Journal of Computers Vol. 36 No. 2, April 2025, pp. 281-296
https://doi.org/10.63367/19911599202504360201 9

Multimodal Emotion Recognition Based on Hierarchical Feature Fusion

Ying-Gang Xie[1*], Nan-Nan Zhou[1], and Shi-Yin Zhu[1]

School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing, China
xieyinggang@bistu.edu.cn

Received 30 March 2025; Revised 9 April 2025; Accepted 16 April 2025

Abstract. Multimodal emotion recognition presents two major challenges: the limited capacity to model higher-order interactions among modalities, and the difficulty of achieving effective fusion due to imbalanced data quality across modalities. To address these issues, this paper proposes a novel model based on hierarchical feature fusion. The model adopts a three-level fusion framework. First, it integrates static fusion with a dynamic weighting mechanism informed by Bayesian uncertainty estimation to achieve initial alignment and importance modeling of modality-specific features. Second, a multi-head cross-modal attention mechanism is introduced to capture contextual dependencies and complementary information across modalities. Finally, gated recurrent units are employed to model temporal dynamics, thereby enhancing the semantic-level fusion representation. Experimental results demonstrate that the proposed method achieves 84.6% accuracy on binary classification tasks using the MOSEI dataset and a weighted F1 score of 69.7% on the IEMOCAP dataset—representing a 2.1% improvement over the representative baseline, COGMEN. Ablation studies further validate the essential contributions of the multi-head attention mechanism, dynamic weighting strategy, and gated fusion module to the overall performance gains.

Keywords: multimodal emotion recognition, hierarchical feature fusion, dynamic emotion modeling, bayesian uncertainty estimation, multi-head attention mechanism

1 Introduction

The widespread application of artificial intelligence technologies has driven the rapid growth of demand for emotion recognition in natural interaction scenarios. The core task of emotion recognition is to accurately determine an individual's emotional state by comprehensively analyzing multiple signal sources, such as speech, text, and video. Among these, multimodal emotion recognition, which integrates information from multiple modalities, is regarded as a key technology for achieving high-performance emotion recognition due to its ability to significantly enhance recognition accuracy [1]. However, the complexity of emotion recognition, such as the dynamic changes in emotional states, the cross-modal misalignment, and individual differences, poses significant challenges in the design of emotion recognition systems. In this context, multimodal information fusion has emerged as an effective strategy that, by exploiting the complementary information from multiple data sources, is expected to alleviate the aforementioned issues and significantly improve the accuracy and robustness of emotion recognition.

Existing multimodal emotion recognition methods often employ simple concatenation or static weight distribution strategies during feature fusion. While these methods enable basic integration of different modality features, they overlook the potential dynamic interactions between modalities, making it difficult to fully capture the complex emotional features. Furthermore, multimodal data often exhibit significant differences in quality, with certain modalities potentially being affected by noise or data missingness, thereby weakening the contribution of specific modalities to the emotion recognition task. In such cases, the model needs the ability to dynamically adjust the modality weights to accommodate inputs with varying data quality. However, current research in this area remains insufficient, causing existing methods to be vulnerable to data noise and modality missingness, resulting in inadequate robustness in practical applications.

* Corresponding Author

To address these issues, this paper proposes a multimodal emotion recognition model based on Hierarchical Feature Fusion (HFF). The model employs a progressive feature fusion strategy, gradually extracting higher-order interaction information between modalities from low-level features [2], thereby enabling the model to capture more complex and expressive emotional patterns. In addition, the model incorporates a Bayesian uncertainty-based dynamic weighting mechanism that adjusts modality contributions based on confidence estimates, significantly enhancing robustness under conditions of imbalanced data quality. To further improve the expressive power and interaction efficiency of modality features, we introduce a multihead cross-modal attention mechanism and a gated adaptive fusion module. This architecture not only strengthens the model's focus on salient modality information but also provides a more flexible and adaptive framework for multimodal feature fusion, ultimately improving both recognition accuracy and system stability.

The main contributions of this paper are summarized as follows:

(1) Proposed Hierarchical Feature Fusion Model: We designed a primary modality alignment module based on a dynamic weighting mechanism, an interaction learning module based on multi-head cross-modal attention, and a higher-order semantic fusion module combining a gated mechanism, which progressively explores higher-order interaction characteristics between modalities, significantly enhancing emotion recognition performance.

(2) Introduced Uncertainty Assessment Mechanism: We employ a Bayesian neural network to dynamically assess the confidence of modality features and adjust the modality weight distribution based on confidence, effectively addressing modality data noise and missingness issues, thereby enhancing the model's robustness.

(3) Validated Method Effectiveness on IEMOCAP and MOSEI Datasets: Through systematic experiments, we validated the superior performance of the method in emotion classification tasks, and ablation experiments confirmed the rationality of the model design and the contributions of key modules.

2 Related Work

2.1 Multimodal Feature Fusion Methods

Research in multimodal emotion recognition primarily focuses on feature extraction [3] and feature fusion [4]. As the core process determining the ability to capture interaction information between modalities, feature fusion methods have evolved from static to dynamic fusion, with recent trends in deep fusion, multi-level interactions, and cross-modal modeling.

Static fusion methods primarily use rule-driven strategies such as feature-level concatenation and weighted averaging, which offer advantages in simplicity and computational efficiency [5]. However, since these methods cannot dynamically adjust the fusion weights based on the semantic or contextual relevance between modalities, their performance is often limited when facing significant heterogeneity or nonlinear interactions between modalities. With the widespread use of deep learning technologies, more and more researchers have turned to learning-based dynamic fusion methods to enhance the model's ability to capture complex associations between modalities and improve generalization.

Dynamic fusion mechanisms often incorporate attention mechanisms [6] to achieve adaptive weight distribution between modalities. Chen et al. proposed a staged fusion strategy, using cross-modal dynamic correlation coefficients to integrate high-level local features and enhance semantic consistency [7]. Hu et al. modeled the dynamic evolution of context across multiple semantic spaces, effectively enhancing the complementarity between modalities [8]. In the task of fusion between facial and speech features, attention weight calculation strategies have been proven to optimize the alignment and interaction of multimodal features, improving fusion performance [9]. For the fusion of acoustic and text modalities, a cross-attention mechanism combined with deep neural network architecture further achieves more refined feature extraction and semantic enhancement [10]. In three-modal fusion tasks, a bidirectional multi-head cross-attention architecture effectively mitigated the impact of information redundancy or missingness between modalities by constructing multi-granularity, multi-directional mappings between modalities [11]. Furthermore, to further improve the modeling of complex cross-modal dependencies, some studies have introduced gating mechanisms (such as group gating [12]) to selectively enhance and dynamically control the contribution of different modalities.

Although the aforementioned fusion strategies have achieved significant results in modeling first-order modality interaction relationships, they still fall short when it comes to modeling higher-order structural relationships [13]. In recent years, graph neural networks (GNNs) have gradually become important tools for handling complex structural and long-range dependency relationships between modalities [14]. For instance, Shirian et

al. transformed speech emotion recognition tasks into a graph classification problem and used graph convolution operations to model both intra-modal and cross-modal structural information, thereby strengthening feature associations [15]; Liu et al. designed a non-Euclidean space modeling framework based on Graph Isomorphism Networks (GIN) and demonstrated the potential of graph structures to maintain global consistency in emotion representations [16]. However, despite the advantages of graph models in terms of expressiveness, issues such as high computational complexity and training instability still limit their practical application in large-scale systems.

In addition to fusion strategies, the temporal synchronization and feature density matching between modalities are also crucial factors influencing fusion performance. To address the multimodal alignment problem, some studies have introduced strategies such as sliding time windows, timestamp reconstruction, and feature re-projection to alleviate information misalignment between modalities. For example, by introducing temporal alignment networks or attention alignment modules, dynamic mapping between modalities with different temporal resolutions can be achieved, thereby enhancing consistency and accuracy during the feature fusion process. However, in extreme scenarios involving modality missingness or significant delays, such alignment mechanisms still face challenges in generalization.

It is worth noting that current multimodal fusion methods often focus on local interaction modeling and lack a systematic, hierarchical fusion mechanism. In this context, hierarchical fusion architectures have gradually attracted the attention of researchers. These methods integrate multimodal features at different granularities and semantic levels, not only enhancing the expression capability of high-dimensional semantic information but also providing more stable and discriminative fusion representations for subsequent context modeling and classification tasks.

2.2 Multimodal Uncertainty Modeling and Evaluation

In multimodal emotion recognition tasks, different modalities are often influenced by noise, information loss, and signal delay during data collection, leading to significant differences in modality quality. This imbalance not only interferes with the effectiveness of feature fusion but also reduces the overall robustness and reliability of the model. Therefore, modeling and leveraging modality uncertainty during fusion has become a key issue for improving recognition accuracy and credibility.

Early research primarily focused on enhancing the robustness of multimodal fusion strategies, yet the concept of uncertainty was not explicitly introduced for modeling the contributions of modalities, which made it difficult to address scenarios with large fluctuations in modality quality. With the development of deep learning and Bayesian methods, uncertainty evaluation has gradually become an important approach to address issues of modality quality imbalance and model prediction confidence in multimodal emotion recognition [17]. The primary functions of uncertainty modeling can be summarized in two aspects: first, by quantifying the uncertainty of modality features, it effectively adjusts the influence weights of low-quality modalities, thereby enhancing the robustness of fusion results; second, by measuring the confidence level of the model's output, it assists in risk control, reducing the negative impact of incorrect predictions.

Bayesian Neural Networks (BNNs) serve as a representative method for modeling prediction uncertainty. By representing network weights as probability distributions, BNNs explicitly model uncertainty in the parameter space of the model [18]. BNNs can dynamically allocate modality weights during inference, allowing for adaptive adjustment of decision paths based on feature quality and prediction confidence, thereby improving the model's robustness in multimodal information fusion [19]. For instance, Tellamekala et al. introduced a modality-level uncertainty measurement mechanism, effectively quantifying the randomness and instability of modality performance in emotion prediction, providing a suppression mechanism for low-quality modalities [20]. Chen et al. further proposed a hierarchical uncertainty modeling framework, where, in the context of dialogue emotion recognition, adaptive noise perturbations were introduced to adjust context attention, modeling context-level uncertainty, while Bayesian Capsule Networks were used to model modality-level uncertainty, thus achieving multi-level uncertainty perception and fusion optimization [21].

However, although these methods theoretically enhance the model's ability to perceive uncertainty, their practical deployment still faces challenges such as high computational complexity and training instability. Especially in large-scale multimodal systems, the high demands of Bayesian Neural Networks on resources and inference time limit their widespread application. Moreover, current uncertainty modeling methods mainly focus on modality-level or context-level uncertainty estimation, and the ability to model collaborative uncertainty between different modalities remains insufficient, with a lack of fine-grained cross-modal uncertainty complementarity mechanisms.

Based on this, this paper proposes a hierarchical fusion framework that combines cross-modal interaction mechanisms with uncertainty weighting strategies, building upon previous work. This approach not only retains the advantages of Bayesian methods in expressing model confidence but also reduces model complexity through structural fusion mechanisms, thereby improving deployability while ensuring performance.

3 Proposed Method

The HFF method proposed in this chapter includes the single-modality feature extraction module, hierarchical feature fusion module, cross-modal dynamic weight adjustment module, and classifier design module, as shown in Fig. 1. The model employs a progressively layered feature fusion strategy to fully explore the complex higher-order interactions between modalities, and dynamically adjusts weights to enhance robustness against uncertainty and the imbalance in data quality.

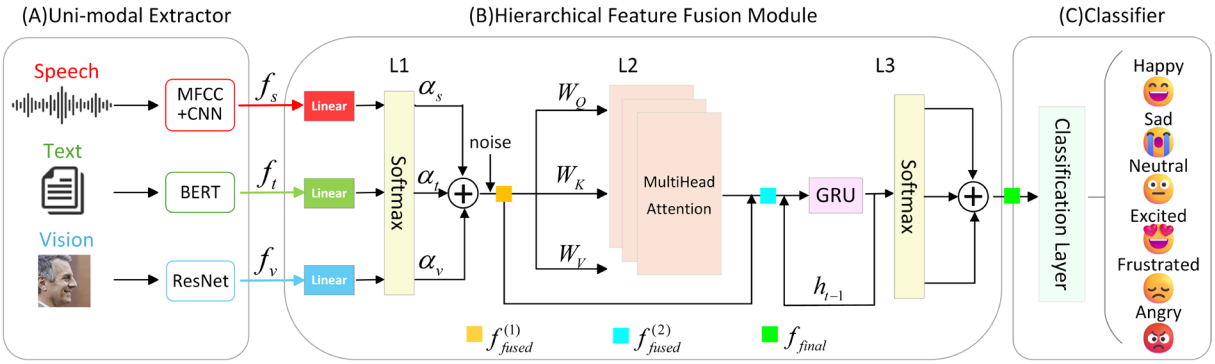

Fig. 1. Hierarchical Feature Fusion (HFF) framework

3.1 Uni-modal Feature Extraction Module

To ensure the quality and adaptability of input features, different feature extraction methods are applied to the speech (Speech), text (Text), and vision (Vision) modalities in this chapter.

(1) Speech Feature Extraction

For the speech modality, Mel-frequency cepstral coefficients (MFCCs) and deep convolutional neural networks (CNNs) are utilized to obtain effective and discriminative audio features. In multimodal emotion recognition tasks, speech signals are typically provided as time-domain waveforms. However, raw time-domain signals do not explicitly convey emotional characteristics. Therefore, it is necessary to convert them into the time-frequency domain to extract more informative representations. The Short-Time Fourier Transform (STFT) is applied to analyze the spectral content of the speech signal across sequential time windows. It is mathematically defined as:

$$X(t,f)=\sum_{n} x_s(n)w(n-t)e^{-j2\pi fn} \tag{1}$$

Where $x_s(n)$ denotes the original speech signal, $w(n)$ is the window function, and $X(t, f)$ represents the time-frequency spectrum. Although STFT provides detailed spectral information, the resulting high-dimensional representation may introduce excessive computational complexity. To better align with the nonlinear characteristics of human auditory perception, a Mel-frequency filter bank is applied to transform the linear frequency spectrum into the Mel scale. The conversion is computed as:

$$M(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

Where f is the linear frequency and $M(f)$ is the corresponding Mel frequency. After weighting the time-frequency spectrum with the Mel filter bank, the Mel spectrogram energy is computed as:

$$S_m = \sum_k |X(t, f_k)|^2 \; H_m(f_k) \tag{3}$$

Where $H_m(f_k)$ denotes the filter response of the m-th Mel filter, and S_m represents the energy in the Mel spectrogram. While this representation captures perceptually relevant frequency information, it may still contain redundant data unsuitable for direct use in classification tasks. Therefore, MFCCs are computed to further refine the features. First, a logarithmic transformation is applied to the Mel spectrogram:

$$S_m^{'} = \log |\, S_m \,| \tag{4}$$

Then, a Discrete Cosine Transform (DCT) is applied to remove the correlation between features and improve their discriminative ability:

$$c_n = \sum_{m=1}^{M} S_m^{'} \cos\left((m - 0.5)\frac{n\pi}{M}\right) \tag{5}$$

Where c_n denotes the n-th coefficient and M is the number of Mel filters. These MFCC features serve as compact and robust inputs for downstream learning models. To extract higher-order features from the MFCCs, a Convolutional Neural Network (CNN) is employed. CNNs are particularly effective in capturing local temporal patterns in audio spectrograms. Given an MFCC feature matrix X_{mfcc}, the CNN computes feature maps as follows:

$$h_k^{(i)} = f\left(\sum_{i=1}^{C_{i-1}} W_i^{(k)} * h_{i-1}^{(k)} + b^{(k)}\right) \tag{6}$$

Where $h_k^{(i)}$ is the feature map of the k-th channel in the i-th layer, $W_i^{(k)}$ and $b^{(k)}$ denote the convolution kernel and bias, with a nonlinear activation function applied after the convolution. To reduce dimensionality and improve computational efficiency, pooling layers are introduced:

$$p_j = \max\left(h_j^{(k)}\right) \tag{7}$$

Where p_j denotes the pooled feature. Finally, the high-level features extracted by the CNN are passed through a fully connected layer to obtain a fixed-dimensional representation:

$$f_s = \mathrm{CNN}(X_{mfcc}) \tag{8}$$

Where f_s is the final speech feature vector. The entire process is visually summarized in Fig. 2, illustrating each stage of the speech feature extraction pipeline.

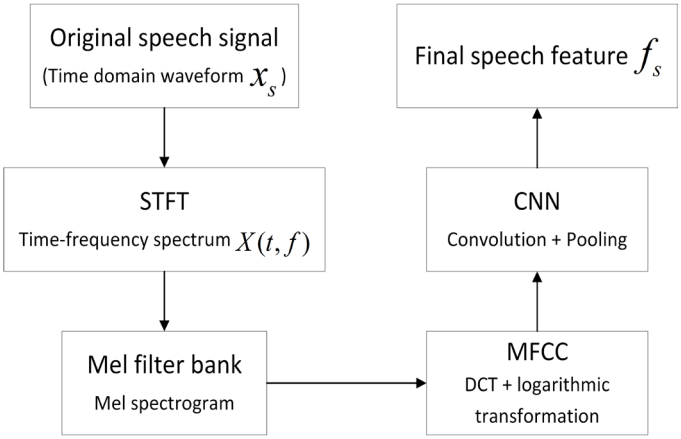

Fig. 2. Speech feature extraction process

(2) Text Feature Extraction

Text feature extraction is performed using the pre-trained BERT model [22], which is based on the Transformer architecture [23]. BERT utilizes a bidirectional self-attention mechanism to capture rich contextual dependencies within the text and generates word-level embedding vectors that effectively represent semantic information. Given an input text sequence $x = \{x_1, x_2, ..., x_n\}$, BERT computes a contextualized embedding vector for each token as follows:

$$h_i = \text{BERT}(x_i), \quad i \in \{1, 2, ..., n\} \tag{9}$$

Where h_i denotes the embedding vector corresponding to the i-th token. These vectors collectively form the word-level embedding representation for the text modality.

BERT's architecture consists of multiple stacked Transformer layers. Each layer comprises a multi-head self-attention mechanism followed by position-wise feedforward neural networks, enabling the model to encode deep semantic relationships in both forward and backward directions. This hierarchical structure allows BERT to model long-range dependencies and generate high-dimensional embeddings for each token in the sequence. To derive a sentence-level representation from the token embeddings, two common strategies are employed:

[CLS] Token Representation: The embedding at the [CLS] token position, which is designed to capture the overall semantic meaning of the entire input sequence, is directly used as a global sentence feature:

$$f_{\text{CLS}} = h_{\text{CLS}} \tag{10}$$

Mean Pooling (Mean-Pooling): The mean of all token embeddings (excluding special tokens if desired) is computed to form a general-purpose sentence representation:

$$f_{\text{mean}} = \frac{1}{n}\sum_{i=1}^{n} h_i \tag{11}$$

where n is the number of tokens in the input sequence.

To further enrich the sentence-level representation, the final text feature vector is constructed by concatenating the [CLS] embedding and the mean-pooled embedding:

$$f_t = \text{Concat}(\text{CLS}(h), \text{Mean - Pooling}(h)) \tag{12}$$

Where f_t denotes the final text feature representation used in downstream tasks such as classification or multimodal fusion. The entire process of text feature extraction using the BERT model is illustrated in Fig. 3, highlighting each stage from input tokenization to the generation of the final sentence-level feature.

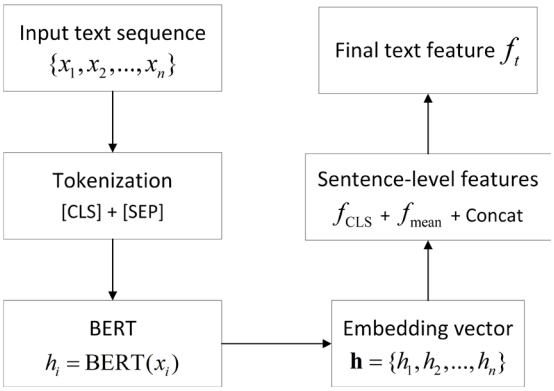

Fig. 3. Text feature extraction process

(3) Visual Feature Extraction

Visual feature extraction is conducted using a pre-trained ResNet model to obtain deep representations from input video frames. Given an input video sequence $V = \{v_1, v_2, ..., v_T\}$, where v_i denotes the i-th frame and T is the total number of frames, the original frames may vary in resolution and color format. Therefore, normalization and resizing operations are applied to ensure consistency in the input data format prior to feature extraction.

Each frame $v_i^{'}$ is individually processed by the ResNet model to extract deep visual features. ResNet employs a hierarchical convolutional architecture augmented with residual connections, which address the vanishing gradient problem by enabling gradient flow across multiple layers. This enhances the stability and efficiency of model training. The forward computation in a residual block is formulated as:

$$y = f(x) + x \tag{13}$$

where x is the input feature, $f(x)$ denotes the residual mapping function consisting of convolutional layers, batch normalization, and ReLU activation. The residual connection x allows the network to learn identity mappings more effectively and accelerates convergence.

For each video frame, ResNet computes a high-dimensional visual feature representation:

$$f_v^{'} = \text{ResNet}(v_i^{'}) \tag{14}$$

where $f_v^{'}$ represents the high-dimensional feature representation computed by ResNet. To reduce the feature dimensionality and enhance discriminative power, Global Average Pooling (GAP) is used to extract a global feature representation. Let N represent the number of feature channels in the ResNet model. The GAP calculation is as follows:

$$f_v^{'} = \frac{1}{N}\sum_{j=1}^{N} f_v^{',j} \tag{15}$$

where $f_v^{',j}$ denotes the activation value in the j-th channel. This operation reduces the number of parameters, enhances the model's generalization ability, and retains global spatial information.

This fusion method helps reduce random noise between frames, resulting in more stable final features. In multimodal emotion recognition tasks, the extracted visual features serve as one of the model inputs, which are jointly learned with other modality data (speech and text) to enhance the overall emotion recognition performance. The visual feature extraction process is illustrated in Fig. 4.

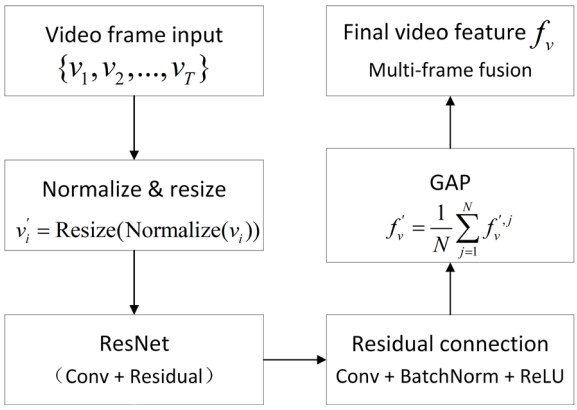

Fig. 4. Visual feature extraction process

3.2 Hierarchical Feature Fusion Module

To fully exploit high-order interactions between modalities and address inconsistencies among multimodal features, this study proposes a hierarchical feature fusion module composed of a three-layer progressive architecture. The module enhances the representation of salient information by dynamically adjusting the weights assigned to each modality. The implementation details are as follows:

(1) Preliminary Fusion and Modality Feature Weighting

The first fusion layer focuses on dimensionality reduction and the initial alignment of modality features. Features from different modalities are projected into a unified low-dimensional space through linear transformations. A dynamic weighting mechanism [24] is then applied to achieve preliminary modality alignment and fusion. Let $\{f_s, f_t, f_v\}$ denote the features extracted from modality, the dimensionally reduced features $\{\hat{f}_s, \hat{f}_t, \hat{f}_v\}$ are obtained via:

$$\hat{f}_m = \text{Linear}(f_m), \quad m \in \{s, t, v\} \tag{16}$$

Next, the features from all modalities are concatenated to generate fusion weights:

$$\alpha = \text{Soft max}(W_{\text{dyn}}[\hat{f}_s, \hat{f}_t, \hat{f}_v]) \tag{17}$$

Here, W_{dyn} is the parameter matrix used to generate dynamic weights, and $\alpha = [\alpha_s, \alpha_t, \alpha_v]$ represents the weight allocation for the audio, text, and visual modalities. The preliminary fused feature representation is computed as:

$$f_{\text{fused}}^{(0)} = \alpha_s \hat{f}_s + \alpha_t \hat{f}_t + \alpha_v \hat{f}_v \tag{18}$$

To account for the inherent uncertainty and variability of modality features, an uncertainty estimation mechanism is introduced. Using a Bayesian Neural Network, the distribution of each modality feature is modeled, and its variance is calculated to estimate uncertainty:

$$w_i = \frac{1}{\sigma_i^2 + \epsilon} \tag{19}$$

Where ϵ is a small constant added to prevent numerical instability. During final fusion, the confidence (inverse of uncertainty) serves as a weighting factor: modalities with higher confidence contribute more to the fused representation, while those with lower confidence are suppressed. The final uncertainty-aware fusion representation is given by:

$$f_{fused}^{(1)} = \mu + \epsilon \cdot \sigma, \quad \epsilon \sim \mathcal{N}(0, \mathrm{I}) \tag{20}$$

(2) Cross-Modality Interaction and Multi-Head Attention Mechanism

The second fusion layer captures deep inter-modal interactions through a multi-head cross-modal attention mechanism [25]. For the preliminarily fused features $f_{\text{fused}}^{(0)}$, the Query (Q), Key (K), and Value (V) matrices are defined as:

$$\begin{aligned} Q &= f_{\text{fused}}^{(0)} W_Q \\ K &= f_{\text{fused}}^{(0)} W_K \\ V &= f_{\text{fused}}^{(0)} W_V \end{aligned} \tag{21}$$

Where W_Q, W_K, and W_V are learnable transformation matrices. The attention weights are calculated as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{22}$$

This process ensures that the model can adaptively focus on the key information from different modalities when processing the input. To further enhance the expressive power of the features, the multi-head attention mechanism is used for fusion, specifically represented as:

$$f_{\text{fused}}^{(2)} = \text{MultiHead}(Q, K, V) + f_{\text{fused}}^{(1)} \tag{23}$$

By introducing the multi-head mechanism, the model can learn in parallel across different subspaces, thereby enhancing its ability to capture features. The fused features $f_{\text{fused}}^{(2)}$ will provide rich contextual information for subsequent tasks, improving the model's performance in various applications.

(3) Temporal Modeling and Dynamic Weight Adjustment

The third-level fusion utilizes Gated Recurrent Units (GRUs) for modeling temporal dependencies, further enhancing the expressive power of the fused features. GRUs are particularly effective in capturing long-term dependencies in sequential data, making them well-suited for handling dynamic changes in time-series data. The update rules for the GRU are as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{24}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{25}$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t]) \tag{26}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{27}$$

Where z_t and r_t are the update and reset gates, respectively, $\tilde{h}_t$ and h_t are the candidate hidden state and the final hidden state, respectively.

By employing GRUs for modeling the temporal features, the model effectively captures the dynamic interaction information between modalities as they evolve over time and passes this information to the next layer for further processing. Ultimately, the outputs from all time steps are aggregated into a global feature representation:

$$f_{\text{final}} = \sum_t \beta_t h_t \tag{28}$$

This final feature representation f_{final} is used as the input for the emotion classification task.

3.3 Classifier

The classifier plays a crucial role in multimodal emotion recognition tasks, as its design directly impacts the final prediction performance of the model. The primary function of the classifier is to map the high-level semantic representations extracted through the hierarchical feature fusion module to specific emotional categories. To achieve this, the classifier typically consists of an input layer, multiple hidden layers, and an output layer. In the input layer, the fused multimodal features are passed as input data to the hidden layers. The hidden layers, through multiple nonlinear activation processes, further extract more complex features, and ultimately, in the output layer, the representations from the hidden layers are mapped to a probability distribution over the emotional categories using the Softmax function.

The hidden layers extract higher-level representations by applying linear transformations followed by nonlinear activation to the input features. Let F denote the fused input features; the output z_i of the i-th hidden layer is computed as:

$$z_i = \mathrm{Re}\,LU(W_i F + b_i) \tag{29}$$

Where W_i and b_i are the weight matrix and bias vector of the i-th layer.

In the final layer of the classifier, the output layer is responsible for mapping the output of the hidden layers to a probability distribution over the emotional categories. Let the output of the last hidden layer be z_{final}, the output layer computes the predicted probabilities for each emotional category using the Softmax function:

$$p_j = \frac{exp(W_{out} z_{final} + b_{out})}{\sum_{k=1}^{C} exp(W_{out} z_{final} + b_{out})} \tag{30}$$

Where C denotes the total number of emotion categories, p_j represents the predicted probability of class j. The Softmax function ensures that the classifier selects the class with the highest predicted probability, thereby determining the emotional label for the input sample.

During training, the classifier's objective is to minimize the discrepancy between the predicted probabilities and the true labels. To achieve this, this paper adopts the cross-entropy loss function as the optimization objective. The cross-entropy loss measures the difference between the model's predicted probability distribution and the actual distribution of the labels, and its computation is given by:

$$\mathrm{L} = -\frac{1}{\mathrm{N}} \sum_{\mathrm{n}=1}^{\mathrm{N}} \sum_{\mathrm{c}=1}^{\mathrm{C}} \mathrm{y}_{\mathrm{n,j}} \log(\mathrm{p}_{\mathrm{n,j}}) \tag{31}$$

Where N is the number of training samples, $y_{n,j}$ is the ground truth label of the n-th sample for class j (encoded as a one-hot vector), and $p_{n,j}$ is the predicted probability of that sample belonging to class j. The cross-entropy loss function ensures that the model continuously adjusts its parameters to bring the predicted probabilities closer to the true labels, thereby improving classification accuracy.

4 Experiments and Analysis

4.1 Experimental Setup

Datasets and Splits. To validate the effectiveness of the proposed multimodal emotion recognition method, this study employs two widely used multimodal emotion recognition datasets: MOSEI and IEMOCAP. Detailed dataset splits are shown in Table 1.

Table 1. Dataset splits for MOSEI and IEMOCAP

Dataset	Training set	Validation set	Test set	Total
MOSEI	16327	1871	4662	22860
IEMOCAP (6-way)	5146	664	1623	7433

The MOSEI dataset [26], composed of 22,860 video clips sourced from YouTube, is widely used in multimodal sentiment and emotion recognition tasks. It provides both a 7-point continuous emotion annotation scale (ranging from -3 [strongly negative] to +3 [strongly positive]) and six categorical emotion labels: happy, sad, angry, fearful, disgusted, and surprised.

The IEMOCAP dataset [27] comprises 7,433 samples and adopts a six-class categorical emotion classification scheme, including happy, sad, neutral, angry, excited, and frustrated. It is primarily used to evaluate the generalization capability of the proposed model.

Experimental Configuration and Evaluation Metrics. All experiments were implemented using the PyTorch deep learning framework [28]. The dimensionalities of the input features were set to 128 for audio, 768 for text, and 512 for visual modalities. These were projected into a unified 128-dimensional feature space via a feature dimensionality reduction module.

The core model employs a Transformer-based encoder architecture consisting of 4 layers, each with 8 attention heads. The hidden size is set to four times the embedding dimension. Dropout regularization is applied throughout the network: a dropout rate of 0.3 is used in the classifier, while a rate of 0.1 is applied to the remaining modules. Layer Normalization is incorporated to facilitate faster convergence and improve model stability.

During optimization, the initial learning rate is set to 0.0003, and dynamic learning rate adjustment is performed using the ReduceLROnPlateau strategy. AdamW is adopted as the optimizer, with weight decay employed to mitigate overfitting. Gradient clipping is applied to prevent gradient explosion. The batch size is set to 64, and the maximum number of training epochs is capped at 50. An early stopping mechanism is adopted, terminating training if no significant improvement is observed on the validation set for five consecutive epochs.

For evaluation, binary accuracy (ACC-2) and 7-class accuracy (ACC-7) are used as metrics on the MOSEI dataset. On the IEMOCAP dataset, overall accuracy (ACC) and weighted F1 score (WF1) are employed as the primary evaluation criteria.

Baseline Models. To comprehensively assess the effectiveness of the proposed approach, we compare it against several classical baseline models commonly used in the field of emotion recognition. For the MOSEI dataset, detailed evaluation results are presented in Table 2. The selected baselines include Tensor Fusion Network (TFN) [29], Low-rank Multimodal Fusion (LMF) [30], Multimodal Factorization Model (MFM) [31], and Interpretable Cross-modal Correlation Network (ICCN) [32]. For the IEMOCAP dataset, the corresponding performance comparisons are shown in Table 3 and Table 4. The baselines evaluated include TFN, Memory Fusion Network (MFN) [33], DialogueRNN [34], DialogueGCN [35], and COGMEN [36].

4.2 Comparative Experimental Analysis

We conducted a systematic performance comparison between the proposed HFF model and the aforementioned baseline models. As shown in the results presented in Table 2 and Table 3, the proposed HFF model consistently outperforms most baseline methods on both the MOSEI and IEMOCAP datasets, achieving notable improvements across key evaluation metrics.

On the MOSEI dataset, the HFF model achieved an accuracy of 52.8% in the ACC-7 task and 84.6% in the ACC-2 task, surpassing the performance of all baseline models. For the ACC-7 task, the classification accuracy of HFF is comparable to that of MFM and ICCN, indicating its ability to effectively model multimodal features and maintain robust performance in complex emotion classification scenarios. In the ACC-2 task, the HFF model outperforms all baseline methods, with an accuracy improvement of over 2% compared to TFN and LMF. This demonstrates the model's superior stability in binary classification tasks and its effectiveness in handling sentiment polarity classification.

On the IEMOCAP dataset, the HFF model achieved an F1-score of 69.7% and an accuracy of 69.2%, demonstrating its strong overall performance. In particular, the model outperforms all competitors in classifying Neutral, Angry, and Excited emotions, reflecting its balanced learning capability across various emotional states. For the Happy category, HFF attained an F1-score of 51.0%, slightly below the 51.9% achieved by COGMEN.

This may be attributed to a more conservative feature weighting strategy in HFF for this category, limiting performance gains. However, in the Sad category, HFF reached an F1-score of 81.4%, comparable to COGMEN, indicating stable classification capability for this emotion.

Table 2. Emotion classification performance on the MOSEI dataset

MOSEI dataset					
Model	TFN	ICCN	MFM	LMF	HFF
ACC - 7 ↑	50.2	51.6	51.3	48.0	52.8
ACC - 2 ↑	82.5	84.2	84.4	82.0	84.6

Table 3. Emotion classification performance on the IEMOCAP (6-way) dataset

Model / Emotion	IEMOCAP (6-way) dataset							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Avg	
	F1	F1	F1↑	F1↑	F1↑	F1	ACC↑	F1↑
TFN	33.7	68.6	55.1	64.2	62.4	61.2	58.8	58.5
MFN	34.1	70.5	52.1	66.8	62.1	62.5	60.1	59.9
DialogueRNN	32.8	78.0	59.1	63.3	73.6	59.4	63.3	62.8
DialogueGCN	42.7	**84.5**	63.5	64.1	63.1	**66.9**	65.2	64.2
COGMEN	**51.9**	81.7	68.6	66.0	75.3	58.2	68.2	67.6
HFF	51.0	81.4	**71.0**	**67.0**	**75.8**	62.4	**69.2**	**69.7**

To provide a more intuitive illustration of the HFF model's performance on the IEMOCAP dataset, Fig. 5 presents its confusion matrix. As shown, the model achieved high classification accuracy for the Sad and Excited categories, reaching 81% and 75%, respectively, indicating its effectiveness in capturing features associated with high-intensity emotions. Conversely, the Happy category shows relatively lower accuracy at 51%, likely due to greater emotional variability in its distribution. For the Neutral and Angry categories, the recognition rates are 71% and 67%, respectively, suggesting a balanced performance across both emotionally neutral and intense classes.

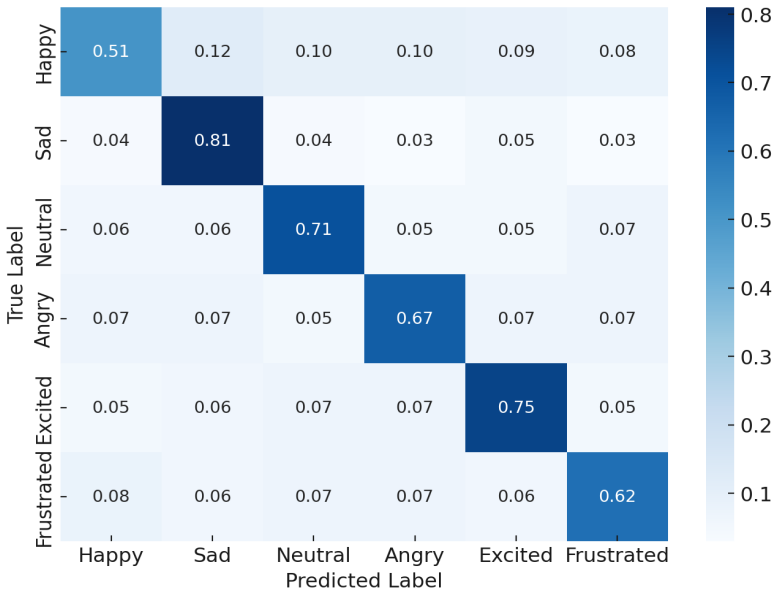

Fig. 5. Confusion matrix of the HFF model on the IEMOCAP dataset

Fig. 5 further illustrates the HFF model's classification performance across different emotional categories, showcasing its ability to learn in the six-way classification task on the IEMOCAP dataset. In conjunction with the results in Table 3, the model demonstrates a relatively balanced distribution of F1-scores across categories, indicating strong and generalized emotion recognition capabilities.

4.3 Ablation Study Analysis

To investigate the contribution of key components within the HFF model and evaluate the impact of different modality combinations on emotion classification performance, we conducted a series of ablation studies on the IEMOCAP (6-way) dataset. These experiments aimed to examine the role of each module in the context of multimodal emotion recognition. The detailed results are presented in Table 4.

As shown in the table, the complete HFF model achieves the highest F1-score of 69.7% under the full multimodal configuration (S + T + V), demonstrating that the model can reach optimal performance when fully leveraging cross-modal interactions, dynamic weighting mechanisms, and temporal modeling components. In contrast, removing the cross-modal alignment module causes the F1-score to drop to 68.5%, indicating the critical role of this module in capturing deeper inter-modal interactions. The impact is particularly evident in unimodal and bimodal settings. For instance, under the S + T configuration, the F1-score drops from 67.6% to 66.3% after removing the cross-modal alignment module, validating its optimization role when modality quality varies.

Table 4. Ablation study results of the HFF model on the IEMOCAP (6-way) dataset

Module / Modality	S	T	V	S+T	S+V	T+V	S+T+V
HFF (Full model)	62.8	65.5	60.4	67.6	66.2	68.1	69.7
Without attention mechanism	60.2	63.7	57.5	64.6	63.2	65.4	65.9
Without dynamic weighting	61.3	64.3	58.3	65.4	64.0	66.2	67.5
Without temporal modeling	61.5	65.0	59.8	66.8	65.5	67.2	67.5
Without cross-modality alignment	61.7	64.8	58.5	66.3	64.9	67.0	68.5

The removal of the attention mechanism results in a noticeable decline in overall model performance, especially in the full multimodal setting, where the F1-score drops to 65.9%. This highlights the mechanism's crucial role in extracting salient modality-specific information and suppressing irrelevant signals. Similarly, ablation of the dynamic weighting mechanism leads to a reduction in performance to 67.5%, indicating its significance in optimizing the distribution of modality contributions. The removal of the temporal modeling module (GRU) also leads to a performance drop. Although the F1-score under the S+T+V configuration remains at 67.5%, the model's ability to capture temporal dynamics is compromised, particularly in modality combinations involving speech (S), where performance degrades more substantially.

In unimodal settings, the text modality yields the best individual performance (F1 = 65.5%), whereas the visual and acoustic modalities show relatively lower standalone effectiveness. This further confirms the stability of textual information in emotion recognition tasks and underscores the necessity of multimodal fusion strategies in handling complex emotion classification scenarios.

These results collectively demonstrate the critical role of each component within the HFF model in enhancing multimodal emotion recognition. Notably, the integration of attention mechanisms, dynamic weight adjustment, and temporal modeling significantly enhances inter-modality information exchange and optimizes feature extraction across different modality combinations, thereby contributing to improved overall emotion classification performance.

4.4 The Impact of Attention Head Count on Model Performance

The multi-head attention mechanism is a core component of the Transformer architecture and plays a vital role in multimodal emotion recognition tasks. Appropriately configuring the number of attention heads directly influences the model's ability to represent features from different modalities and facilitates effective information exchange. To investigate how the number of attention heads affects model performance, we conducted experiments based on the HFF model. While keeping all other hyperparameters constant, we tested models with 4, 6, 8, and

10 attention heads. The evaluation metrics include accuracy (ACC) and weighted F1-score (WF1). The results are presented in Table 5.

As the number of attention heads increases, model performance improves within a certain range. When the number of heads reaches 8, the model achieves its peak performance. Specifically, on the MOSEI dataset, ACC increases from 81.3% to 84.4%, and the F1-score improves from 80.7% to 82.3%. On the IEMOCAP dataset, ACC rises from 67.6% to 69.0%, while the F1-score increases from 66.3% to 69.5%. These improvements indicate that a moderate number of attention heads can enhance the model's capacity for multimodal feature representation and improve the accuracy of emotion classification.

However, when the number of attention heads is further increased to 10, a slight decline in performance is observed. On the MOSEI dataset, ACC drops to 83.0% and the F1-score to 82.1%, while on the IEMOCAP dataset, ACC decreases to 68.9% and the F1-score to 69.2%. This degradation suggests that an excessive number of attention heads may introduce redundant information, diminishing the effectiveness of cross-modal interactions and adversely affecting training stability and generalization performance.

The trends observed across the MOSEI and IEMOCAP datasets are consistent, demonstrating that a well-chosen number of attention heads can significantly enhance the model's capability to model heterogeneous modality-specific features, thereby improving the stability and accuracy of emotion recognition. When the number of heads is set to 8, all performance metrics across both datasets reach optimal levels, indicating a favorable balance between computational complexity and performance improvement. These experimental results further validate the critical role of multi-head attention in multimodal emotion recognition and underscore the importance of appropriately tuning model parameters to achieve optimal recognition performance.

Table 5. Performance of the HFF model under different attention head configurations

Number of attention heads	MOSEI-ACC	MOSEI-F1	IEMOCAP-ACC	IEMOCAP-F1
4	81.3	80.7	67.6	66.3
6	82.0	81.3	68.2	67.1
8	84.4	82.3	69.0	69.5
10	83.0	82.1	68.9	69.2

5 Conclusion

This paper introduces the Hierarchical Feature Fusion (HFF) model, a novel approach for multimodal emotion recognition that addresses the inherent challenges of integrating textual, audio, and visual modality information. The HFF model employs a hierarchical fusion strategy, facilitating efficient interactions between different modalities through a progressive integration mechanism. Additionally, the model incorporates uncertainty estimation, which dynamically adjusts the weighting of each modality, thereby enhancing the robustness and stability of emotion recognition. By independently assigning weights to each modality's features and incorporating advanced mechanisms such as multi-head attention and gated recurrent units (GRUs), the model is able to capture the intricate dependencies between modalities, leading to improved overall emotion recognition performance.

Experimental results validate the efficacy of the HFF model, demonstrating its powerful feature fusion capabilities across multiple datasets. Notably, the HFF model outperforms other state-of-the-art methods in terms of accuracy and stability when evaluated on the MOSEI and IEMOCAP datasets. The model's ability to effectively combine multimodal information and account for temporal dynamics allows it to handle modality heterogeneity and temporal variations with remarkable precision, enabling it to capture subtle nuances in emotional expressions.

However, this study acknowledges several limitations. The introduction of the hierarchical feature fusion mechanism increases computational complexity, which may pose challenges in large-scale datasets or real-time applications due to the significant computational burden. Additionally, the imbalance in data distribution remains a concern, as it continues to affect the classification performance for specific emotion categories.

Future research could explore the incorporation of additional modality information, such as physiological signals or EEG data, to further enhance the model's perceptual capabilities and improve emotion recognition accuracy. To mitigate the issue of computational complexity, it will be crucial to investigate lightweight fusion mechanisms, such as optimizing attention structures or applying model pruning techniques. Furthermore, inte-

grating self-supervised learning and transfer learning approaches, particularly in resource-constrained environments, could enhance the model's adaptability and performance. Future work will continue to refine multimodal emotion recognition technologies, aiming to provide more robust, accurate, and efficient solutions for practical applications.

References

[1] R. Das, T.D. Singh, Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges, ACM Computing Surveys 55(13s)(2023) 1–38. https://doi.org/10.1145/3586075

[2] G. Chen, X. Zeng, Multi-modal emotion recognition by fusing correlation features of speech-visual, IEEE Signal Processing Letters 28(2021) 533–537. https://doi.org/10.1109/LSP.2021.3055755

[3] A. Koduru, H.B. Valiveti, A.K. Budati, Feature extraction algorithms to improve the speech emotion recognition rate, International Journal of Speech Technology 23(1)(2020) 45–55. https://doi.org/10.1007/s10772-020-09672-4

[4] T. Baltrušaitis, C. Ahuja, L.P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2)(2019) 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[5] K. Zhang, Y. Li, J. Wang, Z. Wang, X. Li, Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis, IEEE Signal Processing Letters 28(2021) 1898–1902. https://doi.org/10.1109/LSP.2021.3112314

[6] X.H. Qi, M. Zhi, Review of attention mechanisms in image processing, Journal of Frontiers of Computer Science & Technology 18(2)(2024) 345-362. https://doi.org/10.3778/j.issn.1673-9418.2305057

[7] S. Chen, Y. Liu, J. Wang, Q. Zhang, Z. Zhou, A multi-stage dynamical fusion network for multimodal emotion recognition, Cognitive Neurodynamics 17(3)(2023) 671–680. https://doi.org/10.1007/s11571-022-09851-w

[8] D. Hu, X. Hou, L. Wei, L. Jiang, Y. Mo, MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations, in: Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022. https://doi.org/10.1109/ICASSP43922.2022.9747397

[9] D. Mamieva, A.B. Abdusalomov, A. Kutlimuratov, B. Muminov, T.K. Whangbo, Multimodal emotion detection via attention-based fusion of extracted facial and speech features, Sensors 23(12)(2023) 5475. https://doi.org/10.3390/s23125475

[10] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, C. Fookes, Attention driven fusion for multi-modal emotion recognition, in: Proc. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020. https://doi.org/10.1109/ICASSP40776.2020.9054441

[11] T. Shi, S.L. Huang, MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations, in: Proc. 2023 Annual Meeting of the Association for Computational Linguistics (ACL), 2023. https://doi.org/10.18653/v1/2023.acl-long.824

[12] P.F. Liu, K. Li, H.L. Meng, Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition, in: Proc. Interspeech 2020, 2020. https://doi.org/10.21437/Interspeech.2020-2067.

[13] L. Yuan, J. Wang, L.C. Yu, X. Zhang, Graph attention network with memory fusion for aspect-level sentiment analysis, in: Proc. 2020 Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020. https://doi.org/10.18653/v1/2020.aacl-main.4

[14] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, AI Open 1(2020) 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001

[15] A. Shirian, T. Guha, Compact graph architecture for speech emotion recognition, in: Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021. https://doi.org/10.1109/ICASSP39728.2021.9413876

[16] J. Liu, H. Wang, Graph isomorphism network for speech emotion recognition, in: Proc. 2021 Interspeech Conference, 2021. https://doi.org/10.21437/Interspeech.2021-1154

[17] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Proc. Advances in Neural Information Processing Systems, 2017.

[18] C. Blundell, J. Cornebise, K. Kavukcuoglu, D. Wierstra, Weight uncertainty in neural network, in: Proc. 2015 International Conference on Machine Learning, 2015.

[19] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: Proc. 2016 International Conference on Machine Learning, 2016.

[20] M. K. Tellamekala, S. Amiriparian, B.W. Schuller, E. André, T. Giesbrecht, M. Valstar, COLD fusion: calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 46(2)(2024) 805–822. https://doi.org/10.1109/TPAMI.2023.3325770

[21] F. Chen, J. Shao, A. Zhu, D. Ouyang, X. Liu, H.T. Shen, Modeling hierarchical uncertainty for multimodal emotion recognition in conversation, IEEE Transactions on Cybernetics 54(1)(2024) 187–198. https://doi.org/10.1109/TCYB.2022.3185119

[22] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, in: Proc. 2019 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, 2019.
[23] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, S. Nanayakkara, Multimodal emotion recognition with transformer-based self supervised feature fusion, IEEE Access 8(2020) 176274–176285. https://doi.org/10.1109/ACCESS.2020.3026823
[24] K. Yang, H. Xu, K. Gao, Cm-bert: cross-modal bert for text-audio sentiment analysis, in: Proc. 2020 ACM International Conference on Multimedia, 2020. https://doi.org/10.1145/3394171.3413690
[25] Y. Du, Y. Liu, Z. Peng, X. Jin, Gated attention fusion network for multimodal sentiment classification, Knowledge-Based Systems 240(2022) 108107. https://doi.org/10.1016/j.knosys.2021.108107
[26] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2018. https://doi.org/10.18653/v1/P18-1208
[27] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Language Resources and Evaluation 42(4)(2008) 335–359. https://doi.org/10.1007/s10579-008-9076-6
[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Proc. Advances in Neural Information Processing Systems, 2019.
[29] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.P. Morency, Tensor Fusion Network for Multimodal Sentiment Analysis, in: Proc. 2017 Conference on Empirical Methods in Natural Language Processing, 2017. https://doi.org/10.18653/v1/D17-1115
[30] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L.P. Morency, Efficient Low-Rank Multimodal Fusion with Modality-Specific Factors, in: Proc. 2018 56th Annual Meeting of the Association for Computational Linguistics, 2018. https://doi.org/10.18653/v1/P18-1209
[31] Y.H.H. Tsai, P.P. Liang, A. Zadeh, L.P. Morency, R. Salakhutdinov, Learning Factorized Multimodal Representations, in: Proc. 2019 7th International Conference on Learning Representations, 2019.
[32] Z. Sun, P. Sarma, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: Proc. the AAAI Conference on Artificial Intelligence, 2020. https://doi.org/10.1609/aaai.v34i05.6431
[33] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.P. Morency, Memory fusion network for multi-view sequential learning, in: Proc. AAAI Conference on Artificial Intelligence, 2018. https://doi.org/10.1609/aaai.v32i1.12021
[34] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, DialogueRNN: An attentive RNN for emotion detection in conversations, in: Proc. AAAI Conference on Artificial Intelligence, 2019. https://doi.org/10.1609/aaai.v33i01.33016818
[35] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. https://doi.org/10.18653/v1/D19-1015
[36] A. Joshi, A. Bhat, A. Jain, A. Singh, A. Modi, COGMEN: Contextualized GNN Based Multimodal Emotion Recognition, in: Proc. 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022. https://doi.org/10.18653/v1/2022.naacl-main.306