Current Science
CS 5 (5), 4488-4498 (2025)
Received 15 June 2025 | **Revised 10 July** 2025 | **Accepted 16 August** 2025 | **Online Available 19 September** 2025
https://doi.org0/10.52845/CS/2025-5-5-14

Current Science

Current Science

OPEN ACCESS

ISSN(O)2795-8639

Original Article

Research on Multimodal Emotion Recognition Method Based on Multi-Path Feature Fusion and Comprehensive Optimization Strategy

Zhanpeng Li[1,2], Yuming Qi[1,2*], Sanpeng Deng[1,2], Xiumin Shi[1,2]

[1]Tianjin University of Technology and Education, Tianjin 300222, China

[2]Tianjin Key Laboratory of Intelligent Robot Technology and Application, Tianjin 300350, China

[*]Corresponding Author: Yuming Qi

Abstract:
Multimodal emotion recognition plays a crucial role in human–computer interaction and related fields. However, deep learning models often face severe overfitting due to insufficient feature interaction, data imbalance, and high model complexity. To address these challenges, this paper systematically explores the evolution from simple feature concatenation to advanced attention-based fusion, and proposes a Multi-Path Attention Fusion Network (MPAF-Net), which processes original unimodal features and cross-modal attention–generated interactive features in parallel to achieve more comprehensive information representation. Furthermore, we integrate a set of comprehensive optimization strategies, including multi-level regularization, class-weighted Focal Loss, and learning rate warmup with cosine annealing. Experimental results on the IEMOCAP dataset show that, compared with the baseline, MPAF-Net achieves a macro F1-score of 50.2% on the validation set and 47.42% on the test set, significantly outperforming the baseline. This study not only validates the effectiveness of the proposed method but also demonstrates the great potential of multi-path fusion architectures in enhancing model performance and robustness for complex multimodal tasks.

Key Words: Multimodal, Emotion recognition, Cross modal attention, Feature fusion, Speaker independent

1. Introduction

With the increasing intelligence of human–computer interaction systems, emotion recognition has attracted wide attention as an important technology for understanding user states, enhancing natural interaction, and improving personalized experiences [1]. Human emotion is expressed through multiple modalities. Compared with single-modality information sources, multimodal emotion recognition (MER), which integrates text, speech, and visual information, captures richer emotional cues and demonstrates stronger robustness in complex scenarios.

Despite significant progress, current deep learning–based MER methods still face three major challenges. The first is limited generalization, particularly under the strict speaker-independent setting, where overfitting becomes prominent [2]. In this setting, a model is required to recognize emotions of unseen speakers, which demands learning general patterns across individuals. However, studies have shown that deep models tend to memorize acoustic traits or habits of speakers in the training set instead of learning abstract emotional patterns, leading to a sharp performance decline in real-world scenarios. To mitigate overfitting, regularization techniques have been widely applied. Dropout [3] randomly deactivates neurons to enforce robust feature learning, while weight decay (L2 regularization) constrains model complexity by penalizing large weights. These techniques have become foundational components of deep model design and remain essential in comprehensive optimization strategies.

The second challenge lies in the imbalance of emotional categories in real-world data [4]. In the widely used IEMOCAP dataset [5], classes such as fear (fea) contain far fewer samples than neutral (neu) or frustration (fru). This skewed distribution biases models toward majority classes, degrading balanced performance metrics such as macro F1-score and reducing recognition accuracy for minority classes. Cost-sensitive learning provides an effective solution by adjusting the loss function, often by assigning higher weights to minority-class samples. Focal Loss [6] further introduces a modulation factor that down-weights easily classified samples and focuses learning on harder cases, making it a core algorithmic tool for addressing class imbalance in this study.

The third and most critical challenge concerns the design of efficient feature fusion mechanisms [1,7]. According to the fusion stage, mainstream approaches can be divided into early, late, and hybrid fusion. Early fusion [8], typically realized by simple concatenation, is straightforward but fails to model complex nonlinear correlations between modalities, resulting in performance bottlenecks that have been confirmed in prior research and our baseline experiments. Late fusion [8], which combines decisions at the output stage, offers high flexibility but often sacrifices low-level intermodal interactions. Hybrid fusion, especially attention-based methods [9], has therefore become the focus of current studies. Attention mechanisms enable dynamic weighting of modalities. Cross-modal attention [10], for instance, aligns and reweights information across modalities through query–key–value interactions. Structures such as CAG-MoE [11] further introduce gating and mixture-of-experts strategies to enhance fusion precision while maintaining efficiency. However, most existing methods follow a "replacement-based" paradigm, where newly generated features overwrite the original ones, potentially causing information loss. Tensor Fusion Networks (TFN) [12] capture higher-order interactions but suffer from heavy computational costs and unstable training. Thus, the key challenge is to design a fusion architecture that captures deep interactions while retaining the integrity of original information.

To systematically address these challenges, a progressive set of studies was conducted. A feature-concatenation baseline model was first constructed and rigorously evaluated to quantify the performance bottleneck. After exploring more advanced fusion mechanisms, a novel architecture named Multi-Path Attention Fusion Network (MPAF-Net) was proposed. Guided by the principle of "augmentation rather than replacement," the network processes unimodal features and cross-modal attention–generated interactive features in parallel, achieving a balance between information integrity and deep interaction. To further exploit its potential, comprehensive optimization strategies were integrated, including hierarchical regularization, class-weighted Focal Loss, and learning rate scheduling techniques. Experiments on the IEMOCAP dataset verified the effectiveness of the proposed

The study makes the following contributions:

1. A novel architecture (MPAF-Net): A multi-path fusion network was designed, where unimodal features and cross-modal attention–based features are processed in parallel to balance information integrity and deep interaction.
2. Targeted optimization strategies: Hierarchical regularization, class-weighted Focal Loss, and learning rate warmup with cosine annealing were integrated to improve robustness, handle class imbalance, and stabilize convergence.
3. Extensive validation: Comparative experiments on the IEMOCAP dataset demonstrated that MPAF-Net significantly outperforms the baseline in terms of balanced performance metrics.

2 Method

2.1 Overall Framework

The overall framework of the proposed multimodal emotion recognition model is illustrated in Figure 1. The model follows a hierarchical processing pipeline consisting of three core components: unimodal encoders, a multimodal fusion module, and a classifier.

1. Unimodal Encoders: These modules extract deep feature representations from raw text, audio, and video data.
2. Multimodal Fusion Module: Serving as the core of the framework, this module receives unimodal features from the encoders and generates fused representations that capture intermodal information through specific mechanisms such as simple concatenation or attention-based interaction.
3. Classifier: A multilayer perceptron (MLP) that takes the fused feature vector as input and outputs

the predicted probabilities of emotion categories.

2.2 Baseline Model

To establish a performance reference and quantify the limitations of existing methods, a widely adopted baseline model was constructed and evaluated. This model strictly follows the framework illustrated in Figure 1.

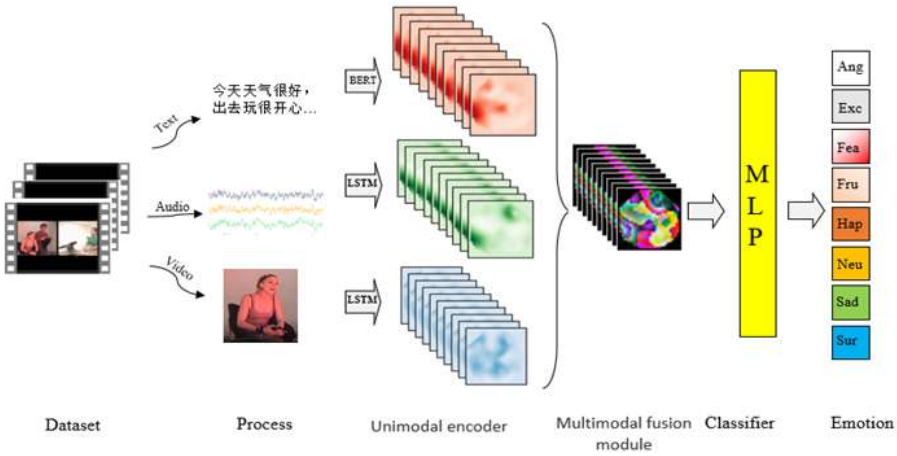

Figure 1 Overall Framework

1. Text Encoder: A pretrained BERT-base-uncased model [13] was employed. The 768-dimensional output corresponding to the [CLS] token in the last layer was used as the semantic representation of the text, denoted as $h\_t \in R^{768}$. To balance computational cost and performance, only the last two layers of BERT were fine-tuned.
2. Audio Encoder: The input was a sequence of Mel-Frequency Cepstral Coefficients (MFCCs). A unidirectional Long Short-Term Memory (LSTM) network [14] processed this sequence, and the hidden state at the last time step was taken as the audio feature representation, denoted as $h\_a \in R^{128}$.
3. Video Encoder: The input was a sequence of facial frames extracted from video. A unidirectional LSTM was used to model the temporal variations of facial expressions, and the hidden state at the last time step was taken as the video feature representation, denoted as $h\_v \in R^{256}$.
4. Fusion and Classification: The model adopted an early fusion strategy. The three feature vectors were directly concatenated into a high-dimensional vector:

$$h\_fused = Concat(h\_t,\ h\_a,\ h\_v) \in R^{1152} \quad (1)$$

2.3 Multi-Path Attention Fusion Network (MPAF-Net)

To address the limitations revealed by the baseline model, a Multi-Path Attention Fusion Network (MPAF-Net) was designed. The central innovation of this model lies in the concept of multi-path feature fusion. Instead of replacing unimodal features with interaction features, MPAF-Net processes unimodal features and cross-modal interaction features in parallel, using both as the basis for final classification. The detailed architecture is illustrated in Figure 2.

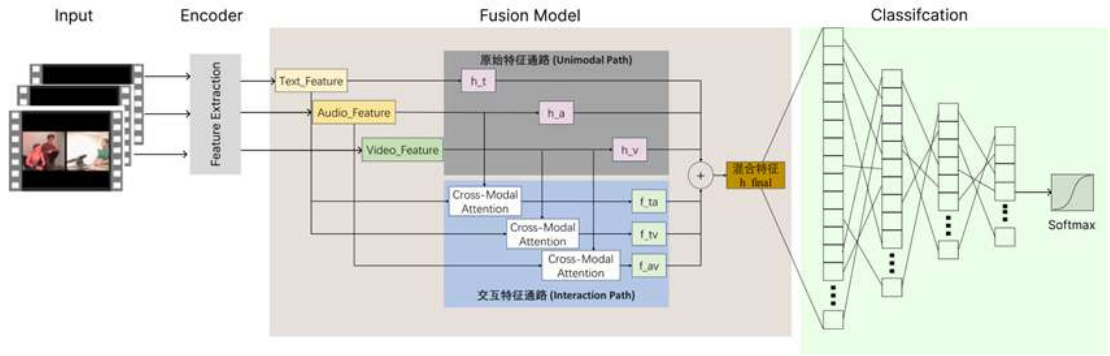

Figure 2 MPAF-Net Framework

2.3.1 Cross-Modal Attention Module

To generate interaction features between modalities, an enhanced pairwise cross-modal attention module was designed, as illustrated in Figure 3.

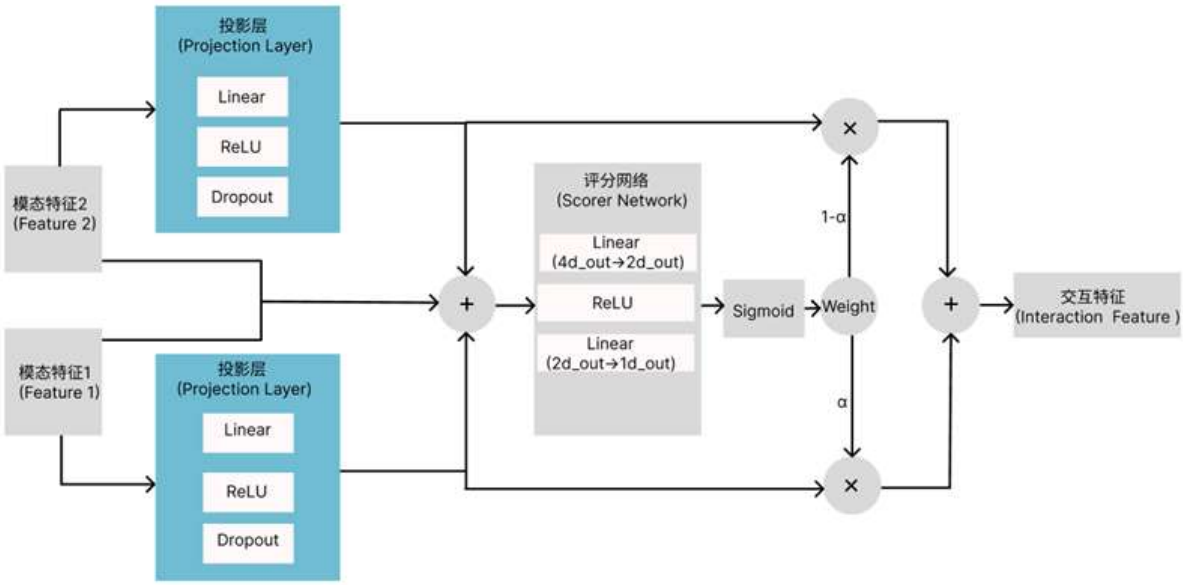

Figure 3 Cross-modal Attention Module

For any two modality feature vectors $f\_1 \in \mathrm{R}^{\mathrm{d}\_1}$ and $f\_2 \in R^{d\_2}$, the module aims to produce an interaction feature $f\_fused \in R^{d\_\mathrm{out}}$. The computation process is as follows:

1. Feature Projection:

Each input feature is projected into a common hidden dimension d_out using a linear layer followed by ReLU activation and Dropout:

$$P\_1 = Dropout(\mathrm{ReLU}(W\_1 * f\_1 + b\_1)) \quad (2)$$

$$P\_2 = Dropout(\mathrm{ReLU}(W\_2 * f\_2 + b\_2)) \quad (3)$$

where W_1, W_2, b_1, b_2 are learnable parameters, and ReLU() denotes the rectified linear unit activation.

2. Attention Score Calculation:

To fully exploit both original and projected features, they are concatenated and fed into a scoring network (a two-layer MLP) to compute a scalar attention score s:

$$s = W_{s\_2} * ReLU\big(W_{s\_1} * Concat(f_{\_1}, f_{\_2}, p_{\_1}, p_{\_2}) + b_{\_s1}\big) + b_{\_s\_2} \quad (4)$$

where W_{s_1}, W_{s_2} are weight matrices, and $f_{\_1}, f_{\_2}, p_{\_1}, p_{\_2}$ are concatenated input vectors.

3. Attention Weight Generation:

The score is normalized by a Sigmoid function to yield the attention weight α, constrained within (0,1)

$$\alpha = \text{Sigmoid}(s) \quad (5)$$

4. Feature Fusion:

Finally, the projected features are combined using the attention weight to form the interaction feature:

$$f\_fused = \alpha * p\_1 + (1-\alpha) * p\_2 \quad (6)$$

where f_fused is the final interaction feature, and α represents the weight assigned to *p_1*.

In MPAF-Net, three such cross-modal attention modules are instantiated to compute the interaction features for text–audio f_ta, text–video f_tv, and audio–video f_av.

2.3.2 Final Fusion and Classification

Unlike previous approaches that replace unimodal features with fused representations, the proposed model follows an information-maximization strategy. All available features—including three dropout-processed unimodal features (h'_t, h'_a, h'_v) and three newly generated interaction features (f_ta, f_tv, f_av) —are concatenated along the feature dimension:

$$h\_final = (h'\_t, h'\_a, h'\_v, f\_ta, f\_tv, f\_av) \quad (7)$$

The resulting high-dimensional and information-rich feature vector h_final is then passed to a lightweight, strongly regularized MLP classifier for final emotion prediction.

2.4 Integrated Training Optimization Strategies

To enhance generalization, address data imbalance, and fully exploit the potential of the proposed architecture, a set of integrated training optimization strategies was employed:

1. Regularization Combination:

Hierarchical Dropout: Different dropout rates were applied at multiple levels, including encoder outputs, projection layers, and each layer of the classifier, thereby improving generalization systematically.

Weight Decay: Differential L2 regularization was applied to parameters of BERT and other components at the optimizer level to constrain model complexity.

2. Cost-Sensitive Loss Function:

To handle class imbalance, the model adopted Focal Loss as the objective function, defined as:

$$FL(p\_t) = -\alpha\_t * (1-p\_t)\text{^}\gamma * \log(p\_t) \quad (8)$$

where p_t is the predicted probability for the ground-truth class. α_t is a class-specific weight determined by the inverse frequency of the class in the training set, giving higher weight to minority classes. The focusing parameter γ set to 2.0 in this work) reduces the loss contribution of easily classified samples, enabling the model to focus on harder cases.

3. Learning Rate Warmup with Cosine Annealing:

To achieve stable and efficient convergence, a scheduling strategy combining warmup and cosine annealing was applied. In the initial stage (e.g., the first three epochs), the learning rate increased linearly from zero to a base value (warmup). In the remaining epochs, it decayed smoothly to a minimum value following a cosine function (cosine annealing). This strategy has been shown to help the model escape local optima and achieve better solutions.

3 Experiments and Analysis

This section systematically evaluates the effectiveness of the proposed MPAF-Net and the associated optimization strategies. We first describe the datasets, evaluation metrics, and implementation details. Then, we compare MPAF-Net with baseline models to demonstrate overall performance improvements. Finally, ablation studies are conducted to analyze the contributions of each optimization module.

3.1 Dataset and Preprocessing

3.1.1 Dataset Description

The experiments are conducted on the IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset [5], one of the most widely used benchmarks in multimodal emotion recognition. The dataset contains approximately 12 hours of audiovisual recordings from 10 actors (5 male and 5 female) performing both improvised and scripted dialogues.

Following the conventional paradigm in this field, nine discrete emotion categories are retained: anger (ang), excitement (exc), fear (fea), frustration (fru), happiness (hap), neutral (neu), sadness (sad), surprise (sur). The disgust (dis) category, which has very few samples, is excluded during training.

To ensure rigorous evaluation, a speaker-independent split is adopted. The 10 actors are partitioned at an 8:1:1 ratio into training (8 actors), validation (1 actor), and testing (1 actor) sets, such that there is no overlap of speakers across these three sets. This

guarantees that the model is evaluated on entirely unseen individuals.

3.1.2 Data Preprocessing

To convert the raw IEMOCAP recordings into a format suitable for multimodal deep learning, we design a comprehensive preprocessing pipeline for text, audio, and video modalities, as illustrated in Figure 4. The pipeline extracts effective features from each modality and standardizes them into a unified format.

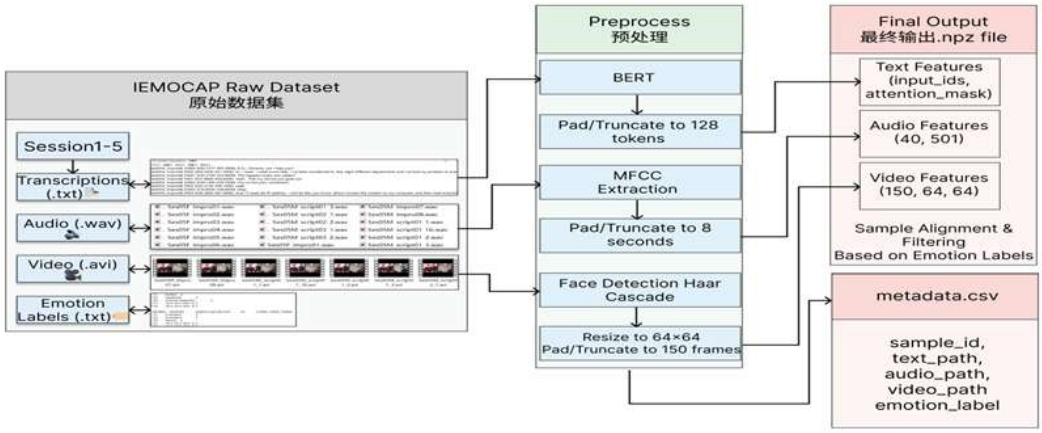

Figure 4 Preprocessing Framework

1. Text Modality:For each utterance, the original transcription is tokenized using the bert-base-uncased tokenizer. Each sentence is converted into a numerical sequence of tokens. To ensure consistent input length for batch processing, sequences are either padded or truncated to a fixed length of 128 tokens. Each sample is stored as an .npz file containing input_ids and attention_mask.
2. Audio Modality:Each .wav file is first resampled to 16 kHz. To address variable utterance lengths, all audio signals are normalized to a fixed duration of 8 seconds (shorter signals padded with silence, longer ones truncated). From each normalized signal, 40-dimensional Mel-frequency cepstral coefficients (MFCCs) are extracted, forming a fixed-size feature matrix, which is saved as a .npy file.
3. Visual Modality:Since facial expressions convey rich emotional cues, each .avi file is processed frame by frame. Face detection is performed using the Haar Cascade classifier from OpenCV. When multiple faces are detected, the largest face is assumed to be the primary speaker. The detected face regions are cropped, converted to grayscale, and resized to 64×64 pixels. To ensure uniform sequence length, all face sequences are standardized to 150 frames by either truncation or padding (repeating the last frame). Each video sample is represented as a tensor of shape (150,64,64)and stored as a .npy file. An example visualization of extracted audio and video features (sample Ses02F_impor04) is shown in Figure 5.

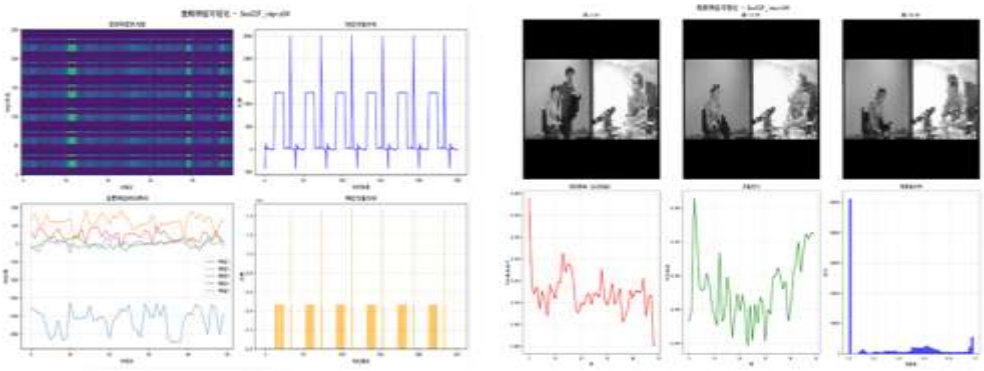

Figure 5 Feature visualization

4. Multimodal Alignment:Finally, all modalities are aligned using the emotion annotation files as reference. Each labeled utterance is matched 4.Multimodal Alignment:Finally, all modalities are aligned using the emotion annotation files as reference. Each labeled utterance is matched with its processed text, audio, and video features. If feature extraction fails for any modality (e.g., no face detected in video), the sample is discarded to preserve dataset integrity. A metadata file (metadata.csv) is generated as the master index, recording the sample_id, paths to the three modality features, and the corresponding emotion_label. The final class distribution after preprocessing is summarized in Table 1.

Table 1 Dataset category classification

emotion	Ang	Dis	Exc	Fea	Fru	Hap	Neu	Sad	Sur
support	1103	2	1041	40	1849	595	1780	1084	107

From Table 1, it is evident that the dataset suffers from a severe class imbalance problem. For instance, the disgust (dis) category contains only 2 samples, while the number of frustration (fru) samples is more than 40 times that of fear (fea). Such imbalance poses a major challenge to the generalization ability of the model.

3.2 Evaluation Metrics and Experimental Settings

Evaluation Metrics:

To comprehensively evaluate model performance, the following metrics are considered:

1. Accuracy: Measures the overall correctness of classification.
2. Macro-Precision: Computes the precision for each class and takes the average.
3. Macro-Recall: Computes the recall for each class and takes the average.
4. Macro-F1 Score: Computes the F1 score for each class and takes the average.

Given the severe class imbalance, Macro-F1 Score is adopted as the primary evaluation metric, since it treats all classes equally and provides a fairer assessment of performance on minority categories.

Experimental Settings:

The multimodal feature extraction strategies for text, audio, and video are described in Section 2.2. All models are implemented in PyTorch. The optimizer is AdamW [15] with an initial learning rate of $2\times10^{-4,}$ and a decay factor of 0.1 is applied to the BERT encoder. The batch size is set to 32.

For MPAF-Net, a learning rate warm-up strategy (3 epochs) followed by a cosine annealing scheduler is adopted. All experiments are conducted on a single NVIDIA RTX 4060 GPU. Detailed hyperparameter settings are summarized in Table 2.

Table 2 Main hyperparameter settings

parameter	Baseline	MPAF-Net
Learning -rate	2e-4	2e-4
Batch Size	32	32
Optimizer	AdamW	AdamW
Epoch	100	100
Dropout	0.2	0.2
Adjust	Reduce LR On Plateau	Warmup + Cosine Annealing
Loss	Cross Entropy (Weighted)	Focal Loss(γ=2.0)

In addition, data augmentation techniques were applied to samples of minority emotion classes. For the audio modality, new samples were generated by adding background noise, adjusting pitch, and other transformations. For the video modality, augmentations such as frame rotation and cropping were applied to create additional samples. These strategies increased the number of minority-class samples and improved the model's ability to learn from underrepresented categories.

3.3 Experimental Results and Analysis

3.3.1 Baseline Model Performance Analysis

The results of the baseline model clearly reveal the limitations of simple fusion methods. As shown in Table 5, the baseline achieves a Macro-F1 score of only 40.13% on the test set. During training, the training loss continuously decreased, while the validation loss began to rise rapidly in the early stage, indicating severe overfitting.

The classification metrics in Table 3 show that the model completely fails to learn minority classes such as fear (fea), achieving an F1 score of 0. This confirms that simple concatenation-based fusion is insufficient for fine-grained emotion recognition under speaker-independent settings.

Table 3. Classification Metrics of Baseline Model

Emotion	Precision	Recall	F1-score	Support
Ang	0.2917	0.1795	0.2222	117
Exc	0.2247	0.2326	0.2286	86
Fea	0.0000	0.0000	0.0000	5
Fru	0.2694	0.2935	0.2810	201
Hap	0.1475	0.1475	0.1475	61
Neu	0.3838	0.4318	0.4064	176
Sad	0.4767	0.4227	0.4481	97
Sur	0.0476	0.1250	0.0690	8

3.3.2 MPAF-Net Performance Analysis

To evaluate the performance of MPAF-Net across emotion categories, Table 4 presents detailed classification metrics on the test set. The model demonstrates robust recognition ability for major emotion classes.

For relatively well-represented categories such as neutral (neu), sadness (sad), and anger (ang), the F1 scores exceed 0.50, with sadness reaching 0.5556, indicating that the multi-path fusion architecture effectively captures and distinguishes general multimodal features of core emotions.

The model also shows improved performance on minority classes. Unlike the baseline, which fails to learn rare classes (F1 = 0), MPAF-Net achieves an F1 score of 0.2667 for fear (fea) (8 samples) and 0.3571 for surprise (sur) (12 samples). This demonstrates that the integrated optimization strategies, particularly class-weighted Focal Loss, effectively guide the model to focus on and learn features from these rare samples.

Overall, the detailed analysis indicates that MPAF-Net not only improves overall performance but also achieves more balanced and comprehensive emotion recognition. It maintains high performance on majority classes while acquiring preliminary recognition capability for difficult and minority classes, which is critical for building robust real-world emotion recognition systems.

Furthermore, as shown in Figure 6, the training process is stable. The validation loss curve gradually stabilizes without sharp rebounds, indicating that the multi-level regularization strategy effectively mitigates overfitting. The performance gap between training and validation sets remains within a reasonable range.

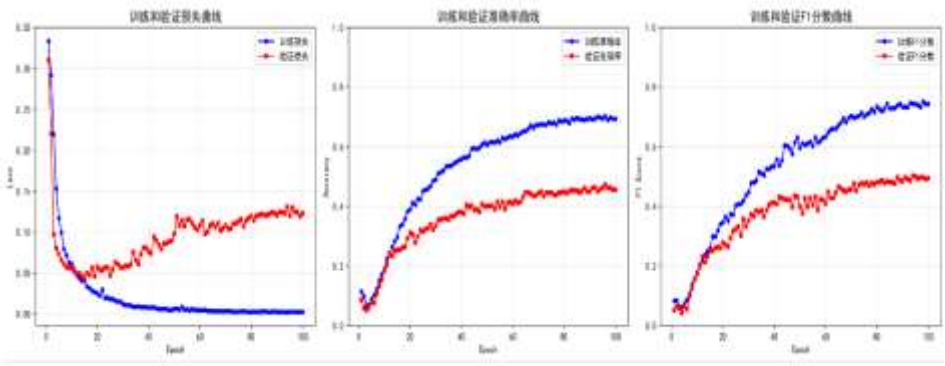

Figure 6 training loss

Table 4. Classification Metrics of MPAF-Net

Emotion	Precision	Recall	F1-score	Support
Ang	0.4225	0.6522	0.5128	92
Exc	0.5042	0.4444	0.4724	135
Fea	0.2857	0.2500	0.2667	8
Fru	0.4910	0.4740	0.4824	173
Hap	0.2857	0.3099	0.2973	71
Neu	0.5789	0.5156	0.5455	192
Sad	0.6044	0.5140	0.5556	107
Sur	0.3125	0.4167	0.3571	12

To provide a more comprehensive evaluation, MPAF-Net was further compared with three representative fusion methods—Early Fusion LSTM (EF-LSTM), Late Fusion LSTM (LF-LSTM), and Tensor Fusion Network (TFN)—on the same test set. Only accuracy and Macro-F1 scores on the test set are reported for these comparisons.

Table 5. Comparative Experimental Results

Model	Validation Acc(%)	Validation Macro-F1 (%)	Test Acc (%)	Test Macro-F1 (%)
Baseline	30.19	30.03	38..48	40.13
EF-LSTM			35.12	29.29
LF-LSTM			40.58	36.66
TFN			44.30	44.93
MPAF-Net	**51.59**	**54.14**	**50.13**	**47.42**

As shown in Table 5, MPAF-Net demonstrates excellent performance on the validation set, achieving a Macro-F1 score of 54.14%, which is significantly higher than the baseline. This highlights the superiority of combining the multi-path feature fusion architecture with advanced training strategies. The peak performance on the validation set indicates that the potential upper bound of the proposed method is considerably higher than that of the baseline.

From the comparative results, the following conclusions can be drawn:

1. Hybrid fusion strategies exhibit clear advantages in complex interaction tasks. EF-LSTM, which performs early fusion without considering modality interactions, achieves the worst performance, with a test set Macro-F1 of only 29.29%. LF-LSTM preserves more information by processing each modality independently, outperforming EF-LSTM, but is still limited by information loss during decision-level fusion. TFN and MPAF-Net, representing hybrid fusion approaches, significantly outperform the first two models, demonstrating the importance of modeling inter-modal interactions at the feature level.
2. MPAF-Net achieves the best performance among all compared methods. It attains the highest test set accuracy (50.13%) and Macro-F1 score (47.42%). Compared with the strong TFN model (44.93%), MPAF-Net achieves an improvement of approximately 2.5 percentage points. This advantage mainly arises from the multi-path fusion architecture: by feeding both original and interaction features in parallel to the classifier, the model leverages independent modality information while capturing deep inter-modal interactions, resulting in superior performance in speaker-independent multimodal emotion recognition tasks.

To further analyze model behavior, Figure 7 presents the confusion matrix of MPAF-Net on the test set. Compared with the baseline, MPAF-Net shows improved discrimination for most categories. For example, F1 scores for neutral (neu) and excitement (exc) increase significantly.

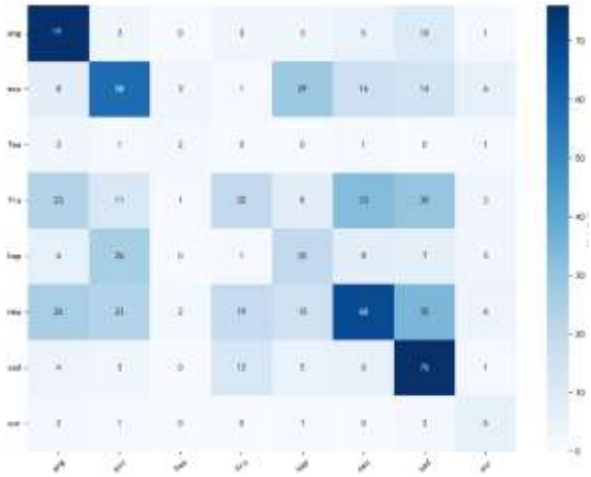

Figure 7 confusion matrix

However, some challenges remain. The model occasionally confuses semantically similar emotions, misclassifying certain samples. In addition, the learning performance on extremely rare classes, such as fear (fea), remains limited. This suggests future research directions, including enhancing the representation of fine-grained features or incorporating external knowledge to improve the recognition of ambiguous emotions.

In summary, the comparative experiments strongly demonstrate that the proposed MPAF-Net, centered on multi-path attention fusion and complemented with a systematic set of training optimization strategies, achieves superior performance over classical fusion methods in challenging speaker-independent and multi-class imbalanced emotion recognition tasks.

4 Conclusion and Future Work

4.1 Conclusion

This study systematically addresses the core challenges in multimodal emotion recognition (MER), including model overfitting, data imbalance, and insufficient inter-modal feature fusion. A simple feature concatenation baseline model was first constructed and analyzed to quantify its performance limitations under a speaker-independent setting.

Based on this, the MPAF-Net model was proposed, incorporating a multi-path attention fusion architecture along with a comprehensive training optimization strategy. Through parallel cross-modal attention mechanisms, the model preserves original unimodal features while effectively generating deep inter-modal interaction features. Coupled with model lightweight design, multi-level regularization, class-weighted Focal Loss, and learning rate warm-up with cosine annealing, the model achieves significant performance improvements on the IEMOCAP dataset for multi-class classification.

Experimental results show that, compared with the baseline and several classical fusion methods (e.g., EF-LSTM, LF-LSTM, TFN), MPAF-Net attains a Macro-F1 score of 47.42%, demonstrating strong overall performance and effectively mitigating overfitting.

The main contributions of this study are:

1. Validation of the multi-path feature fusion architecture: The strategy of retaining original features while parallelly generating interaction features proves effective for complex multimodal tasks.
2. Provision of a systematic and reproducible optimization framework: The study illustrates how combining multiple training techniques can effectively address overfitting and data imbalance challenges.

4.2 Limitations and Future Work

Despite the promising results, several limitations remain, which point to future research directions:

1. Recognition of extremely rare classes remains limited: Future work may explore advanced data augmentation techniques, such as Generative Adversarial Networks (GANs) or Diffusion Models, to generate more high-quality training samples for these rare categories.
2. Generalization gap still exists: Future studies could investigate Domain Adaptation or Meta-Learning techniques to improve the model's transferability from one speaker (or a group of speakers) to previously unseen speakers.

Funding: This work was supported by the Tianjin Key R&D Program – Institute-City Collaborative Projects (No. 24YFYSHZ00090, 23YFYSHZ00280) and the Tianjin Municipal Education Commission Scientific Research Program – Key Projects in Natural Science (No. 2022ZD032, 2022ZD026).

References

1. Poria, S., Cambria, E., Hazarika, D., & Majumder, N. (2022). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 82, 98-125.
2. Lotfian, R., & Busso, C. (2019). A speaker-exclusive distillation approach for speaker-independent acoustic emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(11), 1836-1848.
3. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.
4. Li, B., Liu, T., & Wang, G. (2022). A survey on deep learning for multimodal data fusion. Neural Computation, 34(10), 2215-2277.
5. Busso, C., Bulut, M., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.
6. Lin, T. Y., Goyal, P., Girick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).
7. Sun, Z., Yu, L., Shi, J., & Lu, G. (2020). Lf-mmtm: A memory-based multi-task model for multimodal language fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 8968-8975.
8. Zadeh A, Liang P, Poria S, et al. Multi-attention recurrent network for human communication comprehension // Proceedings of the 32th AAAI Conference on Artificial Intelligence. Palo Alto, 2018: 5642–5649.
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
10. Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6558-6569).
11. Mengara A G M , Moon Y , He M E .CAG-MoE: Multimodal Emotion Recognition with Cross-Attention Gated Mixture of Experts[J]. 2025.
12. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1103-1114).
13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
14. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9 (8), 1735-1780.
15. Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.