Multimodal Personality-aware Depression Detection using Machine Learning

Kaushal Damania (kd2990), Niranjan Sundararajan (ns3888), Aryamaan Saha (as7482)
Advanced Topics in DL, Columbia University

Abstract

Depression is a significant global health concern, and its objective detection remains a challenge. This project investigates the potential of multimodal machine learning for depression detection using the MPDD Challenge dataset. We explore various fusion architectures, including early, intermediate and late fusion approaches, using audio (Wav2Vec), video (OpenFace), and personalized text embedding features. Advanced multimodal architectures, including various Transformer fusion models enhanced with specific regularization techniques, were compared to address data limitations and improve stability. Experiments focused on binary classification (depression vs. nondepression). The results indicate that an attention fusion transformer with strong regularization achieved the best performance, with a Macro F1 score of 0.828. However, significant challenges were encountered, including class imbalance, sensitivity to hyperparameters, and potential overfitting, highlighting the complexities of multimodal depression detection on limited datasets.

Introduction

Major Depressive Disorder (MDD) affects millions worldwide, significantly impacting quality of life and posing a substantial socioeconomic burden. Traditional diagnosis relies heavily on subjective clinical interviews and self-report questionnaires, which can be prone to biases and inconsistencies. Consequently, there is a growing interest in developing objective, data-driven methods for depression detection and severity assessment.

Human behavior provides rich cues related to mental state. Depression often manifests through changes in speech patterns, facial expressions (e.g., reduced expressiveness, specific action units), and even written language or expressed personality traits. Multimodal machine learning offers a promising avenue by integrating information from these diverse sources (audio, video, text) to potentially capture a more holistic and robust representation of an individual's state. Combining modalities can compensate for noise or limitations within a single modality and capture complementary or correlated information, leading to improved diagnostic accuracy.

This project aims to explore and evaluate various multimodal fusion strategies for binary depression detection using the Multimodal Personality-aware Depression Detection (MPDD) Challenge dataset. We investigate how different architectural choices (early, intermediate, late fusion) and model components (MLPs, SVMs [1], LSTMs [2], Attention, Transformers [3], MoE) impact performance when integrating pre-extracted audio, video, and personalized features. The goal is to identify effective fusion techniques and understand the challenges associated with building such systems, particularly concerning data characteristics like class imbalance.

Related Work

Numerous studies have explored depression detection using multimodal and language-based cues. Williamson et al. [4] proposed a system for the AVEC 2016 Challenge that fused vocal, facial, and semantic features extracted from human-avatar interviews. Their approach incorporated neurophysiologically motivated

biomarkers (e.g., vocal tract constriction, facial coordination) and achieved a strong F1 score of 0.81 using Gaussian staircase regression. Similarly, Mallol-Ragolta et al. [5] focused on transcribed interviews from the DAIC-WoZ dataset and applied hierarchical attention networks, showing that contextual attention mechanisms are effective even without multimodal inputs.

To address the challenge of small, labeled datasets in mental health, Shen et al. [6] introduced the EATD-Corpus, a Chinese multimodal dataset containing emotional audio and textual content. Their model combined GRUs for audio and BiLSTM with attention for text, achieving state-of-the-art results while avoiding reliance on preset questions. In another approach emphasizing medical interpretability, Agarwal et al. [7] involved psychiatrists to annotate the DAIC-WoZ dataset and trained transformer-based models aligned with clinical judgment. This expert-in-the-loop design boosted the credibility and performance of the automated system.

AudiBERT [8] is a recent deep learning framework that combines pretrained BERT and audio models through dual self-attention layers to enhance depression detection from speech. The model demonstrates strong generalizability and achieves F1 scores up to 0.92 using informal question-answer data, reinforcing the value of joint audio-text modeling.

Collectively, these works inform our design choices: employing multimodal fusion architectures, leveraging pre-extracted embeddings from pretrained models (e.g., Wav2Vec, RoBERTa), and exploring attention-based mechanisms for robust classification on the MPDD dataset.

Dataset

This project utilizes the MPDD Challenge dataset (Elderly track). The dataset provides multimodal recordings and associated labels for depression detection.

- **Modalities:** We utilize the following pre-extracted features:
  - **Audio:** Wav2Vec [9] features, capturing rich speech representations learned from large unlabeled datasets.
  - **Video:** OpenFace [10] features, providing facial landmarks, action units, head pose, and eye gaze information.
  - **Personalized Features:** These embeddings are derived from rich textual descriptions associated with each individual, encompassing information grounded in PHQ-9 depression scores, BigFive-10 personality traits, and detailed demographic attributes such as age, family situation, economic status, and disease profiles. The textual data was encoded using a pretrained RoBERTa [11] model to generate 1024-dimensional embeddings. This representation aims to provide a holistic, context-aware feature vector for each participant, enabling more personalized and robust depression detection. [12]

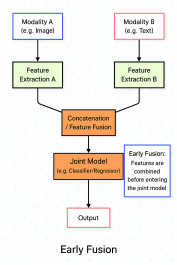
Methodology

Our methodological approach began with benchmarking a series of non-sequential baseline models, followed by progressively incorporating sequence-aware architectures and sophisticated attention-based mechanisms. This progression allowed us to both evaluate the individual contributions of different model classes and understand the role of fusion strategies in integrating multimodal information.

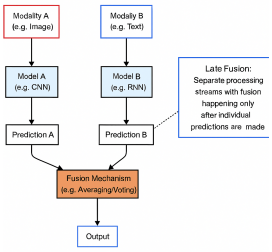
Non-Sequential Models and Fusion Strategies

We began our exploration with simpler, non-sequential models such as multilayer perceptrons (MLPs) and support vector machines (SVMs). These models operate on fixed-length feature vectors and do not account for temporal structure. Consequently, the audio and video modalities, which are inherently sequential, were

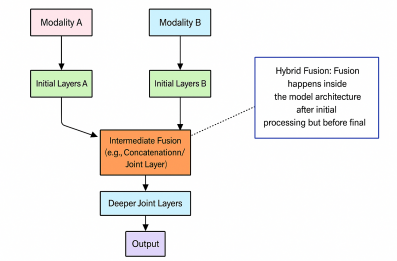
aggregated using mean pooling along the sequence dimension to obtain compact representations. Personalized features, derived from RoBERTa embeddings of demographic and trait-related text, were already static.



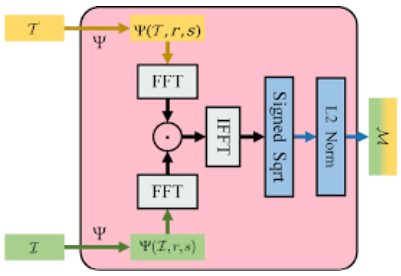
(a) Early Fusion



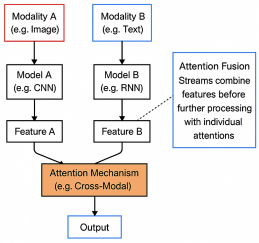
(b) Late Fusion



(c) Intermediate Fusion



(d) Compact Bilinear Pooling



(e) Attention-based Fusion

Figure 1: Architectural schematics for various fusion strategies: (a) Early Fusion, (b) Late Fusion, (c) Intermediate Fusion, (d) Compact Bilinear Pooling (CBP)[13], and (e) Attention-based Fusion. These mechanisms integrate multimodal information at different processing stages.

In this regime, we systematically investigated several fusion strategies:

- **Early Fusion:** Features from all modalities were pooled (if necessary) and concatenated at the input level. This approach is simple and enables joint representation learning from the start. However, it assumes that modalities align well in feature space, which may not hold in practice [14].
- **Late Fusion:** Each modality was processed independently with dedicated classifiers, and their logits or probabilities were combined (typically averaged) at the decision level. This preserves modality-specific information but lacks cross-modal interactions [14].
- **Intermediate Fusion:** Each modality was first encoded via a separate MLP, and their embeddings were concatenated and passed through a final classifier. This method balances the benefits of modality-specific encoding and joint fusion, allowing the model to learn richer interactions without early entanglement of raw features [14].
- **Compact Bilinear Pooling (CBP):** To capture higher-order interactions between modality pairs without incurring excessive computational overhead, we employed Compact Bilinear Pooling [15]. CBP approximates the outer product of feature vectors from different modalities using a low-dimensional sketch, enabling multiplicative feature interactions that go beyond simple concatenation. This results in more expressive fused representations while maintaining a manageable model size and training cost.
- **Attention-based Fusion:** Instead of static concatenation or averaging, we used trainable attention weights to combine modality embeddings. This allows the model to learn which modalities (or even which specific timesteps or feature dimensions) are more informative for a given prediction. Such dynamic fusion is especially beneficial in scenarios where modalities carry varying degrees of relevance across samples [16].

These baselines were crucial in identifying the relative importance of each modality and the strengths and limitations of static vs. adaptive fusion schemes. While simple to implement and interpret, these models were limited by their inability to capture temporal dynamics present in the speech and facial behavior of participants, both of which are critical cues in depression detection.

Sequential Modeling with LSTMs

To address the shortcomings of non-sequential models, we incorporated Long Short-Term Memory (LSTM) [2] networks to model the temporal evolution of audio and video features. LSTMs are well-suited for depression detection as they can capture subtle temporal patterns such as prolonged silences, intonation changes, and facial microexpressions, often indicative of depressive behavior.

Fusion strategies were adapted to the sequential nature of the data:

- In the **early fusion LSTM model**, the audio and video sequences were independently processed using LSTM encoders. The final hidden states were concatenated with the static personalized embedding, and classification was performed using a multilayer perceptron (MLP).
- The **late fusion LSTM model** treated each modality with its own sequence model (or an MLP for the personalized feature), and averaged the output logits across modalities.
- The **attention-based fusion LSTM model** introduced a shared attention mechanism over the sequence outputs of the LSTM encoders and the personalized embedding. This allowed the model to dynamically learn fusion weights over the audio, video, and personalized representations.

LSTM-based models offered a significant performance boost over static baselines, particularly in capturing modality-specific temporal nuances. However, they still relied on fixed fusion mechanisms and could be limited in modeling long-range dependencies or nuanced inter-modal interactions.

Transformer-Based Architectures

To push the limits of multimodal modeling, we implemented a suite of Transformer-based architectures. These models are capable of capturing both intra-modal and inter-modal relationships through self-attention and cross-attention mechanisms [3]. Unlike LSTMs, Transformers allow for non-sequential access patterns, enabling richer global interactions across both time and modalities.

- The **late fusion Transformer model** processed each modality—audio, video, and personalized text—independently using separate Transformer encoders. A special classification (CLS) token was used to summarize each modality. These representations were then concatenated and passed through an MLP for final prediction. While this model preserved modality-specific representations, it did not facilitate explicit cross-modal interaction.
- The **intermediate fusion Transformer model** extended the late fusion approach by introducing an additional fusion Transformer layer. A dedicated fusion CLS token attended over the modality-specific encodings, enabling more expressive and context-aware reasoning across modalities [17].
- The **attention-based fusion Transformer model** employed a unified attention mechanism across all modalities. Rather than relying on fixed fusion operations, it learned to assign attention weights to each modality dynamically, enabling more flexible and adaptive information integration.
- The **cross-attention Transformer model** incorporated both self-attention and cross-attention layers within the Transformer blocks. Audio and video sequences were allowed to attend to one another during encoding, capturing intricate inter-modal temporal dependencies prior to their fusion with the personalized features. This model enabled tightly integrated, bidirectional reasoning across modalities.
- The **adapter-based mixture-of-experts model** was a lightweight and modular design that used adapter modules trained on top of frozen modality embeddings. Each adapter learned a task-specific transformation, and the adapted features were combined via a dense mixture-of-experts (MoE) layer.

This approach allowed for efficient parameter tuning and modality-specific expert routing, improving scalability and interpretability.

Transformer-based models represented the most expressive and capable architectures in our study. They supported modality-specific encoding, flexible attention-based fusion, and complex cross-modal interactions, all essential for robust performance in multimodal depression detection. These models also demonstrated improved regularization and training stability, especially when equipped with dropout and layer normalization.

Impact of Regularization

Given the limited size of the MPDD data set and the complexity of the models explored, particularly the transformer architectures, regularization was crucial to mitigate overfitting and improve generalization. Initial analyzes revealed potential overfitting, especially in attention-based models, motivating a systematic application of various regularization strategies. We incorporated several techniques targeting different aspects of the training process.

For Transformer models, specific architectural choices served a regularizing role: the use of a CLS token aggregation strategy forced the model to summarize information effectively, and employing the GELU activation function, known to benefit Transformer performance, offered smoother gradients compared to ReLU. Optimization and training dynamics were carefully managed using techniques such as AdamW, which decouples weight decay from the gradient update, providing more effective regularization than standard Adam. We also implemented gradient clipping to prevent exploding gradients and stabilize training. A learning rate schedule that combined warm-up and cosine decay was used, balancing initial stability and cautious exploration with gradual convergence.

Furthermore, to address the significant class imbalance, Focal Loss with class weighting was used to prioritize learning hard examples and downweight easy majority-class samples, complemented by a *WeightedRandomSampler* to create more balanced mini-batches during training. Collectively, these regularization efforts helped stabilize training, reduce the gap between training and validation performance, and ultimately contributed to achieving higher Macro F1 scores, particularly for the more complex Transformer-based fusion models.

Experiments

Experimental Setup

All models were implemented in PyTorch and trained on GPU-accelerated systems, including NVIDIA CUDA-enabled devices and Apple Silicon hardware. A unified training and evaluation pipeline was used across all architectures to ensure consistent experimental conditions.

Training Protocol

We used both Adam and AdamW optimizers depending on the model family: Adam for MLPs, SVMs, and LSTM-based models, and AdamW for Transformer-based and Mixture-of-Experts architectures. To address class imbalance, we employed focal loss as the primary objective, tuning both the focusing parameter γ and class-balancing factor α for best performance.

Training stability was improved using regularization techniques such as dropout and weight decay. For deeper models, particularly Transformers, we applied cosine learning rate scheduling with warmup phases. In some cases, we rebalanced training data to emphasize minority classes.

Hyperparameter tuning was conducted using Bayesian optimization via Optuna, targeting macro F1-score as the objective. Early stopping was employed to select optimal checkpoints while reducing computational overhead.

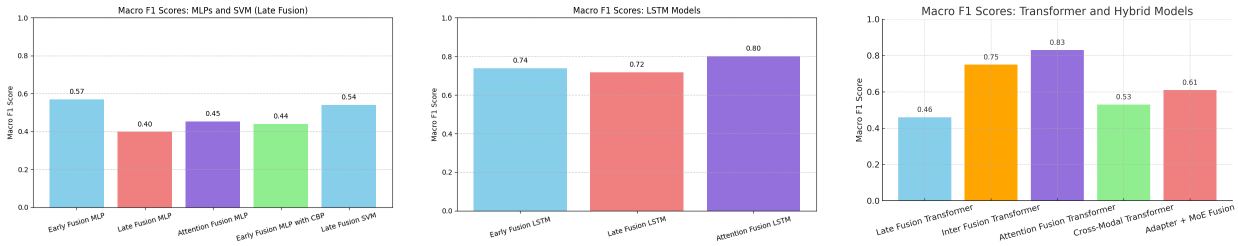
Evaluation Metrics

We report three key evaluation metrics: **macro F1-score**, **weighted F1-score**, and overall **accuracy**. Macro F1, our primary metric, captures the average F1-score across both classes without bias from class imbalance—making it especially suitable for our binary depression classification task. Weighted F1 complements this by accounting for class frequencies, while accuracy provides a general sense of overall correctness.

Results

Table 1: Summary of Model Performance.

Model Architecture	Macro Recall	Macro F1
SVM / MLP Models		
Late Fusion SVM	0.55	0.54
Early Fusion MLP (Without Focal Loss)	0.50	0.41
Early Fusion MLP	**0.56**	**0.57**
Early Fusion MLP + CBP	0.44	0.44
Late Fusion MLP	0.53	0.40
Attention Fusion MLP	0.53	0.45
LSTM-Based Models		
Early Fusion LSTM	0.72	0.74
Late Fusion LSTM	0.70	0.72
Attention Fusion LSTM	**0.78**	**0.80**
Transformer-Based and Hybrid Models		
Late Fusion Transformer	0.62	0.46
Inter Fusion Transformer	0.80	0.75
Attention Fusion Transformer	**0.87**	**0.83**
Cross-Modal Transformer	0.53	0.53
Adapter + MoE Fusion	0.62	0.61



(a) MLP and SVM Models (b) LSTM-Based Models (c) Transformer and Hybrid Models

Figure 2: Macro F1 Score Comparison Across Model Families: (a) Non-sequential models (MLPs and SVMs); (b) Sequence-aware LSTM models; (c) Transformer-based and hybrid architectures.

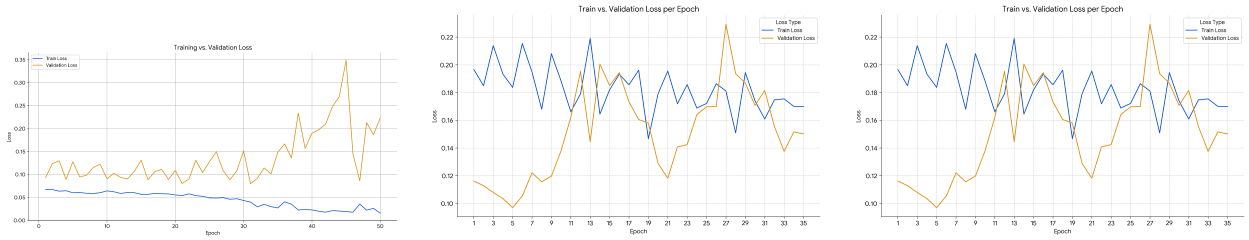
The results of our experiments, summarized in Table 1 and Figures 2a–2c, reveal clear trends in model effectiveness across architectural categories.

Non-sequential models such as multilayer perceptrons (MLPs) and support vector machines (SVMs) exhibited limited performance, with macro F1-scores generally below 0.60. While the application of focal loss improved these models, raising the early fusion MLP's performance to 0.57, they overall lacked the capacity to capture temporal dependencies critical for interpreting behavioral signals in depression detection.

Sequence-aware models based on Long Short-Term Memory (LSTM) networks demonstrated significantly stronger performance. All LSTM variants outperformed their non-sequential counterparts, with the attention-based fusion LSTM reaching a macro F1-score of 0.80. These results underscore the value of temporal

modeling for multimodal signals, particularly when coupled with adaptive fusion strategies.

Transformer-based models offered the highest potential, with the attention fusion Transformer achieving a macro F1-score of 0.83 and macro recall of 0.87—the best across all experiments. However, this performance was not uniform across all Transformer variants. For instance, the cross-modal Transformer achieved only moderate results, on par with baseline models. These disparities highlight the challenge of training large-capacity architectures in data-limited regimes. Given their reliance on large-scale data and higher model complexity, Transformers are more susceptible to overfitting and training instability when applied to modest-sized datasets such as MPDD.



(a) (b) (c)

Figure 3: Training and validation trajectories for (a) unregularized LSTM, (b) regularized attention-based Transformer, and (c) regularized intermediate fusion Transformer.

Initial experiments, particularly with complex models like LSTMs and Transformers, revealed signs of overfitting, characterized by a noticeable divergence between training and validation performance (Figure 3). Validation metrics often became erratic while training metrics continued to improve. Applying the comprehensive regularization techniques detailed in the methodology significantly mitigated this issue. Subsequent training runs demonstrated improved stability and closer alignment between training and validation curves, leading to better generalization. This stabilization was key to unlocking the full performance of models like the attention fusion Transformer, which ultimately achieved the highest macro F1-score.

We employed Focal Loss [18] to mitigate the effects of class imbalance present in the dataset. Its positive impact is evident when comparing the Early Fusion MLP model trained with Focal Loss to an identical architecture trained without it, as shown in Table 1. The model incorporating Focal Loss achieved higher scores on both metrics, with Macro Recall increasing to 0.56 (from 0.50) and Macro F1 reaching 0.57 (compared to 0.41 without Focal Loss).

In summary, while sequential modeling is essential for effective multimodal depression detection, careful architectural design and regularization are necessary to harness the full power of Transformer-based models—particularly in low-resource settings.

Conclusion

In this work, we investigated a range of multimodal architectures for binary depression detection using the MPDD dataset, spanning from simple non-sequential models to LSTMs and advanced Transformer-based systems. Our experiments showed that sequence-aware models substantially outperform static baselines, and that Transformer models, when properly regularized, achieve the best overall performance.

These results emphasize the importance of modeling temporal dynamics and incorporating adaptive fusion strategies. However, the limited dataset size remains a key constraint, especially for training high-capacity architectures like Transformers.

Future work can explore data-efficient Transformer variants, domain-informed priors, and interpretability techniques to enhance clinical applicability. Our study establishes a strong foundation for scalable and effective multimodal approaches to mental health assessment.

Contributions

This project was a collaborative effort among all three authors, with each member contributing to different stages of model development, experimentation, and reporting.

- **Aryamaan Saha** led the implementation and evaluation of non-sequential models, including multilayer perceptrons (MLPs), support vector machines (SVMs), and Compact Bilinear Pooling architectures. He also contributed significantly to data preprocessing.
- **Niranjan Sundararajan** developed the LSTM-based architectures and conducted early experiments with Transformer models. He also co-managed the cloud infrastructure setup and ensured reproducibility across environments.
- **Kaushal Damania** designed and implemented the advanced Transformer-based architectures, including attention-based and mixture-of-experts models. He also contributed to data preprocessing and shared responsibility for setting up the experimental infrastructure.
- All three authors contributed to writing and refining the final report, ensuring clarity, rigor, and alignment with the experimental results.

Novelty: Our work presents a comprehensive comparison of multimodal fusion strategies, ranging from simple static baselines to advanced Transformer-based fusion, for depression detection. We systematically analyze the effect of fusion technique, temporal modeling, and regularization on performance. To the best of our knowledge, this is one of the most complete architectural evaluations on the MPDD Challenge dataset, and it offers practical insights for building robust multimodal models in low-resource mental health screening scenarios. We also apply multiple regularization techniques observed in various literature to combat the class imbalance and limit data availability.

Code and Recorded Presentation

All code used in this project is available at https://github.com/aryamaansaha/MPDD-ACM-challenge

Our recorded presentation is available at this drive link

References

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017.

[4] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H. T. Kung, C. Dagli, and T. F. Quatieri, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 11–18, ACM, 2016.

[5] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," in *Proc. Interspeech*, pp. 1–5, 2019.

[6] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model," *arXiv preprint arXiv:2202.08210*, 2022.

[7] N. Agarwal, K. Milintsevich, L. Metivier, M. Rotharmel, G. Dias, and S. Dollfus, "Analyzing symptom-based depression level estimation through the prism of psychiatric expertise," in *Proceedings of the LREC-COLING 2024*, pp. 974–983, 2024.

[8] Anonymous, "Audibert: A deep transfer learning multimodal classification framework for depression screening," *Unpublished abstract*, 2024.

[9] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[10] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 2016.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[12] HaciLab, Tianjin University, "Mpdd challenge: Multimodal personality-aware depression detection," 2023. Accessed: 2025-05-05.

[13] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 317–326, 2016.

[14] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Machine Vision and Applications*, vol. 32, no. 121, 2021.

[15] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," *CoRR*, vol. abs/1511.06062, 2015.

[16] S. Liu, S. Yao, J. Li, D. Liu, T. Wang, H. Shao, and T. Abdelzaher, "Giobalfusion: A global attentional deep learning framework for multisensor information fusion," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, Mar. 2020.

[17] Y. Zou, S. Yi, Y. Li, and R. Li, "A closer look at the cls token for cross-domain few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13658–13667, IEEE, 2022.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.