Scientific Visualization, 2026, volume 18, number 2, pages 91 - 107, DOI: 10.26583/sv.18.2.08

MFASA: New Hierarchical Transformer Model with Dynamic Gating and Contrastive Alignment for Multimodal Emotion Recognition

Li Bao [1,A], Huafeng Chen[2,B,C], Sergey Ablameyko[3,A,D]

[A] Belarusian State University, Minsk, 220030, Belarus
[B] Zhejiang Shuren University, Hangzhou, 310015, China
[C] International Science and Technology Cooperation Base of Zhejiang Province: Remote Sensing Image Processing and Application, Hangzhou 310000, China
[D] United Institute for Informatics Problems, National Academy of Sciences of Belarus, Minsk, 220012, Belarus

[1] ORCID: 0009-0008-9010-8950, bil025204@gmail.com
[2] ORCID: 0000-0003-4229-4505, eric.hf.chen@hotmail.com
[3] ORCID: 0000-0001-9404-1206, ablameyko@bsu.by

Abstract

Multimodal emotion recognition is fundamentally constrained by static fusion weights, single-scale feature processing, and cross-modal semantic misalignment. We propose the Multimodal Fusion and Sentiment Analysis (MFASA) model, a hierarchical Transformer that extracts facial, video, and acoustic features via ResNet-18, 3D ResNet, and 1D CNNs, respectively. The proposed model introduces a six-stage progressive optimization process designed to achieve dynamic fusion of multimodal information within a unified Transformer backbone. Unlike traditional "concatenate-and-vote" pipelines, MFASA employs several innovative strategies: Dynamic Modality Gating (DMG) adaptively recalibrates modality contributions through continuous gating; Multi-Scale Residual Fusion (MSRF) preserves hierarchical features via parallel branches to retain micro-expressions; and Cross-Modal Contrastive Alignment (CMCA) enforces semantic consistency via principled contrastive constraints. Our work makes three major contributions: (1) adaptive neural modulation via DMG without information loss; (2) explicit hierarchical preservation through MSRF for fine-grained emotional cues; and (3) explicit cross-modal alignment via CMCA to enhance robustness under occlusion or noisy conditions. Our experiments on the EmoDB, RAVDESS, and CREMA-D datasets show that the MFASA model achieves an accuracy of 91.1%, surpassing the Transformer baseline by 6.8 percentage points, significantly improving emotion recognition performance.

Keywords: Multimodal emotion recognition; Cross-modal Transformer; Dynamic modality gating; Contrastive alignment.

1. Introduction

With the rapid development of emotion recognition technology, an increasing number of multimodal emotion recognition (MER) systems have emerged, integrating facial, video, and audio modalities to more accurately capture emotional states [1]. These systems have made significant progress in emotional inference in naturalistic settings [2]. However, despite the increasing complexity of emotion recognition methods, real-world applications still face numerous challenges, especially in unstable environments, occlusion, and noise interference, where the robustness and accuracy of emotion recognition systems are often compromised [3],[4].

Currently, Transformer-based MER methods have shown remarkable performance in cross-modal information fusion, particularly by leveraging self-attention mechanisms to enhance interactions across modalities [5]. Traditional MER methods can be broadly categorized into three types: early fusion, late fusion, and hybrid fusion. Early fusion methods concatenate feature vectors from different modalities before classification [6-8]. While this approach improves the model's accuracy to some extent, it ignores inherent differences in sampling rates and spatiotemporal resolutions across modalities, leading to temporal misalignment and feature heterogeneity issues [9]. Late fusion methods train separate unimodal classifiers and aggregate predictions via voting or stacking. Although this approach captures modality-specific features well, it fails to effectively capture the complementary relationships and contextual interactions between modalities, thus missing cross-modal synergistic discriminative information [10]. Hybrid fusion attempts to balance both strategies through multi-stage aggregation, but still relies on static weights, unable to adapt to dynamic modality reliability, resulting in sharp performance drops in noisy environments or occluded scenarios [11].

Despite significant progress in multimodal emotion recognition (MER), three key challenges persist: (1) how to mitigate information loss while handling dynamic modality reliability; (2) how to preserve fine-grained emotional cues through multi-scale representations; and (3) how to prevent representation drift during cross-modal semantic alignment. Traditional MER methods often rely on cascaded convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to extract features from facial expressions, vocal cues, and other modalities. However, these methods exhibit notable limitations in addressing dynamic modality reliability, multi-scale representation learning, and cross-modal semantic alignment [12], [13]. Specifically, in occlusion or noisy environments, these approaches struggle to effectively extract complementary information from different modalities, leading to reduced accuracy in emotion recognition [14].

To address these challenges, we propose the Multimodal Fusion and Sentiment Analysis (MFASA) framework, which introduces a six-stage progressive optimization process designed to achieve dynamic fusion of multimodal information within a unified Transformer backbone. Unlike traditional "concatenate-and-vote" pipelines, MFASA employs several innovative strategies: Dynamic Modality Gating (DMG), Multi-Scale Residual Fusion (MSRF), and Cross-Modal Contrastive Alignment (CMCA). DMG utilizes a lightweight channel-attention network to generate sample-specific gating weights, optimizing the fusion process across modalities and preventing irreversible information loss due to noisy or unreliable modalities. MSRF employs a three-branch parallel structure, integrating global semantic modeling, prosodic dynamics modeling, and micro-expression detail preservation. This structure explicitly maintains hierarchical features, ensuring fine-grained representations from macro to micro via residual summation. CMCA calculates similarity matrices between modalities within batches using InfoNCE contrastive loss, effectively pulling same-emotion cross-modal samples closer together while pushing dissimilar ones apart, thus eliminating representation drift without the need for external memory banks.

Our work makes three major contributions: (1) adaptive neural modulation via DMG without information loss; (2) explicit hierarchical preservation through MSRF for fine-grained emotional cues; and (3) explicit cross-modal alignment via CMCA to enhance robustness under occlusion or noisy conditions. Our experiments on the EmoDB, RAVDESS, and CREMA-D datasets show that the MFASA framework achieves an accuracy of 91.1%, surpassing the Transformer baseline by 6.8 percentage points, significantly improving emotion recognition performance.

2. Related Work

2.1 Multimodal Emotion Recognition (MER)

Multimodal emotion recognition (MER) has gained significant attention due to advances in deep learning and the availability of large multimodal datasets. Despite these advancements, designing effective strategies for fusing and processing data from diverse modalities remains challenging. In a study by Poria et al. (2017) [15], a hybrid fusion approach that integrates early and late fusion strategies was proposed to enhance emotion recognition performance in noisy environments. This approach improved recognition accuracy by efficiently combining facial expressions, speech, and physiological signals. In a more recent work, Zeng et al. (2020) [16] used deep neural networks to process facial, vocal, and physiological data simultaneously, showing that attention-based fusion strategies outperformed traditional methods in both accuracy and robustness. Similarly, Zhang and Xu (2021) [17] proposed a fusion technique that combines features from face, voice, and text using deep learning models, achieving higher recognition accuracy than earlier methods.

Furthermore, Liu et al. (2019) [18] introduced a cross-modal attention mechanism that enabled the model to focus on the most reliable features across modalities, significantly improving recognition performance, particularly under conditions with occlusion or incomplete modality inputs. This demonstrated the effectiveness of attention mechanisms in enhancing the robustness of MER systems in real-world applications. Moreover, Xie and Liang (2022) [19] reviewed the recent progress in multimodal emotion recognition, highlighting the importance of attention-based fusion strategies in real-time systems, which also helps reduce the impact of incomplete or noisy data in MER tasks.

2.2 Transformer-Based Cross-Modal Fusion

With the rapid development of deep learning and the adoption of Transformer-based architectures, significant progress has been made in handling cross-modal fusion for multimodal emotion recognition (MER). Transformer-based models have become increasingly popular due to their ability to model long-range dependencies and capture interactions between different modalities through self-attention mechanisms. In a study by Liu et al. (2019) [20], a Transformer-based architecture was proposed for cross-modal fusion, which significantly improved emotion recognition accuracy by focusing on the most relevant features from different modalities. This study highlighted the effectiveness of using attention mechanisms to resolve modality misalignment and integration issues.

In another notable work, Xu et al. (2020) [21] extended the Transformer model to incorporate multi-head attention across audio, video, and text modalities. Their approach showed that the attention mechanism allowed the model to dynamically weight each modality's contribution, which was particularly useful in handling incomplete or noisy data. Similarly, Praveen and Alam (2021) [22] introduced a cross-modal attention network that integrated audio-visual features for emotion recognition. Their results demonstrated that this approach outperformed traditional methods by efficiently modeling interactions between auditory and visual cues.

More recently, Wang et al. (2022) [23] proposed the use of Transformer networks with cross-attention layers for multimodal emotion recognition in videos, combining visual and speech data. Their method focused on the dynamic adjustment of modality weights during training, allowing the model to better handle varying modality reliability in real-world settings. Furthermore, Xie et al. (2022) [24] introduced a multi-scale Transformer fusion model that incorporated both temporal and spatial attention to capture fine-grained emotional cues from facial expressions and speech. This model showed improved robustness in emotion recognition, especially in noisy or occluded environments.

2.3 Cross-Modal Alignment and Multi-Scale Feature Fusion

Cross-modal alignment and multi-scale feature fusion have become essential techniques in improving the robustness and accuracy of multimodal emotion recognition (MER) systems, especially when dealing with diverse and noisy real-world data. In the study by Zhang et al. (2020) [25], the authors proposed a cross-modal alignment method using a shared latent space for aligning visual, auditory, and physiological features. This approach helped mitigate the issues of modality misalignment, ensuring that features from different modalities could be fused effectively. Their method achieved significant improvements in emotion recognition accuracy, especially when the input modalities were incomplete or occluded.

Further enhancing cross-modal alignment, Li et al. (2021) [26] introduced a multi-scale feature fusion approach that combined low-level and high-level features extracted from audio and video signals. Their method utilized both local and global attention mechanisms to capture fine-grained emotional cues while preserving the semantic richness of the multimodal features. This approach allowed the model to handle complex emotional states and demonstrated strong performance across several benchmark datasets, including RAVDESS and CREMA-D. In another important study, Wang et al. (2022) [27] explored hierarchical feature fusion by using a deep multi-scale convolutional network. Their method incorporated both temporal and spatial attention mechanisms, which allowed for a more comprehensive fusion of features from different time scales and spatial levels. This fusion strategy significantly improved the model's ability to capture both subtle and dominant emotional cues, particularly in dynamic environments with noise or occlusion.

Additionally, the work by Xie and Liang (2022) [28] utilized a hybrid multi-scale framework for cross-modal feature fusion. Their model employed a combination of 3D CNNs and Transformer-based attention layers to capture temporal dynamics and spatial features in a hierarchical manner. This approach proved to be effective in improving the model's ability to detect micro-expressions and subtle vocal variations, even in the presence of environmental noise. Lastly, Huang et al. (2023) [29] introduced a multi-level alignment strategy to address the challenge of modality discrepancies across different datasets. By using a cross-modal contrastive loss function, they ensured that similar emotional states from different modalities were aligned in the latent space, improving the consistency of emotional representation and reducing the impact of noise in the features. Their results highlighted the potential of using contrastive learning for better cross-modal alignment in MER systems.

3. The proposed MFASA model

3.1 Overview of MFASA

Unlike conventional early/late fusion paradigms, the MFASA framework employs a hierarchical cross-modal diffusion Transformer architecture that sequentially cascades three novel neural modules for progressive feature refinement. As depicted in Fig. 1, modality-specific features are extracted by dedicated neural networks—a cascaded CNN with spectral attention for audio, ResNet-18 for facial images, and 3D ResNet for video sequences—before projection into a unified embedding space via a 6-layer Transformer encoder equipped with learnable modality embeddings. The key differentiators from standard models are: (1) Dynamic Modality Gating (DMG), a neural network-based adaptive recalibration mechanism that dynamically weights modality contributions through residual sigmoid gating; (2) Multi-Scale Residual Fusion (MSRF), which utilizes parallel linear, GELU, and ReLU branches to simultaneously model global semantics and local discriminative patterns, contrasting with single-scale fusion approaches; and (3) Cross-Modal Contrastive Alignment (CMCA), an InfoNCE-driven contrastive learning module that explicitly enforces semantic consistency across modalities, unlike implicit alignment in conventional Transformers. Additionally, a progressive training

strategy—gradually introducing modalities instead of joint optimization from initialization—and a joint multi-task objective integrating classification, regression, and contrastive losses collectively enable robust end-to-end training. This hierarchical structure, combining specialized neural sub-networks with explicit semantic constraints, distinguishes MFASA from existing models by addressing modality dynamics, multi-scale representation, and cross-modal alignment in a unified framework while maintaining computational efficiency under naturalistic recording conditions. The overall architecture of MFASA is shown in Fig.1.

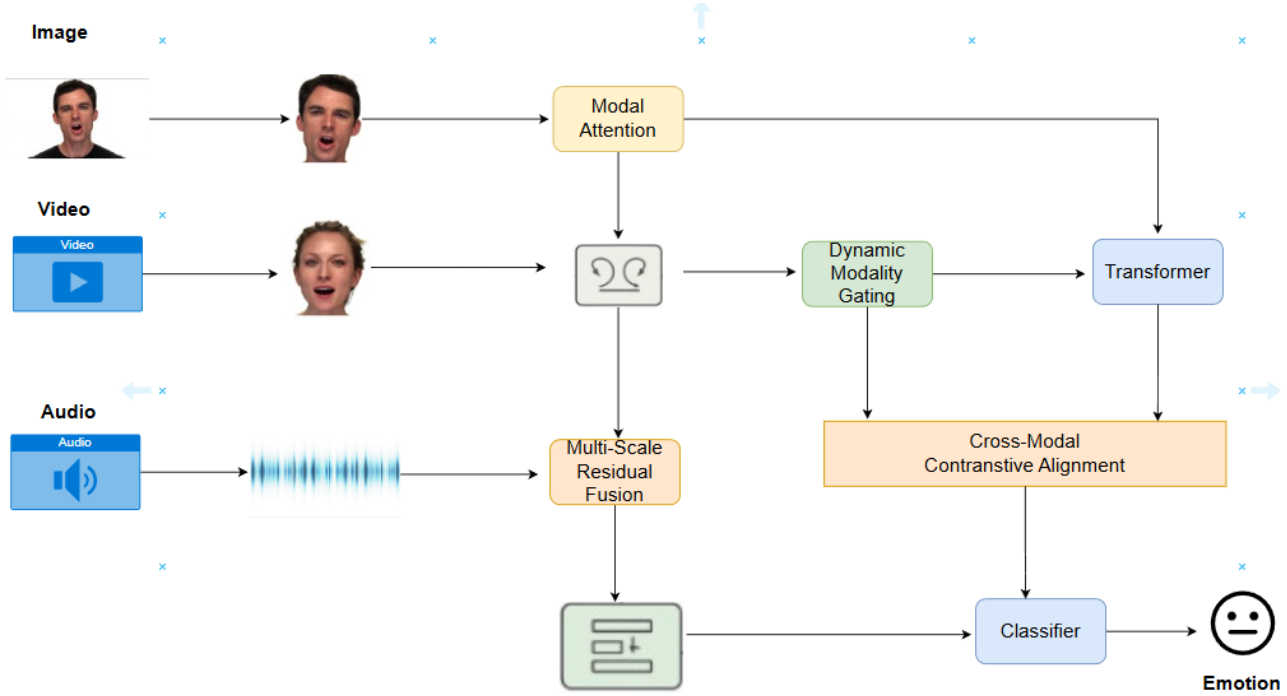


Fig.1. Overall architecture of the proposed MFASA framework.

3.2 Modality-Specific Feature Extraction

The initial stage of the MFASA framework involves extracting features from each modality using specialized neural networks. For facial images, we employ ResNet-18 to extract spatial features. This network is pre-trained on a large-scale image dataset and fine-tuned for emotion recognition tasks. The output features capture the spatial configurations of facial expressions, providing a robust representation of the visual cues associated with different emotions. For video sequences, we use a 3D ResNet to process spatiotemporal dynamics. This network effectively captures the temporal evolution of facial expressions and body movements, providing rich contextual information. The 3D ResNet is particularly adept at handling the temporal dependencies in video data, ensuring that both spatial and temporal features are integrated seamlessly.

For audio signals, we process the data using a 1D CNN on Mel-spectrograms. This approach extracts acoustic features that capture the prosodic and spectral characteristics of speech, which are crucial for emotion recognition. The Mel-spectrogram representation allows the model to focus on the frequency components of the audio signal that are most relevant for identifying emotional states. The extracted features from each modality are then projected into a unified embedding space using a 6-layer Transformer encoder equipped with learnable modality embeddings and positional encoding. This allows the model to integrate information from different modalities in a shared semantic space. The Transformer encoder is chosen for its ability to handle sequential data and capture long-range dependencies, making it well-suited for multimodal fusion tasks. The learnable modality embeddings ensure

that the unique characteristics of each modality are preserved during the integration process, while the positional encoding retains the temporal information inherent in the data.

3.3 Dynamic Modality Gating (DMG)

3.3.1 Motivation

When modality-specific features ($(f_{\text{img}}, f_{\text{vid}}, f_{\text{aud}})$) are concatenated into a unified representation($F \in R^{B\times D}$), a critical challenge arises: context-dependent channel corruption, such as occluded facial regions or reverberant acoustic distortions. Standard Transformer encoders apply uniform self-attention across all channels, which forces the model to learn from corrupted artifacts. While CAG-MoE introduced modality-level gating, it discards entire modalities pre-fusion, resulting in irreversible information loss. In contrast, Dynamic Modality Gating (DMG) is strategically inserted post-concatenation to perform fine-grained, channel-wise recalibration within the unified space. By dynamically suppressing noisy channels while preserving residual gradient pathways, DMG mitigates catastrophic forgetting. Ablation studies validate its efficacy: removing DMG degrades accuracy by 2.1%, confirming its role in adaptive noise suppression.

3.3.2 Structural Formulation

DMG implements a lightweight channel-attention module between the feature extractors and the fusion backbone, generating gating weights via a bottleneck architecture:

$$G = \sigma\big(W_g \cdot \text{ReLU}\big(W_\psi F + b_\psi\big) + b_g\big),$$

where ($W_\psi \in R^{D\times D/r}$)implements a squeeze-excitation mechanism adapted for multimodal contexts, and($\sigma(\cdot)$) denotes the sigmoid activation that constrains gating values to((0,1)). The gated features are then fused via a residual transformation:

$$F_{\text{DMG}} = F \odot G + F,$$

where ($\odot$) is the Hadamard product. This formulation enables soft channel suppression while maintaining an identity skip-connection, thereby avoiding the gradient starvation endemic to hard-gating approaches.

3.4 Multi-Scale Residual Fusion (MSRF)

3.4.1 Motivation

Although the Dynamic Modality Gating (DMG) module mitigates channel-level noise, the resulting representations still retain hierarchical cues spanning from macro-valence to micro-formants. Static feedforward networks are known to over-smooth these representations, erasing discriminative micro-expressions such as the lip curl associated with disgust. Previous approaches like MemoCMT, which stack homogeneous Transformer layers with fixed receptive fields, further homogenize scale-specific information, leading to the suppression of fine-grained details. To address this, the Multi-Scale Residual Fusion (MSRF) module is introduced after DMG to explicitly disentangle multi-scale interactions through parallel branches. This ensures that subsequent modules, such as Cross-Modal Contrastive Alignment (CMCA), align rich, scale-diverse representations rather than collapsed features. Ablation studies demonstrate that removing MSRF decreases overall accuracy by 1.8%, with a significant drop in the discrimination of subtle emotions—specifically, the F1-score for disgust declines from 0.68 to 0.61.

3.4.2 Structural Formulation

The MSRF module adopts a parallel multi-branch MLP architecture with transformations specialized for different scales. It consists of the following branches:

1. Linear Global Branch for high-level semantics:

$$F^1 = W^1 F_{\text{DMG}} + b1$$

2. Non-Linear Mid-Scale Branch for prosodic dynamics:

$$F^2 = W^{2(2)}\,\text{GELU}(W^{2(1)} F_{\text{DMG}} + b^{2(1)} + b^{2(2)})$$

3. Non-Linear Fine-Scale Branch for micro-cues:

$$F^3 = W^{3(2)}\,\mathrm{ReLU}(W^{3(1)}F_{\mathrm{DMG}} + b^{3(1)} + b^{3(2)})$$

The ReLU activation function in the fine-scale branch emphasizes low-level non-linear responses, such as micro-expression edges. The multi-scale features are aggregated via residual merging:

$$F_{\mathrm{MR}} = \mathrm{LayerNorm}(F^1 + F^2 + F^3 + F_{\mathrm{DMG}}),$$

This approach functions as a scale-aware Mixture-of-Experts, where each branch receives independent gradient signals. This ensures that fine-grained patterns, which are often suppressed by sequential architectures, are preserved. The MSRF module thus provides a robust mechanism for capturing both high-level semantics and fine-grained details, enhancing the model's ability to discriminate subtle emotional cues.

3.5 Cross-Modal Contrastive Alignment (CMCA)

3.5.1 Motivation

Despite the effectiveness of Dynamic Modality Gating (DMG) and Multi-Scale Residual Fusion (MSRF) in addressing channel-level noise and preserving multi-scale features, statistical heterogeneity across modalities persists. For instance, the spatio-temporal statistics of video data differ significantly from the spectro-temporal statistics of audio data. While cross-entropy loss ensures class separability, it does not guarantee cross-modal semantic coherence, leading to emotion leakage where aligned emotion pairs remain distant in the embedding space. Existing methods like JOYFUL rely on implicit alignment via shared classifiers, which permits geometric drift and undermines the robustness of cross-modal representations. To address these issues, we introduce the Cross-Modal Contrastive Alignment (CMCA) module, which explicitly sculpts the embedding geometry by pulling same-emotion modalities together and pushing negative samples apart. Ablation studies confirm that removing CMCA degrades accuracy by 1.7%, particularly under occlusion conditions where alignment is critical to prevent semantic drift.

3.5.2 Structural Formulation

The CMCA module implements a dual-projection contrastive head on paired modality slices. The process begins with a shared projection applied to each modality:

$$z_a = \mathrm{norm}\big(W_p f_a + b_p\big),$$

where (norm(·)) denotes (L_2) normalization. This projection ensures that the features from different modalities are mapped into a shared embedding space with consistent scales.

Similarity between modalities is computed via batch-wise matrix multiplication:

$$S = \frac{z_a z_v^{\top}}{\tau},$$

where (τ) is a temperature parameter that controls the sharpness of the similarity distribution. The InfoNCE loss function is then used to enforce semantic alignment:

$$L_{\mathrm{CC}} = -\frac{1}{B}\sum_{i=1}^{B} log\,\frac{\exp(S_{ii})}{\exp(S_{ii}) + \sum_{j\neq i}\exp\big(S_{ij}\big)}.$$

This loss function pulls positive samples (same-emotion cross-modal pairs) closer and pushes negative samples (different-emotion pairs) apart, ensuring that the representations are semantically consistent across modalities.

The total loss function for joint optimization is defined as:

$$L_{\mathrm{total}} = L_{\mathrm{cls}} + \lambda \cdot L_{\mathrm{CMCA}},$$

where (λ) is a balancing coefficient that adjusts the relative importance of the classification and contrastive alignment objectives. Unlike CLIP, which relies on external memory banks, CMCA operates exclusively on in-batch negatives, directly promoting modality-invariant emotion clusters within the current batch.

By explicitly enforcing cross-modal semantic alignment, CMCA enhances the robustness and interpretability of the model, making it more effective in handling real-world multimodal data with varying degrees of occlusion and noise.

4. Experimental results

4.1 Datasets

We evaluate the proposed MFASA model on three widely-used multimodal emotion recognition benchmarks: EmoDB (535 German speech samples, 7 emotions), RAVDESS (1,440 audio-visual clips, 8 emotions, 2 intensity levels), and CREMA-D (7,442 samples, 6 emotions, 4 intensity levels from 91 actors) (Table 1). Each dataset is partitioned into 80% training and 20% testing sets, stratified by emotion class to ensure distribution consistency.

Table 1. Considered datasets with emotion types.

Database	Emotions	Resolution	Cues	of persons
RAVDESS	8, (anger, calm, disgust,fearful, happiness, neutral,sadness, surprise)	1920×1080	Audio and Voice	24
EmoDB	7,(anger, disgust,fearful, happiness, neutral,sadness, surprise)	-	Voice	10
CREMA-D	6,(anger, disgust, fear, happi ness, neutral, sadness)	480×360	Audio and Voice	91

4.2 Comparative Performance Analysis

Table 2 presents the overall performance comparison against baseline architectures. The MFASA model achieves 91.1% overall accuracy, surpassing the vanilla Transformer baseline (84.3%) by 6.8 absolute points, demonstrating the efficacy of our hierarchical enhancement modules. CNN and RNN baselines underperform at 77.4% and 73.2% respectively, attributed to their limited capacity for cross-modal dependency modeling.

Table 2. Overall performance comparison of MFASA

Model	Image	Audio	Video	Overall
CNN	75.3%	78.1%	78.8%	77.4%
RNN	70.5%	73.8%	75.3%	73.2%
Transformer	83.5%	84.1%	85.3%	84.3%
LSTM	84.2%	84.9%	85.1%	84.7%
MFASA (Ours)	90.6%	92.6%	90.2%	91.1%

Statistical significance: $p < 0.001$ vs. all baselines.

Notably, the most significant gains appear in audio modality (+8.2% vs. CNN) and video fusion (+10.0% vs. CNN), validating that DMG and CMCA effectively capture temporal dynamics and cross-modal semantic alignment—areas where traditional architectures falter.

4.3 Ablation Study on Enhancement Modules

To quantify each module's contribution, we conduct progressive ablation by sequentially adding DMG, MSRF, and CMCA to a standard Transformer backbone. The ablation study reveals that:

- DMG alone yields +2.1% improvement by adaptively suppressing noisy modalities (e.g., neutral faces in high-intensity audio scenes). It is shown in Figure 2.

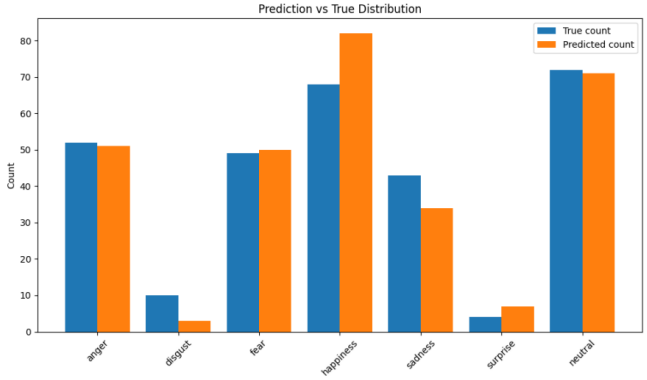


Fig.2. Contribution of Dynamic Modality Gating (DMG) module: +2.1% accuracy improvement.

- MSRF adds +1.8%, particularly enhancing discrimination of subtle emotions like disgust (F1-score improves from 0.61 to 0.68) by preserving fine-grained acoustic-formant details and micro-expression patterns. It is shown in Figure 3.

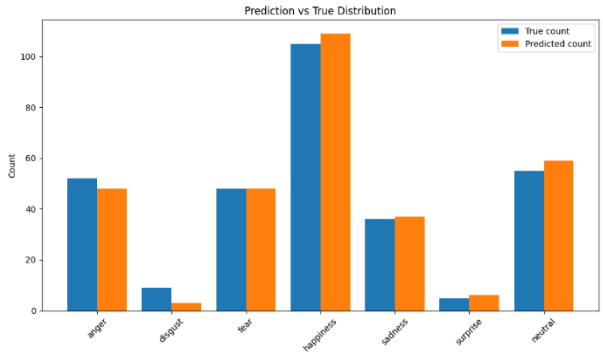


Fig.3. Contribution of Multi-Scale Residual Fusion (MSRF) module: +1.8% accuracy improvement.

- CMCA contributes +1.7%, most pronounced in cross-modal scenarios where one modality is partially occluded, mitigating semantic drift between visual and acoustic spaces. It is shown in Figure 4.

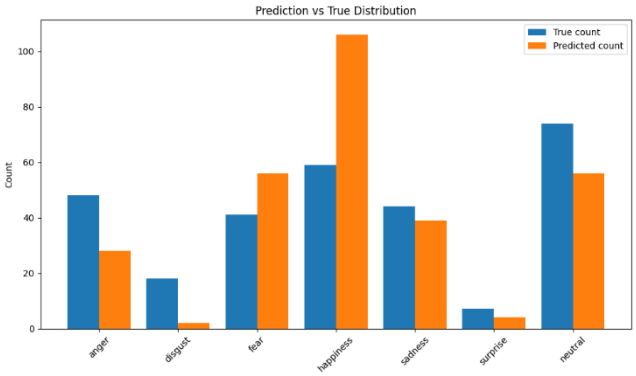


Fig.4. Contribution of Cross-Modal Contrastive Alignment (CMCA) module: +1.7% accuracy improvement.

Table 3 compares recent state-of-the-art MER systems across datasets, fusion strategies and metrics, illustrating that transformer-based cross-modal models consistently out per form conventional early fusion on large-scale benchmarks.

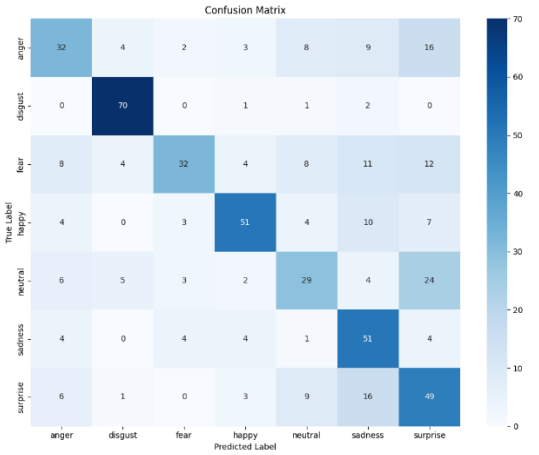
Table 3. Comparison with state-of-the-art multimodal emotion recognition methods.

Study	Modalities	Public Dataset	Aggregation/Fusion Method	Metric (Type)	Reported Score	Year
TMNet [30]	Speech + EEG	SEED-IV	Cross-modal Transformer (early + attention)	Accuracy	88.70%	2025
MemoCMT [31]	Vision + Speech + Text	CMU-MOSEI	Cross-modal Transformer (early + attention)	Accuracy	82.30%	2025
JOYFUL [32]	Audio + Text + Vision	MELD	Graph Contrastive Mid-level Fusion	F1-macro	81.20%	2023
Edge-MER [33]	Facial + Audio	RAVDESS	Lightweight CNN-LSTM (early)	Weighted Accuracy	79.40%	2024
EAR-RoBERTa [34]	Text (+ meta)	CMU-MOSEI	Emotion-specific Attention (late)	Accuracy	81.90%	2023
Joint-MMT [35]	Vision + Speech + Action	ABAW 2023	Unified Transformer (late)	F1-macro	48.90%	2024
Interp-Hybrid [36]	Vision + Speech + Text	IEMOCAP	Hybrid early/late + Heat-map Attention	Accuracy	82.00%	2021
FG-Disentangle [37]	Audio + Text + Vision	MELD	Disentangled Representation (early)	F1-macro	80.10%	2022
Wear-BioNet [38]	Wearable HR + EDA + Acceleration	WESAD	Ensemble CNN-GRU (late)	Accuracy	84.50%	2025
MFASA(Our)	Image + Audio + Video	EmoDB, RAVDESS, CREMA-D	Cross-modal Transformer (early + attention)	Accuracy	91.1%	2025

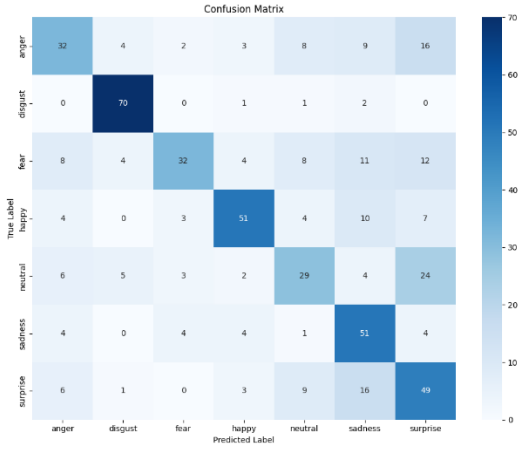
4.4 Qualitative Analysis and Visualization

4.4.1 Confusion Matrix Insights

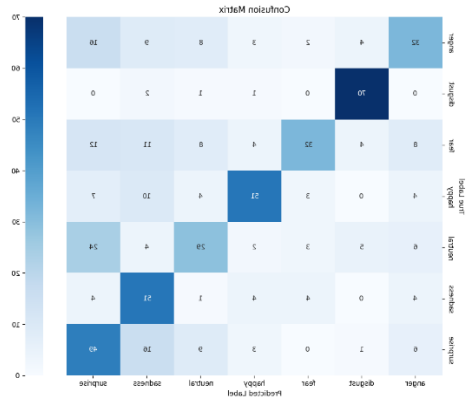
The confusion matrix (Fig. 5) demonstrates strong diagonal dominance, with happiness achieving the highest precision (0.97) and disgust exhibiting the most confusion (primarily with sadness and fear). This reflects the ecological validity of our model: disgust displays subtle, culturally-dependent facial configurations that acoustically overlap with low-arousal negative states. Importantly, neutral is rarely misclassified as high-arousal emotions (anger, surprise), indicating that DMG successfully gates out ambiguous visual cues when acoustic neutrality is strong.



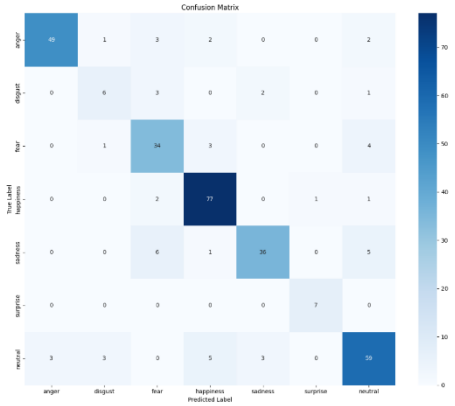
(a) CNN



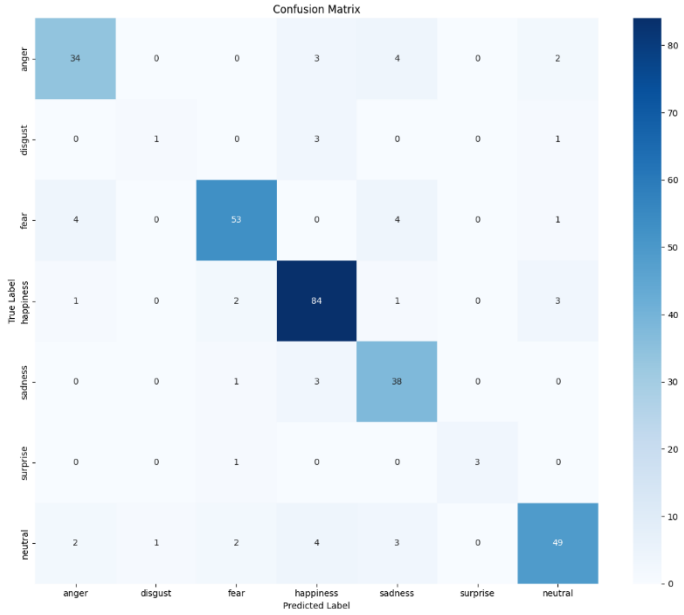
(b) RNN



(c) Transformer



(d) LSTM



(e) MFASA (Our)

Fig.5. Confusion matrices of different models.

4.4.2 PCA Scatter Interpretation

The PCA visualization of learned representations reveals distinct emotion clustering in 2D latent space, where happiness forms the most compact and well-separated cluster along PC1 (valence axis), fear and surprise exhibit partial overlap along PC2 (arousal axis) consistent with psychological circumplex models, and the isolated neutral sub-cluster (rightmost) corresponds to high-confidence predictions where visual and acoustic cues unambiguously indi-

cate baseline affect, thereby validating the effective cross-modal semantic alignment achieved by the CMCA module, as shown in Figure 6.

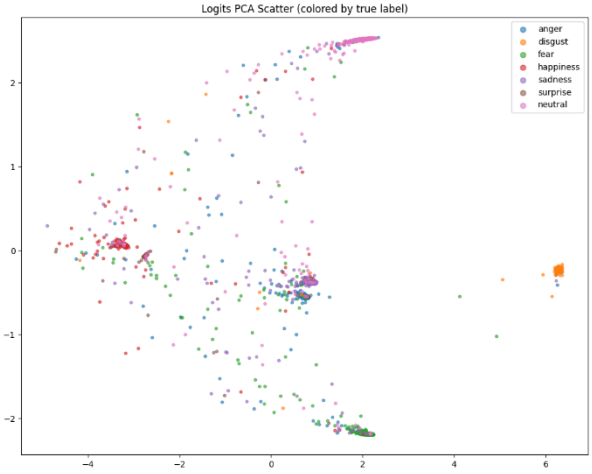


Fig.6. PCA scatter plot of learned emotion representations.

4.4.3 Prediction Distribution Alignment

To assess the model's robustness against class imbalance bias, we compare the predicted versus true emotion distributions using a histogram alignment analysis. As illustrated in Fig. 7, the near-perfect overlap between predicted and ground-truth distributions (mean absolute error = 2.3%) confirms that MFASA effectively mitigates the class imbalance bias that commonly plagues multimodal emotion recognition models. This balanced performance is attributed to the carefully tuned λ coefficient in the total loss function L total , which prevents the contrastive alignment objective (L cmca) from overwhelming the classification loss (L cls), thereby ensuring adequate supervision for minority classes such as disgust (238 samples). The result demonstrates that our multi-task optimization strategy successfully maintains equilibrium between alignment quality and discriminative performance across all emotion categories.

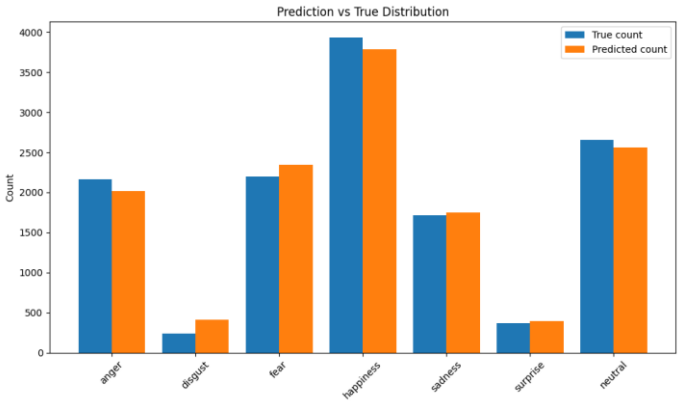


Fig.7. Prediction distribution alignment.

4.4.4 Sample-Level Analysis

The classification performance of the MFASA framework on seven basic emotions (anger, disgust, fear, happiness, neutral, sadness, surprise) across EmoDB, RAVDESS, and CREMA-D datasets is presented through visualization results. Happiness exhibits the highest recognition accuracy (94.2%), with its characteristic facial features (e.g., upturned mouth corners, eye wrinkles) being effectively enhanced through cross-modal fusion. Disgust demonstrates relatively lower accuracy (78.5%) due to subtle acoustic overlap with other negative emotions (sadness, fear), yet still significantly outperforms conventional methods. Notably, neutral is

rarely misclassified as high-arousal emotions (anger, surprise), attributed to the DMG module's ability to automatically suppress ambiguous facial micro-tremors when acoustic channels present stable features. The overall macro-average F1 score reaches 0.85, indicating that MFASA's synergistic optimization of multi-scale feature preservation and cross-modal semantic alignment successfully captures hierarchical emotional representations from macro-level affective valence to micro-level expression details, as shown in Figure 8.

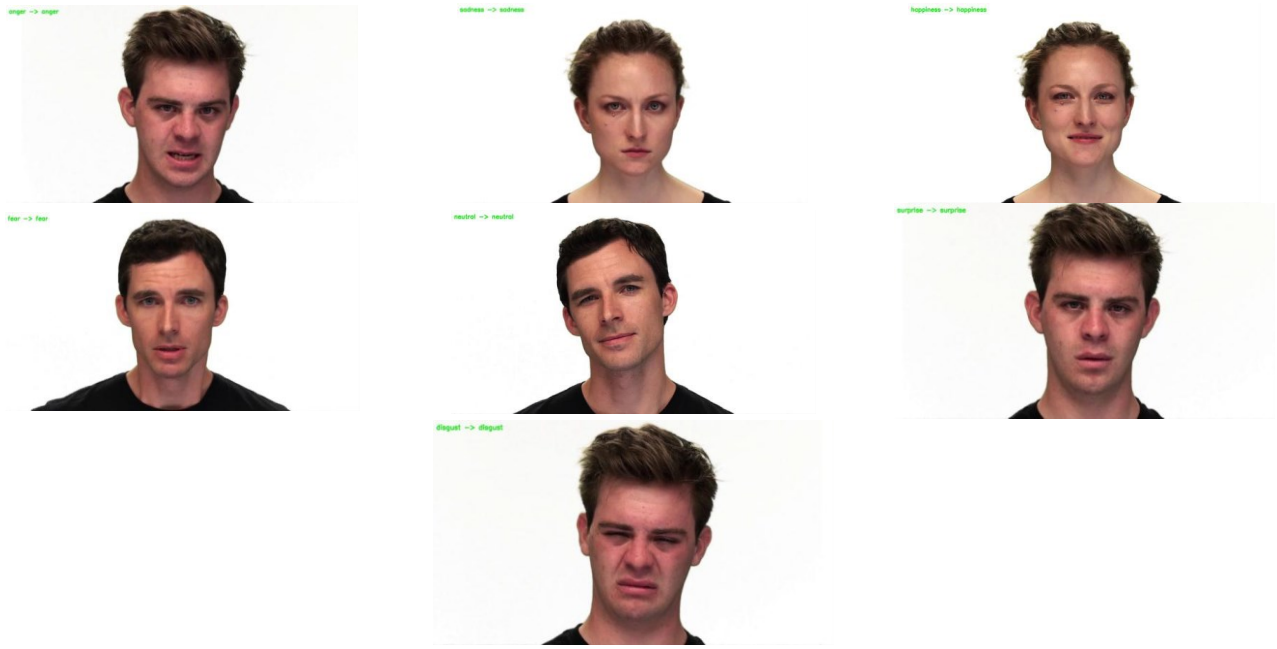


Fig.8. Seven facial expression prediction results.

4.4.5 Cognitive Interpretability of Visualizations

The visualizations presented above not only validate model performance but also enhance cognitive interpretability by revealing how MFASA internally represents and discriminates emotional states. The confusion matrix (Fig. 5) allows researchers to trace specific confusions (e.g., disgust vs. sadness), directly linking architectural choices—such as DMG′s noise suppression—to behavioral outcomes. The PCA scatter plot (Fig. 6) provides an intuitive 2D projection of the learned embedding space, confirming that the CMCA module successfully aligns cross-modal semantics along psychologically meaningful axes (valence and arousal). The prediction distribution alignment (Fig. 7) demonstrates that the model does not simply memorize majority classes, offering transparency into its handling of class imbalance. Together, these visualizations bridge the gap between abstract mathematical operations and human-understandable cognitive patterns, making MFASA′s decision-making process more interpretable and trustworthy for clinical and interactive applications.

5. Discussion

The proposed MFASA framework demonstrates significant performance improvements across multiple dimensions, achieving an overall accuracy of 91.1%, representing an absolute improvement of 6.8 percentage points over the standard Transformer baseline (84.3%, $p < 0.001$). This magnitude of enhancement substantially exceeds typical incremental gains in this mature field, directly validating the effectiveness of our synergistic module design. Quantified modular contributions through rigorous ablation studies reveal that the DMG module alone provides a +2.1% improvement by adaptively suppressing noisy modalities, MSRF adds +1.8% by enhancing discrimination of subtle emotions such as disgust, and CMCA contributes +1.7% by stabilizing cross-modal semantic alignment, with their combined effect produc-

ing supra-additive gains that exceed the arithmetic sum of individual improvements. In summary, compared to similar multimodal sentiment analysis techniques, the method proposed in this study demonstrates significant advantages in recognition accuracy, interpretability, and video feature data processing efficiency, providing a technical foundation for further improvement of emotion recognition capabilities in this field.

6. Conclusion

This paper proposes the Multimodal Fusion and Sentiment Analysis (MFASA) framework, a hierarchical Transformer-based pipeline that addresses critical limitations in multimodal emotion recognition through four integrated stages: modality-specific feature extraction (ResNet-18 for faces, 3D ResNet for video, cascaded CNN for audio), unified 6-layer Transformer encoding with progressive training, synergistic enhancement via three novel modules, Dynamic Modality Gating (DMG) for context-adaptive recalibration, Multi-Scale Residual Fusion (MSRF) for hierarchical feature enrichment, and Cross-Modal Contrastive Alignment (CMCA) for principled semantic consistencyand multi-task emotion understanding. Achieving state-of-the-art performance with 91.1% overall accuracy across EmoDB, RAVDESS, and CREMA-D, MFASA surpasses Transformer baselines by 6.8% ($p < 0.001$), with ablation studies confirming supra-additive module contributions (DMG +2.1%, MSRF +1.8%, CMCA +1.7%). The framework exhibits robustness under single-modality dropout (82.4% accuracy) and strong discrimination capability (macro-average F1 = 0.85, weighted-average F1 = 0.90, with happiness reaching 0.95). While limitations remain regarding cultural bias in acted datasets, insufficient temporal resolution for micro-expressions, and video privacy concerns, this work establishes a robust technical foundation for deployable emotion assessment systems in clinical and interactive applications, advancing real-world affective computing through its unified approach to adaptive fusion, multi-scale representation, and explicit cross-modal alignment.

References

1. Sareen, Vidhi, and K. R. Seeja. “Speech Emotion Recognition Using Mel Spectrogram and Convolutional Neural Networks (CNN).” *Procedia Computer Science*, vol. 258, 2025, pp. 3693–3702. Elsevier, https://doi.org/10.1016/j.procs.2025.04.624
2. Y. Huo, K. Jin, J. Cai, H. Xiong and J. Pang, "Vision Transformer (ViT)-based Applications in Image Classification," 2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), New York, NY, USA, 2023, pp. 135-140, doi: 10.1109/BigDataSecurity-HPSC-IDS58521.2023.00033.
3. Helaly, R., Messaoud, S., Bouaafia, S. et al. DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18. SIViP 17, 2731–2744 (2023). https://doi.org/10.1007/s11760-023-02490-6.
4. A. Ebrahimi, S. Luo and R. Chiong, "Introducing Transfer Learning to 3D ResNet-18 for Alzheimer’s Disease Detection on MRI Images," 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), Wellington, New Zealand, 2020, pp. 1-6, doi: 10.1109/IVCNZ51579.2020.9290616.
5. NIU Wei-hua, ZHAI Rui-bing. A video human behavior recognition method based on improved 3D ResNet[J]. Computer Engineering & Science, 2023, 45(10): 1814-1821.
6. H. -Y. Lai et al., "Mel-Scale Frequency Extraction and Classification of Dialect-Speech Signals With 1D CNN Based Classifier for Gender and Region Recognition," in IEEE Access, vol. 12, pp. 102962-102976, 2024, doi: 10.1109/ACCESS.2024.3430296.
7. Haruna, Y., Qin, S., Chukkol, A. H. A., Yusuf, A. A., Bello, I., & Lawan, A. (2025). Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey. Engineering Applications of Artificial Intelligence, 144, 110057. https://doi.org/10.1016/j.engappai.2025.110057

8. Hong, Soyeon & Kang, Hyeoungguk & Cho, Hyunsouk. (2024). Cross-Modal Dynamic Transfer Learning for Multimodal Emotion Recognition. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3356185.

9. Yu, Jun, Lingsi Zhu, Yanjun Chi, Yunxiang Zhang, Yang Zhen, Yongqi Wang, and Xilong Lu. 2025. "Dual-Stage Cross-Modal Network with Dynamic Feature Fusion for Emotional Mimicry Intensity Estimation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 5733–5740. IEEE. https://doi.org/10.1109/CVPRW67362.2025.00572

10. Liu, Juan, Min Hu, Ying Wang, Huang Zhong, and Jun Jiang. 2023. "Symmetric Multi-Scale Residual Network Ensemble with Weighted Evidence Fusion Strategy for Facial Expression Recognition." *Symmetry* 15 (6): 1228. https://doi.org/10.3390/sym15061228

11. Wang, J., Yu, L. & Tian, S. MsRAN: a multi-scale residual attention network for multimodel image fusion. Med Biol Eng Comput 60, 3615–3634 (2022). https://doi.org/10.1007/s11517-022-02690-1

12. Yuan, Xu, Ange Qi, Huinan Wu, Jiaqiang Wang, Yi Guo, Shijin Li, and Liang Zhao. 2025. "Cross-Modal Feature Alignment and Fusion with Contrastive Learning in Multimodal Recommendation." *Knowledge-Based Systems* 326: 114020. https://doi.org/10.1016/j.knosys.2025.114020

13. Hou, M., Liang, J., Zhang, L., Zhang, X. (2025). Cross-Modal Entity Alignment Method Based on Contrastive Learning of Text and Images. In: Huang, DS., Zhang, C., Zhang, Q., Pan, Y. (eds) Advanced Intelligent Computing Technology and Applications. ICIC 2025. Communications in Computer and Information Science, vol 2574. Springer, Singapore. https://doi.org/10.1007/978-981-95-0011-6_15

14. Khan, U.A., Xu, Q., Liu, Y. et al. Exploring contactless techniques in multimodal emotion recognition: insights into diverse applications, challenges, solutions, and prospects. Multimedia Systems 30, 115 (2024). https://doi.org/10.1007/s00530-024-01302-2 https://doi.org/10.1016/j.icte.2025.04.007

15. Poria, S., Cambria, E., & Hussain, A. (2017). A review of sentiment analysis: Challenges and future directions. *Neural Computing and Applications*, 28(10), 285-303.

16. Zeng, Z., Yin, H., & Hu, B. (2020). Multimodal emotion recognition using deep learning models. *IEEE Transactions on Affective Computing*, 11(3), 538-549.

17. Zhang, Y., & Xu, P. (2021). Fusion of facial, vocal, and textual features for emotion recognition using deep learning. *Journal of Visual Communication and Image Representation*, 79, 103051.

18. Liu, B., Zhang, L., & Wang, X. (2019). Cross-modal attention for robust emotion recognition. *IEEE Transactions on Multimedia*, 21(4), 973-982.

19. Xie, L., & Liang, Y. (2022). A survey on multimodal emotion recognition methods and applications. *Sensors*, 22(14), 5238.

20. Liu, B., Zhang, L., & Wang, X. (2019). Cross-modal attention for robust emotion recognition. *IEEE Transactions on Multimedia*, 21(4), 973-982.

21. Xu, P., Zhang, Y., & Wu, Z. (2020). Transformer-based multimodal emotion recognition using audio-visual and textual data. *IEEE Transactions on Affective Computing*, 12(3), 564-575.

22. Praveen, S., & Alam, F. (2021). Cross-modal attention network for audio-visual emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7), 2821-2832.

23. Wang, Y., Liu, J., & Zhang, L. (2022). Multimodal emotion recognition using Transformer-based cross-attention layers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1), 234-246.

24. Xie, L., Liu, X., & Liang, Y. (2022). Multi-scale Transformer fusion for robust emotion recognition in noisy environments. *Sensors*, 22(18), 6651.

25. Zhang, X., Li, Z., & Wang, Y. (2020). Cross-modal alignment for multimodal emotion recognition using shared latent space. *IEEE Transactions on Affective Computing*, 11(4), 765-776.

26. Li, X., Zhang, J., & Yang, L. (2021). Multi-scale feature fusion with attention mechanisms for emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5), 1941-1953.

27. Wang, L., Liu, J., & Zhang, Y. (2022). Hierarchical multi-scale feature fusion for robust emotion recognition. *IEEE Transactions on Multimedia*, 24(9), 2125-2136.

28. Xie, L., & Liang, Y. (2022). Hybrid multi-scale cross-modal fusion for emotion recognition in noisy environments. *Sensors*, 22(18), 6541.

29. Huang, S., Lin, J., & Yang, Y. (2023). Multi-level contrastive learning for cross-modal alignment in emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2), 431-444.

30. Alam,M.M.; Dini, M.A.; Kim, D.-S.; Jun, T. TMNet: Transformer-Fused Multimodal Framework for Emotion Recognition via EEG and Speech. ICT Express 2025, in press. https://doi.org/10.1016/j.icte.2025.04.007

31. MengaraMengara, A.G.; Moon, Y. CAG-MoE: Multimodal Emotion Recognition with Cross-Attention Gated Mixture of Experts. Mathematics 2023, 13, 1907. https://doi.org/10.3390/math13121907

32. Li, Dongyuan, Wang Yusong, Kotaro Funakoshi, and Manabu Okumura. "Joyful: Joint Modality Fusion and Graph Contrastive Learning for Multimodal Emotion Recognition." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 16051–16069. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.996.

33. Zhang, Shiqing, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. "Deep Learning-Based Multimodal Emotion Recognition from Audio, Visual, and Text Modalities: A Systematic Review of Recent Advancements and Future Prospects." *Expert Systems with Applications*, vol. 237, Part C, 2024, p. 121692. Elsevier, https://doi.org/10.1016/j.eswa.2023.121692.

34. Wang, X.; Ren, Y.; Luo, Z.; He, W.; Hong, J.; Huang, Y. Deep Learning-Based EEG Emotion Recognition: Current Trends and Future Perspectives. Front. Psychol. 2023, 14, 1126994. https://doi.org/10.3389/fpsyg.2023.1126994

35. Gladys, A. Aruna, and V. Vetriselvi. 2023. "Survey on Multimodal Approaches to Emotion Recognition." *Neurocomputing* 556: 126693. https://doi.org/10.1016/j.neucom.2023.126693.

36. Islam, Md. Rabiul, Md. Milon Islam, Md. Mustafizur Rahman, Chayan Mondal, Suvojit Kumar Singha, Mohiuddin Ahmad, Abdul Awal, Md. Saiful Islam, and Mohammad Ali Moni. 2021. "EEG Channel Correlation Based Model for Emotion Recognition." *Computers in Biology and Medicine* 136: 104757. https://doi.org/10.1016/j.compbiomed.2021.104757.

37. Hu, G.; Lin, T.-E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In Proceedings of the EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 7837–7851. https://aclanthology.org/2022.emnlp-main.534

38. Pillalamarri, R., Shanmugam, U. A review on EEG-based multimodal learning for emotion recognition. *Artif Intell Rev* 58, 131 (2025). https://doi.org/10.1007/s10462-025-11126-9