while precision, recall, and F1-score are reported as complementary metrics. Key hyperparameters include: learning rate = 1e-5, batch size = 4, epoch = 15, and dropout rate = 0.3. All experiments were conducted on an Ubuntu server with an NVIDIA RTX 5090 GPU using Python 3.13.5 and PyTorch 2.9.1.

[4]https://huggingface.co/openai/whisper-base

Comparison study

To evaluate the effectiveness of the proposed method, we compare it with representative approaches from the recent literature. Tables 1 reports the results on the ADReSS and ADReSSo datasets. All comparison results of existing methods are directly taken from their original publications. The results on both datasets indicate that the proposed model provides competitive performance across evaluation metrics. On the ADReSS dataset, the proposed model achieves the best overall performance, surpassing the next-best method by 2.09% in accuracy, 3.93% in precision, and 1.85% in F1 score, which demonstrates strong discriminative capability. On the ADReSSo dataset, the proposed model further improves performance, achieving an accuracy gain of 1.96% over the next-best baseline, while also delivering the highest recall and F1 score among all compared methods. Although Pu et al. report a higher precision, our approach maintains relatively balanced performance.

Ablation Study

We conduct ablation experiments on the ADReSS dataset, and the results are presented in the following subsections.

Modality Ablation To assess the contribution of each modality, we conduct ablation experiments with individual and combined inputs. As shown in Table 2, the text modality (T) outperforms audio (A) and image (I) modalities when used independently, indicating that linguistic representations provide more discriminative cues for the task and exhibit relatively reliable standalone performance in the absence of other modalities. Multimodal fusion consistently improves performance over single-modality models, with the trimodal setting (T+A+I) achieving the best results. These findings suggest that audio and text offer complementary prosodic–semantic cues, while image features provide additional structural information that further enriches semantic representation.

Module Ablation To assess the contribution of each module, we perform ablation experiments. As shown in Table 3, replacing the SAA module with a generic cross-modal attention mechanism results in the largest performance drop, with accuracy decreasing by 5.84%, highlighting the importance of semantic-aware fusion. Excluding the CFA module, where cross-scale features are simply aggregated, results in a 5.39% accuracy decline. In contrast, removing cross-scale features entirely yields an accuracy improvement of 2.05% compared to the variant without CFA, suggesting that cross-scale representations may introduce redundancy when not coupled with

Table 1: **Comparison study:** Our method achieves the best overall performance on both the ADReSS and ADReSSo datasets compared with representative prior approaches, demonstrating its superior model performance. (LDA = Linear Discriminant Analysis, SVM = Support Vector Machine, RF = Random Forest, LogR = Logistic Regression). "Multi." indicates whether the model uses multimodal input. Dashes indicate that the corresponding evaluation metric was not reported.

Dataset	Multi.	Model	Accuracy	Precision	Recall	F1
ADReSS	No	LDA (Luz et al., 2020)	75.00	76.50	74.50	74.50
	No	BERT (Guo et al., 2021)	82.10	–	–	–
	No	BERT+RF (Haulcy & Glass, 2021)	85.40	–	–	–
	Yes	LogR (Martinc & Pollak, 2020)	77.08	–	–	–
	Yes	Transformer+VGG (Koo et al., 2020)	81.25	82.67	81.25	81.05
	Yes	Transformer+BERT (Pan et al., 2024)	85.42	86.20	85.42	85.47
	Yes	BiLSTM+DistilBERT. (Zhang et al., 2024)	89.58	88.00	**91.67**	89.80
	Yes	BERT+Whisper (Abid et al., 2025)	86.32	86.14	85.28	86.69
	Yes	**BERT+Wav2vec 2.0+ViT (Ours)**	**91.67**	**91.93**	**91.67**	**91.65**
ADReSSo	No	SVM (Luz et al., 2021)	78.87	78.90	78.90	78.87
	No	BERT(Z. S. Syed et al., 2021)	84.51	–	–	–
	No	BERT(Pan et al., 2021)	84.51	84.73	84.57	84.50
	No	BERT(Pu & Zhang, 2025)	83.10	**92.60**	71.40	80.60
	Yes	BERT+Wav2vec 2.0 (Ying et al., 2023)	83.70	83.80	83.80	83.80
	Yes	Wav2vec 2.0+Word2Vec (Yang et al., 2024)	85.50	85.90	85.40	85.00
	Yes	BERT+Wav2vec 2.0 (Bang et al., 2024)	87.32	88.06	87.32	87.25
	Yes	BERT+Whisper (Abid et al., 2025)	88.18	89.27	88.54	87.86
	Yes	**BERT+Wav2vec 2.0+ViT (Ours)**	**90.14**	89.31	**89.29**	**89.36**

Table 2: **Modality ablation study:** The combination of text, audio, and image achieves the best overall performance, validating the complementary contributions of multiple modalities. (T = Text, A = Audio, I = Image).

T	A	I	Accuracy	Precision	Recall	F1
✓			80.16	81.23	82.65	81.58
	✓		72.51	72.25	72.58	77.74
		✓	75.63	77.33	78.37	80.14
✓	✓		87.34	87.17	87.34	87.42
✓		✓	85.42	85.98	85.42	85.36
	✓	✓	81.25	81.30	81.25	81.24
✓	✓	✓	**91.67**	**91.93**	**91.67**	**91.65**

effective alignment mechanisms. Overall, these results show that cross-scale representations are effective only with proper alignment, while semantic-aware fusion is crucial for multimodal integration.

Table 3: **Module ablation study**: Removing individual modules leads to performance degradation, demonstrating the effectiveness of each component. (w/o = without).

Module	Accuracy	Precision	Recall	F1
w/o SAA	85.83	86.26	85.66	85.79
w/o CFA	86.28	87.00	86.27	86.45
w/o cross-scale	88.33	88.68	88.42	88.19
full model	**91.67**	**91.93**	**91.67**	**91.65**

Number of Scales Selection To study the effect of scale diversity, we vary the number of fused scales, where each scale corresponds to features from different encoder layers with increasing contextual scope. As shown in Fig. 5, integrating three scales achieves the best performance. Compared with this optimal configuration, using only one or two scales results in accuracy drops of 3.34% and 2.77%, respectively, indicating that insufficient scale coverage limits the model's ability to capture complementary information. In contrast, increasing the number of fused scales to four or five leads to performance decreases of 2.09% and 6.25%, respectively, suggesting that excessive scale aggregation introduces redundant or misaligned cues.

Scale Combination Selection To evaluate the effect of combining representations at different scales, we examine different scale combination strategies. As shown in Table 4, performance improves from local- to contextual- and global-scale representations, indicating that coarser scales capture more discriminative information. While most cross-scale alignment strategies further improve performance, directly combining local- and global-scale representations results in a 2.38% accuracy drop compared to using global-scale features alone, suggesting that excessive scale discrepancies hinder effective integration. In contrast, jointly integrating local-, contextual-, and global-scale representations achieves the best performance, outperforming the second-best configuration by 1.39% in accuracy. These results suggest that the effectiveness of cross-scale integration is influenced by the choice of scale combinations.

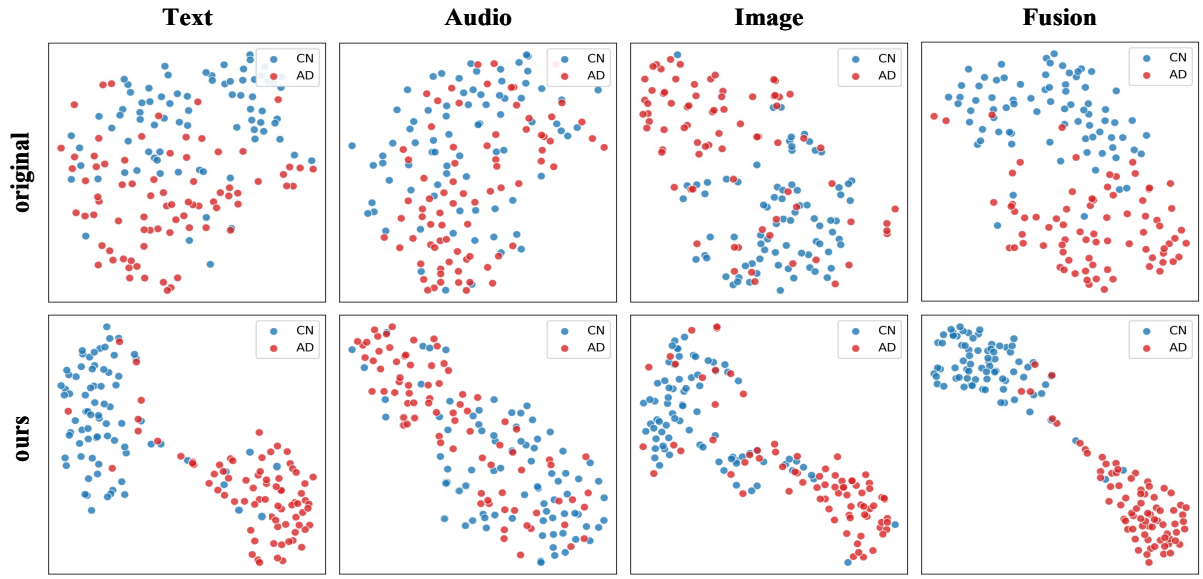

Figure 4: t-SNE visualization of learned feature representations. Top: traditional approach using last-layer single-modality features and cross-modal attention fusion. Bottom: proposed approach with cross-scale alignment and semantic-aware attention. Blue and red dots denote CN and AD samples, respectively.

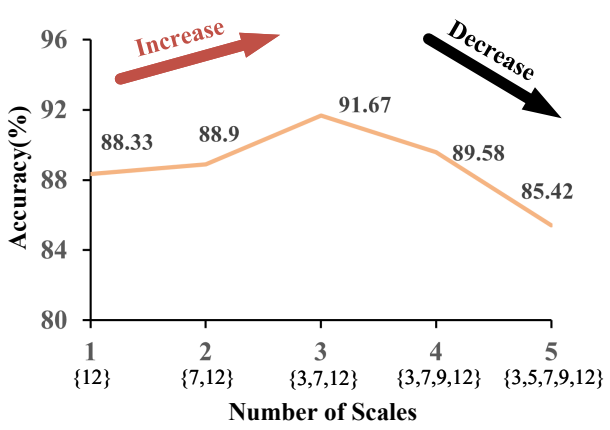

Figure 5: **Scale number study**: The three-scale configuration yields the best performance among different scale settings.

Table 4: **Scale combination study**: Combining local, contextual, and global scales yields the best overall performance. (L = Local-scale, C = Contextual-scale, G = Global-scale).

Level Set	Levels	Accuracy	Precision	Recall	F1
L	2, 3, 4	67.42	68.45	67.28	67.71
C	6, 7, 8	83.33	84.29	83.33	83.22
G	10, 11, 12	88.58	88.65	88.41	88.36
L + C	3, 4, 8	85.42	85.98	85.42	85.36
L + G	3, 4, 12	86.20	86.76	86.50	85.48
C + G	6, 7, 12	90.28	90.21	89.58	90.54
L + C + G	3, 7, 12	**91.67**	**91.93**	**91.67**	**91.65**

Visualization Analysis

To visually analyze the effectiveness of the proposed components, we present t-SNE visualizations of multimodal representations learned under different feature modeling strategies, as shown in Fig. 4. Using only last-layer features yields dispersed and unstructured representations, indicating that single-modality characteristics are not sufficiently captured. Incorporating cross-scale feature alignment yields more compact intra-class distributions across modalities, highlighting the benefit of multi-scale integration. While conventional cross-modal attention improves discriminability to some extent, the resulting representations remain loosely structured and exhibit limited class separation. In contrast, our semantic-aware attention produces the most compact clusters with clearer class boundaries. These observations confirm that both proposed components improve the structure and discriminative power of multimodal representations.

Conclusion

This paper presents SCMF-Net, a multimodal framework for automatic AD detection that addresses key limitations in existing approaches through cross-scale feature alignment and semantic-aware fusion. Experimental results on ADReSS and ADReSSo datasets demonstrate its effectiveness, achieving accuracies of 91.67% and 90.14%, respectively, outperforming the best baseline methods on each dataset by 2.09% and 1.96% in accuracy. Despite these results, the framework's dependence on high-quality multimodal data and limited testing on diverse populations require further study. Future work will incorporate clinical metadata to improve interpretability and validate the approach on larger, cross-institutional datasets.

Acknowledgments

This research was funded by the National Key Research and Development Program of China (Grant No. 2022YFC2405600), the National Natural Science Foundation of China (Grants No. T2225025). This research work was supported by the Big Data Computing Center of Southeast University.

References

Abid, S., Zhang, D., Shehzad, A., Ren, J., Yu, S., Lin, H., & Xia, F. (2025). Speechhgt: A multimodal hypergraph transformer for speech-based early alzheimer's disease detection. *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 9131–9139. https://doi.org/10.24963/IJCAI.2025/1015

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460. https://doi.org/10.48550/arXiv.2006.11477

Bang, J.-U., Han, S.-H., & Kang, B.-O. (2024). Alzheimer's disease recognition from spontaneous speech using large language models. *ETRI Journal*, *46*(1), 96–105. https://doi.org/10.4218/etrij.2023-0356

Bouazizi, M., Zheng, C., Yang, S., & Ohtsuki, T. (2023). Dementia detection from speech: What if language models are not the answer? *Information*, *15*(1), 2. https://doi.org/10.3390/info15010002

Drijvers, L., & Holler, J. (2023). The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review*, *30*(2), 792–801. https://doi.org/10.3758/s13423-022-02178-x

Emery, V. O. B. (2000). Language impairment in dementia of the alzheimer type: A hierarchical decline? *The International Journal of Psychiatry in Medicine*, *30*(2), 145–164.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190–202.

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., & Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *cortex*, *55*, 43–60.

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2015). Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's disease*, *49*(2), 407–422.

Guo, Y., Li, C., Roan, C., Pakhomov, S., & Cohen, T. (2021). Crossing the "cookie theft" corpus chasm: Applying what bert learns from outside data to the adress challenge dementia detection task. *Frontiers in Computer Science*, *3*, 642517.

Haulcy, R., & Glass, J. (2021). Classifying alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology*, *11*, 624137.

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in cognitive sciences*, *23*(8), 639–652.

Ilias, L., & Askounis, D. (2023). Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech. *Knowledge-Based Systems*, *277*, 110834.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does bert learn about the structure of language? *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Khan, T. K., & Alkon, D. L. (2015). Alzheimer's disease cerebrospinal fluid and neuroimaging biomarkers: Diagnostic accuracy and relationship to drug efficacy. *Journal of Alzheimer's Disease*, *46*(4), 817–836.

Koo, J., Lee, J. H., Pyo, J., Jo, Y., & Lee, K. (2020). Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition. *arXiv preprint arXiv:2009.04070*.

Liu, Y.-L., Feng, R., Yuan, J.-H., & Ling, Z.-H. (2024). Clever hans effect found in automatic detection of alzheimer's disease through speech. *arXiv preprint arXiv:2406.07410*.

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., Ballard, C., Banerjee, S., Burns, A., Cohen-Mansfield, J., et al. (2017). Dementia prevention, intervention, and care. *The lancet*, *390*(10113), 2673–2734.

Luz, S., Haider, F., De la Fuente, S., Fromm, D., & MacWhinney, B. (2021). Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint arXiv:2104.09356*.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: The adress challenge.

Martinc, M., & Pollak, S. (2020). Tackling the adress challenge: A multimodal approach to the automated recognition of alzheimer's dementia. *Interspeech*, 2157–2161.

Pan, Y., Mirheidari, B., Harris, J. M., Thompson, J. C., Jones, M., Snowden, J. S., Blackburn, D., & Christensen, H. (2021). Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech. *Interspeech*, 3810–3814.

Pan, Y., Shi, Y., Zhang, Y., & Lu, M. (2024). Swin-bert: A feature fusion system designed for speech-based alzheimer's dementia detection. *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, 1–5.

Pappagari, R., Cho, J., Joshi, S., Moro-Velázquez, L., Zelasko, P., Villalba, J., & Dehak, N. (2021). Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios. *Interspeech*, *2021*, 3825–3829.

Pasad, A., Chou, J.-C., & Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 914–921.

Pompili, A., Rolland, T., & Abad, A. (2020). The inesc-id multi-modal system for the adress 2020 challenge. *arXiv preprint arXiv:2005.14646*.

Pu, Y., & Zhang, W.-Q. (2025). Integrating pause information with word embeddings in language models for alzheimer's disease detection from spontaneous speech. *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. S. (2010). The interspeech 2010 paralinguistic challenge.

Syed, M. S. S., Syed, Z. S., Lech, M., & Pirogova, E. (2020). Automated screening for alzheimer's dementia through spontaneous speech. *Interspeech*, *2020*, 2222–6.

Syed, Z. S., Syed, M. S. S., Lech, M., & Pirogova, E. (2021). Tackling the adresso challenge 2021: The muet-rmit system for alzheimer's dementia recognition from spontaneous speech. *Interspeech*, 3815–3819.

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in aging neuroscience*, *7*, 195.

Tang, B., & Tian, Y. (2025). Mam-gan: Multimodal association modeling based on generative adversarial networks for alzheimer's disease diagnosis. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Wang, N., Wu, M., Gu, W., & Chai, Z. (2024). Hybrid self-aligned fusion with dual-weight attention network for alzheimer's detection. *IEEE Signal Processing Letters*.

Wortmann, M. (n.d.). Dementia: A global health priority—highlights from an adi and world health organization report. alzheimers res ther. 2012; 4 (5): 40.

Yang, X., Hong, K., Zhang, D., & Wang, K. (2024). Early diagnosis of alzheimer's disease based on multi-attention mechanism. *Plos one*, *19*(9), e0310966.

Ying, Y., Yang, T., & Zhou, H. (2023). Multimodal fusion for alzheimer's disease recognition. *Applied Intelligence*, *53*(12), 16029–16040.

Zhang, Z., Wang, T., Hu, Z., Yang, L.-Z., & Li, H. (2024). Dementia: A hybrid attention-based multimodal and multi-task learning framework with expert knowledge for alzheimer's disease assessment from speech. *IEEE Journal of Biomedical and Health Informatics*.

Zhu, Y., Lin, N., Liang, X., Batsis, J. A., Roth, R. M., & MacWhinney, B. (2023). Evaluating picture description speech for dementia detection using image-text alignment. *ACM Transactions on Computing for Healthcare*.