

Article

Not peer-reviewed version

---

# LPGNet: A Lightweight Network with Parallel Attention and Gated Fusion for Multimodal Emotion Recognition

---

[Zhining He](#) \* and [Yang Xiao](#)

Posted Date: 11 August 2025

doi: 10.20944/preprints202508.0678.v1

Keywords: multimodal emotion recognition; parallel attention; dual-gated fusion; speaker-independent; lightweight network



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

# LPGNet: A Lightweight Network with Parallel Attention and Gated Fusion for Multimodal Emotion Recognition

Zhining He <sup>1,\*</sup> and Yang Xiao <sup>2</sup>

<sup>1</sup> Guangzhou University, Guangzhou, China  
<sup>2</sup> The University of Melbourne, Melbourne, Australia  
\* Correspondence: hezhining0624@163.com;

## Abstract

Emotion recognition in conversations (ERC) aims to predict the emotional state of each utterance by using multiple input types, such as text and audio. While Transformer-based models have shown strong performance in this task, they often face two major issues: high computational cost and heavy dependence on speaker information. These problems reduce their ability to generalize in real-world conversations. To solve these challenges, we propose LPGNet, a Lightweight network with Parallel attention and Gated fusion for multimodal ERC. The main part of LPGNet is the Lightweight Parallel Interaction Attention (LPIA) module. This module replaces traditional stacked Transformer layers with parallel dot-product attention, which can model both within-modality and between-modality relationships more efficiently. To improve emotional feature learning, LPGNet also uses a dual-gated fusion method. This method filters and combines features from different input types in a flexible and dynamic way. In addition, LPGNet removes speaker embeddings completely, which allows the model to work independently of speaker identity. Experiments on the IEMOCAP dataset show that LPGNet reaches over 87% accuracy and F1-score in 4-class emotion classification. It outperforms strong baseline models while using fewer parameters and showing better generalization across speakers.

**Keywords:** multimodal emotion recognition; parallel attention; dual-gated fusion; speaker-independent; lightweight network

## 1. Introduction

Emotion recognition in conversations (ERC) [1,2] focuses on identifying the emotional state of each utterance by using multiple types of input, such as text, audio, and visual signals. It plays an important role in human-computer interaction (HCI) [3,4], especially in building systems that can understand and respond to users' emotions. ERC is a key part of many real-world applications, including empathetic dialogue systems, emotion-aware virtual assistants, and mental health support tools. These systems are widely used in areas like intelligent voice assistants, customer service robots, and telemedicine. Compared to traditional emotion recognition based on a single input type, ERC brings more challenges. This is because it requires understanding context over time, handling frequent speaker changes, and processing emotional signals across different modalities. These challenges make ERC a complex but important problem in affective computing.

Recent ERC models based on Transformer [5] and graph neural network (GNN) [6–8] architectures have shown strong performance. However, they often face two key challenges. First, deep sequential layers lead to high computational cost, which limits deployment in real-time systems. Second, many models depend heavily on speaker identity features, making them less effective in open-domain or anonymous settings, where speaker information is missing or unavailable. These limitations highlight the need for efficient and speaker-independent ERC models.

To improve multimodal fusion in ERC, several attention-based approaches have been developed. These methods aim to better integrate features from different input types, such as text and audio. For example, Tang et al. [9] introduced an audio-text interactive attention mechanism that strengthens cross-modal understanding. Zhao et al. [10] proposed collaborative attention to align emotional information across pre-trained features. Maji et al. [11] designed a cross-modal Transformer block to capture both semantic and temporal dependencies across modalities. Although these techniques have demonstrated good performance, they often come with a high cost. Their architectures are usually deep and complex, which increases both training and inference time. As a result, these models are not well-suited for real-time or resource-constrained environments, such as mobile devices.

In addition to attention-based models, graph neural networks (GNNs) have also been applied to ERC. Methods such as DialogueGCN [6] and MMGCN [12] aim to capture speaker interactions by constructing dialogue graphs. These models can represent complex conversational structures but rely heavily on speaker identity labels and manually crafted graphs. This dependency reduces their flexibility and limits their usefulness in open-domain applications, where such information may be unavailable or inconsistent.

These limitations become especially important when considering real-world ERC applications. In practical settings, systems must often deal with unknown, anonymous, or changing speakers [13–15]. They also need to perform emotion recognition in a zero-shot manner, without prior training on speaker-specific data. For instance, voice assistants must understand a user's emotions during their first interaction, without relying on stored speaker profiles [16]. Similarly, customer service bots interact with a wide range of users, many of whom are anonymous or change frequently. In such cases, speaker embeddings offer little benefit. Multi-user devices, such as smart home terminals, require models to handle frequent speaker switching without identity tracking.

Recent ERC methods based on large language models (LLMs) [17–19] also reveal inconsistencies in how speaker information is handled. InstructERC [20] treats ERC as a generation task using emotion templates, but it does not explicitly model speaker roles. DialogueLLM [21] uses dialogue context and visual cues but omits speaker identity. BiosERC [22] goes further by showing that privacy constraints often prevent the use of identity labels, meaning that models must rely only on the dialogue context. These findings underline the importance of speaker-independent emotion modeling for real-world applications.

Alongside attention and graph-based techniques, self-distillation [23–25] has emerged as a promising solution for improving model performance without increasing complexity. Unlike traditional distillation methods that rely on an external teacher model, self-distillation enables a network to learn from its own intermediate features. For example, Li et al. [26] introduced a teacher-free feature distillation method in vision tasks, where a model reuses useful internal representations to boost its own learning. This strategy reduces both memory and computation costs, making it well-suited for lightweight deployment.

Based on these insights, we propose LPGNet, a lightweight and speaker-independent framework for multimodal emotion recognition in conversations. LPGNet is designed to address three major issues: high inference cost, reliance on speaker information, and the need for efficient deployment in practical scenarios.

Our key contributions are as follows:

- We design a Lightweight Parallel Interaction Attention (LPIA) module to replace stacked Transformer layers, allowing efficient modeling of intra- and inter-modal relationships.
- We introduce a dual-gated fusion strategy to refine and dynamically combine multimodal features.
- We remove speaker embeddings entirely, enabling better generalization to unseen speakers and anonymous environments.
- We apply a self-distillation mechanism to internally extract and reuse useful emotional knowledge without requiring external teacher models.

Extensive experiments on the IEMOCAP benchmark show that LPGNet achieves over 87% accuracy and F1-score for 4-class emotion classification. It not only outperforms strong baseline models but also shows better generalization to unknown speakers, all while using fewer parameters.

## 2. Related Work

### 2.1. Multimodal Fusion Strategies in Emotion Recognition

Multimodal emotion recognition (MER) [27] aims to integrate signals from different modalities such as text, speech, and vision. Early models relied on recurrent neural networks and memory-based fusion techniques. For example, CMN [28] used GRUs to model speaker history and applied attention to extract relevant emotion patterns from past utterances. ICON [29] extended this by introducing interactive memory modules to model affective influence between speakers. These methods successfully leveraged dialogue context but involved complex architectures, making them difficult to scale for large or real-time applications.

More recently, Transformer-based methods have become dominant. MER-HAN [30] uses a three-stage attention framework to model intra-modal, cross-modal, and global features. While this improves interpretability, it increases model complexity and may suffer from misalignment across modalities. To improve efficiency, models like SDT [5] apply parallel intra- and cross-modal attention combined with gated fusion and self-distillation. XMBT [31] introduces shared bottleneck tokens for cross-modal interaction, which helps reduce inference cost and allows flexible modality integration.

Despite performance gains, these methods still face key limitations. Many use deep sequential Transformers that are computationally expensive. Others lack dynamic weighting, treating all modalities equally—even when one modality carries more emotional information than others. To address this, LPGNet introduces a Lightweight Parallel Interaction Attention (LPIA) module to replace stacked layers and reduce computation. It also includes a dual-gated fusion mechanism to dynamically filter and combine multimodal features based on signal strength and context.

### 2.2. Speaker Dependency and Real-World Limitations

Several advanced MER models, including COSMIC [32] and DialogXL [33], incorporate speaker identity information or model speaker-role interactions through GNNs. While this improves performance in speaker-labeled datasets, it reduces flexibility in real-world conditions where speaker labels are unavailable or privacy constraints exist. For instance, voice assistants and multi-user devices often encounter new or anonymous speakers. In such cases, models depending on speaker embeddings may fail to generalize. Recent research has explored speaker-independent modeling. SIMR [34] attempts to remove identity bias by disentangling style and semantic content in non-verbal modalities. However, most systems still include speaker-level features or require predefined speaker roles, limiting their adaptability in open-domain dialogue.

Existing approaches do not fully eliminate speaker dependence. In contrast, LPGNet is designed to be speaker-independent by default. It removes all speaker-related inputs and structures, allowing the model to generalize across unknown speakers and dynamically changing roles—an essential requirement for scalable and privacy-aware systems.

### 2.3. Lightweight Learning with Self-Distillation

As emotion recognition moves toward real-time and mobile applications, efficiency has become a core concern. Traditional models like MER-HAN and SDT require deep Transformer stacks, which increase latency and memory usage. Knowledge distillation (KD) has been proposed to compress large models while maintaining accuracy. For example, DMD [35] uses cross-modal distillation to decompose and transfer shared and unique features. MT-PKDOT [36] applies multi-teacher distillation to guide student models using diverse sources. While effective, these strategies often require additional teacher models and increase training complexity. Some models, like SDT, use self-distillation to reuse internal knowledge, but they still rely on heavy architectures.

Few models combine self-distillation with lightweight structures. LPGNet adopts a teacher-free self-distillation approach within a compact design. It transfers knowledge across internal layers without needing extra networks, improving performance and efficiency. This allows LPGNet to remain accurate while significantly lowering inference cost and parameter size, making it suitable for real-world, resource-constrained deployment.

### 3. Proposed Method

#### 3.1. Task Definition

The input to the ERC task is a conversation composed of  $N$  utterances  $\{u_1, u_2, \dots, u_N\}$ , where each utterance  $u_i$  contains both textual and acoustic modalities. After feature extraction, the text and audio features are projected to a common latent space:

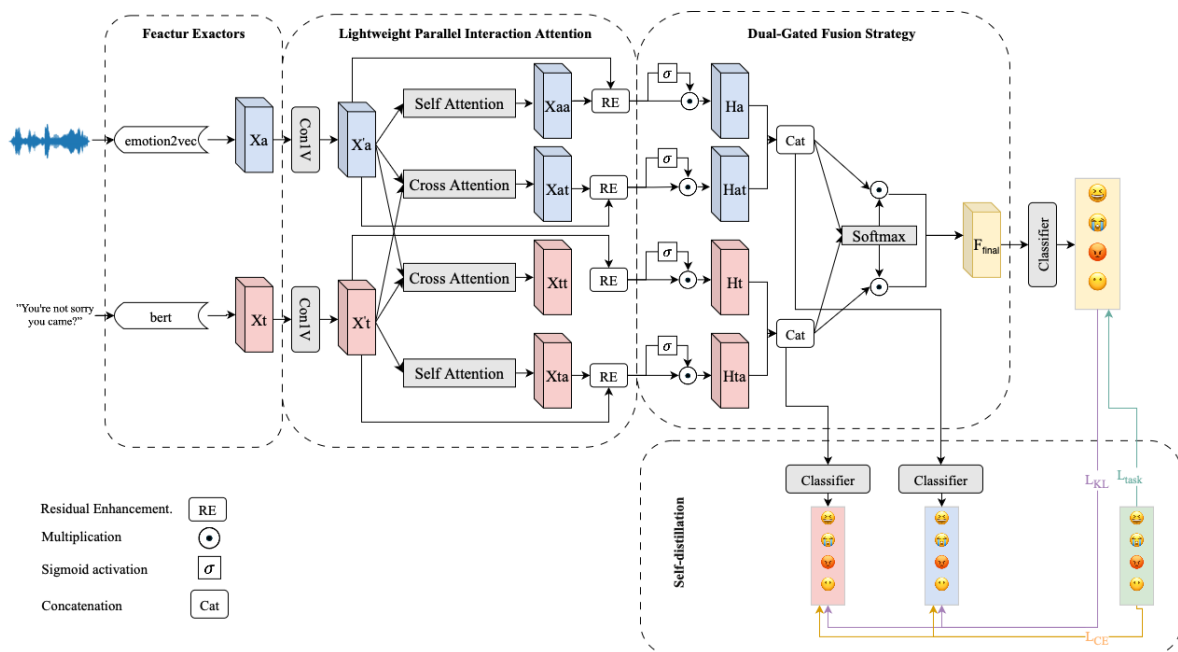
$$\mathbf{X}_t \in \mathbb{R}^{B \times U_i \times F}, \quad \mathbf{X}_a \in \mathbb{R}^{B \times U_i \times F} \quad (1)$$

Here,  $B$  is the batch size,  $U_i$  denotes the number of utterances, and  $F$  is the feature dimension. The goal of ERC is to predict an emotion label for each utterance  $u_i$  from a set of predefined emotion classes.

#### 3.2. Overview

Figure 1 provides an overview of the proposed **LPGNet** model. After extracting utterance-level acoustic and textual features, LPGNet consists of four major modules:

- **Lightweight Parallel Interaction Attention (LPia)** module for capturing both intra- and inter-modal interactions simultaneously;
- **Dual-Gated Fusion** module for refining unimodal representations and adaptively fusing multi-modal information;
- **Emotion Classifier** that predicts emotion labels from the final fused representation;
- **Self-Distillation mechanism** that aligns unimodal branches with the fused multimodal output using both hard and soft label supervision.



**Figure 1.** Overall architecture of LPGNet. After extracting utterance-level unimodal features, it consists of four key components: LPia Block, Dual Gated Fusion, Classifier, and Self-distillation



### 3.3. Lightweight Parallel Interaction Attention (LPIA)

To effectively model both intra- and inter-modal dependencies in a lightweight manner, we propose a parallel attention structure named **Lightweight Parallel Interaction Attention (LPIA)**. This module aims to fully leverage the complementary characteristics of acoustic and textual modalities through parallel attention mechanisms.

#### (1) Modality-wise Local Encoding via 1D Convolution.

We first apply a 1D convolution with kernel size 1 to both textual and acoustic input sequences. This operation serves two purposes: (1) to project modality-specific features into a unified latent space of dimension  $d$ ; and (2) to enable localized context modeling across adjacent utterances within each modality:

$$\mathbf{X}'_s = \text{Conv1D}(\mathbf{X}_s), \quad s \in \{t, a\} \quad (2)$$

where  $\mathbf{X}_s \in \mathbb{R}^{B \times U \times F_s}$  denotes the original feature sequence for modality  $s$  (text or audio),  $B$  is the batch size,  $U$  is the number of utterances, and  $F_s$  is the input feature dimension.

#### (2) Four-Way Parallel Attention Modeling.

To simultaneously capture intra- and inter-modal relationships, we define four parallel attention blocks based on scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (3)$$

The four interaction paths are:

- $\mathbf{X}_{tt} = \text{Attention}(\mathbf{X}'_t, \mathbf{X}'_t, \mathbf{X}'_t)$ : intra-text attention for modeling semantic continuity;
- $\mathbf{X}_{aa} = \text{Attention}(\mathbf{X}'_a, \mathbf{X}'_a, \mathbf{X}'_a)$ : intra-acoustic attention for modeling prosodic dynamics;
- $\mathbf{X}_{at} = \text{Attention}(\mathbf{X}'_a, \mathbf{X}'_t, \mathbf{X}'_t)$ : cross-modal audio-to-text interaction;
- $\mathbf{X}_{ta} = \text{Attention}(\mathbf{X}'_t, \mathbf{X}'_a, \mathbf{X}'_a)$ : cross-modal text-to-audio interaction.

This fully parallel structure avoids sequential stacking, reduces inference latency, and allows efficient modeling of bidirectional modality interactions.

#### (3) Mask-Aware Attention Computation.

To ensure robustness in variable-length conversations, we adopt a mask-aware attention mechanism. Let  $\mathbf{M} \in \{0, 1\}^{B \times U}$  denote a binary mask indicating valid utterances. It is expanded to  $\mathbf{M}' \in \{0, 1\}^{B \times U \times U}$  for masking attention scores:

$$\mathbf{S}_{i,j}^{(b)} = \begin{cases} \frac{\mathbf{Q}_i^{(b)} \cdot \mathbf{K}_j^{(b)}}{\sqrt{d}}, & \text{if } \mathbf{M}_{i,j}^{(b)} = 1 \\ -10^9, & \text{otherwise} \end{cases} \quad (4)$$

The masked scores are normalized by softmax:

$$\mathbf{A}^{(b)} = \text{Softmax}(\mathbf{S}^{(b)}) \quad (5)$$

This formulation ensures that attention is only paid to valid utterances, suppressing the influence of padding noise.

#### (4) Two-Stage Residual Enhancement.

After computing the attention output  $\hat{\mathbf{A}} \in \mathbb{R}^{B \times U \times d}$ , we apply two refinement stages:

- **1. Global Compression:** We first compress the global context using a lightweight convolutional block composed of  $1 \times 1$  convolution, batch normalization, and LeakyReLU activation:

$$\mathbf{A}_{\text{conv}} = \text{Conv2D}(\text{Mean}(\hat{\mathbf{A}})) \quad (6)$$

- **2. Position-wise FFN with Residual Fusion:** The compressed feature is broadcast back to the sequence dimension and fused with the original query and a position-wise feedforward network:

$$\text{FFN}(x) = \mathbf{W}_2(\text{Dropout}_1(\text{GELU}(\mathbf{W}_1 \cdot \text{LayerNorm}(x)))) \quad (7)$$

$$\mathbf{X}_{\text{final}} = \mathbf{Q} + \mathbf{A}_{\text{conv}} + \text{FFN}(\mathbf{A}_{\text{conv}}) \quad (8)$$

Here,  $\mathbf{W}_1 \in \mathbb{R}^{d \times d_{ff}}$  and  $\mathbf{W}_2 \in \mathbb{R}^{d_{ff} \times d}$  are learnable weight matrices, where  $d$  is the hidden dimension of the model and  $d_{ff}$  is the intermediate feedforward dimension (typically  $2d$  or  $4d$ ). The GELU function introduces smooth non-linearity, and Dropout is used for regularization.

This two-stage refinement integrates both global context and local non-linearity, enhancing feature expressiveness and gradient stability.

### 3.4. Dual-Gated Fusion Strategy

To enable both fine-grained intra-modal enhancement and flexible cross-modal integration, we introduce a hierarchical **Dual-Gated Fusion** mechanism, composed of two sequential stages: *Unimodal Gated Fusion* and *Multimodal Gated Fusion*.

#### (1) Unimodal Gated Fusion

Given the outputs from the intra-modal and inter-modal attention modules, we apply a learnable gating operation to each stream independently. Let  $\mathbf{H} \in \mathbb{R}^{B \times U \times d}$  be the hidden representation of an attention output. The refined representation is computed as:

$$\mathbf{Z} = \sigma(\mathbf{W}_g \mathbf{H}), \quad \mathbf{H}_{\text{gated}} = \mathbf{Z} \odot \mathbf{H} \quad (9)$$

where  $\mathbf{W}_g \in \mathbb{R}^{d \times d}$  is a learnable weight matrix and  $\sigma(\cdot)$  denotes the sigmoid activation function. This mechanism allows the model to emphasize informative features while suppressing irrelevant components.

For each modality, we gate the intra-modal and cross-modal branches separately:

$$\mathbf{H}_t = \text{GatedFusion}(\mathbf{X}_{\text{output}_{tt}}), \quad \mathbf{H}_{at} = \text{GatedFusion}(\mathbf{X}_{\text{output}_{at}}) \quad (10)$$

$$\mathbf{H}_a = \text{GatedFusion}(\mathbf{X}_{\text{output}_{aa}}), \quad \mathbf{H}_{ta} = \text{GatedFusion}(\mathbf{X}_{\text{output}_{ta}}) \quad (11)$$

Then, we concatenate and linearly project each modality-specific pair:

$$\mathbf{T}_{\text{fused}} = \text{Linear}([\mathbf{H}_t; \mathbf{H}_{at}]), \quad \mathbf{A}_{\text{fused}} = \text{Linear}([\mathbf{H}_a; \mathbf{H}_{ta}]) \quad (12)$$

#### (2) Multimodal Gated Fusion

The refined textual and acoustic features are then passed into a multimodal gating mechanism that adaptively assigns weights to each modality. Given the fused features  $\mathbf{T}_{\text{fused}}, \mathbf{A}_{\text{fused}} \in \mathbb{R}^{B \times U \times d}$ , the attention weights are computed as:

$$\alpha_T = \text{Softmax}(\mathbf{W} \mathbf{T}_{\text{fused}}), \quad \alpha_A = \text{Softmax}(\mathbf{W} \mathbf{A}_{\text{fused}}) \quad (13)$$

Let  $\mathcal{F} = [\mathbf{T}_{\text{fused}}, \mathbf{A}_{\text{fused}}] \in \mathbb{R}^{B \times U \times 2 \times d}$ , the final multimodal representation is computed as:

$$\mathbf{F}_{\text{final}} = \alpha_T \cdot \mathbf{T}_{\text{fused}} + \alpha_A \cdot \mathbf{A}_{\text{fused}} \quad (14)$$

This two-stage gating architecture enables the model to dynamically refine unimodal streams and adaptively integrate multimodal context based on conversational cues.

### 3.5. Emotion Classifier

To obtain the emotion distribution over  $C$  categories, the final multimodal representation  $\mathbf{F}_{\text{final}} \in \mathbb{R}^{N \times d}$  is passed through a fully connected layer followed by a softmax activation:

$$\mathbf{E} = \mathbf{W}_c \cdot \mathbf{F}_{\text{final}} + \mathbf{b}_c \in \mathbb{R}^{N \times C} \quad (15)$$

$$\hat{\mathbf{Y}} = \text{softmax}(\mathbf{E}) \quad (16)$$

Here,  $\mathbf{W}_c \in \mathbb{R}^{d \times C}$  and  $\mathbf{b}_c \in \mathbb{R}^C$  are trainable parameters. Let  $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1; \hat{\mathbf{y}}_2; \dots; \hat{\mathbf{y}}_N]$ , where  $\hat{\mathbf{y}}_i$  denotes the predicted emotion probability vector for utterance  $u_i$ . The predicted label is determined by  $\arg \max(\hat{\mathbf{y}}_i)$ .

The task loss is defined using the standard cross-entropy objective:

$$\mathcal{L}_{\text{task}} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C y_{i,j} \log \hat{y}_{i,j} \quad (17)$$

### 3.6. Self-Distillation

To further improve modal expressiveness, we adopt a self-distillation strategy that transfers knowledge from the full model (teacher) to each unimodal branch (student). Specifically, we treat the complete model as the teacher and introduce two student branches—textual ( $t$ ) and acoustic ( $a$ )—trained only during training.

For each modality  $m \in \{t, a\}$ , the student predicts the emotion distribution via:

$$\mathbf{E}^m = \mathbf{W}'_m \cdot \text{ReLU}(\mathbf{H}'_m) + \mathbf{b}'_m \in \mathbb{R}^{N \times C} \quad (18)$$

$$\hat{\mathbf{Y}}^m = \text{softmax}(\mathbf{E}^m), \quad \hat{\mathbf{Y}}^m_\tau = \text{softmax}(\mathbf{E}^m / \tau) \quad (19)$$

where  $\tau$  is the temperature for softening the output distribution. Higher values of  $\tau$  yield softer class probabilities.  $\mathbf{W}'_m \in \mathbb{R}^{d \times C}$  and  $\mathbf{b}'_m \in \mathbb{R}^C$  are learnable parameters.

Each student is optimized with two losses: The first one is the Cross-Entropy Loss (hard labels) as:

$$\mathcal{L}_{\text{CE}}^m = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C y_{i,j} \log \hat{y}_{i,j}^m \quad (20)$$

And the soft label loss is KL Divergence Loss as:

$$\mathcal{L}_{\text{KL}}^m = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C \hat{y}_{\tau,i,j}^m \log \frac{\hat{y}_{\tau,i,j}^m}{\hat{y}_{\tau,i,j}^{\text{teacher}}} \quad (21)$$

Finally, the Total Loss could be formulated as: The overall training loss combines task supervision and both hard/soft label distillation:

$$\mathcal{L} = \lambda_{\text{task}} \cdot \mathcal{L}_{\text{task}} + \lambda_{\text{CE}} \cdot \sum_m \mathcal{L}_{\text{CE}}^m + \lambda_{\text{KL}} \cdot \sum_m \mathcal{L}_{\text{KL}}^m \quad (22)$$

## 4. Experiments and Results

### 4.1. Dataset and Evaluation

We evaluate our model using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [37] dataset, a widely adopted benchmark in multimodal emotion recognition. This dataset consists of approximately 12 hours of audio-visual recordings collected from ten professional actors (five male



and five female) performing both scripted and improvised dialogues. Each utterance in the dataset is manually annotated with emotional labels across multiple modalities, including text, audio, video, and motion capture data.

In our study, we focus only on the transcribed text and speech signals, which are the most accessible and commonly used modalities in practical applications. Following standard practice in previous works, we define the emotion recognition in conversation (ERC) task as a four-class classification problem. The target emotion categories include *angry*, *happy*, *sad*, and *neutral*. The excited class, which shares similar characteristics with happy, is merged into it to simplify the classification task. After preprocessing and filtering, we obtain a total of 5,531 annotated utterances. The detailed emotion distribution across the dataset is presented in Table 1, providing insight into the class balance and evaluation.

Table 1. Emotion distribution in the IEMOCAP dataset.

Split	Happy	Sad	Neutral	Angry	Total
Train + Val	1262	828	1319	872	4290
Test	442	245	384	170	1241
Total	1704	1073	1703	1042	5531

For evaluation, we adopt average binary accuracy and F1-score as our main performance metrics, following the standard protocol used in prior work. Specifically, we compute results using a one-vs-all classification strategy for each emotion class. The final scores are then averaged across all classes. This approach ensures that performance is measured fairly, especially in the presence of class imbalance, by giving equal weight to each class regardless of its frequency in the dataset.

4.2. Feature Extraction

**Text Modality:** To represent the text modality, we use the BERT-base-uncased model as our text encoder. This model consists of 12 Transformer layers, each with 768-dimensional hidden states. We fine-tune BERT on the emotion recognition in conversation task using the dialogue data. For each utterance, we extract the final-layer embedding of the [CLS] token, which serves as a summary representation of the entire sentence.

**Acoustic Modality:** For the audio modality, we adopt Emotion2vec, a self-supervised acoustic representation model trained on 262 hours of emotional speech data. Emotion2vec is designed to learn both utterance-level and frame-level emotional features through a joint optimization process based on knowledge distillation. From this model, we extract 768-dimensional utterance-level embeddings, which provide rich and generalizable representations of the speech signal.

4.3. Implementation Details

Our model is implemented in PyTorch 1.8.1 and RTX 3090. We use the Adam optimizer with a learning rate of  $3e-4$ , a batch size of 32, and a hidden dimension of 768. The temperature parameter  $\tau$  for self-distillation is set to 1. We apply L2 weight decay of  $1 \times 10^{-5}$  and dropout with a rate of 0.1. Each model is trained for 150 epochs, and we report average results over 10 independent runs for robustness.

4.4. Results

Table 2 presents a detailed comparison between our proposed LPGNet and several state-of-the-art baselines on the IEMOCAP dataset under the 4-class emotion classification setting. Among existing methods, CFIA and MemoCMT show strong performance, with CFIA achieving 83.37% accuracy using the same BERT and Emotion2vec features as our model. However, our full model, LPGNet (Utterance-level), achieves the highest performance with 87.99% accuracy and 87.96% F1-score, representing a

relative improvement of over 4% compared to CFIA. These results highlight the effectiveness of our lightweight and speaker-independent design.

To ensure that this improvement is due to architectural innovation rather than feature quality, we build a minimal baseline using a simple linear classifier applied to the same BERT and Emotion2vec embeddings. This baseline reaches only 81.68% accuracy, which is 6.31% lower than our full model. This gap confirms that LPGNet’s performance gains are not solely due to better features, but come from its efficient multimodal fusion and context modeling capabilities.

We further evaluate a frame-level version of our model, called LPGNet (Frame), where both text tokens and acoustic frames are padded or truncated to a fixed length of 512. Despite using finer temporal resolution, LPGNet (Frame) still achieves strong results with 83.87% accuracy and 83.68% F1-score, outperforming most baselines. This demonstrates that our model can generalize across different input granularities without losing effectiveness. In summary, LPGNet consistently outperforms existing methods across both utterance- and frame-level settings. These results confirm its robustness, efficiency, and practical potential for real-world multimodal emotion recognition tasks.

**Table 2.** Overall performance comparison on IEMOCAP (4-class).

Model	Upstream(A)	Upstream(T)	ACC (%)	F1 (%)
GatedxLSTM [38]	CLAP	CLAP	–	75.97±1.38
MER-HAN [30]	Bi-LSTM	BERT	74.20	73.66
MMI-MMER [39]	Wav2Vec2	BERT	77.02	–
MGCMA [40]	Wav2Vec2	BERT	78.87	–
MEP [41]	openSMILE	BERT	80.18	80.01
Bi-GRU [42]	Acoustic model	GloVe	80.63	–
MemoCMT[43]	HuBERT	BERT	81.85	–
Linear[44]	emotion2vec	BERT	81.68	80.75
CFIA[45]	emotion2vec	BERT	83.37	–
<b>LPGNet(Frame)</b>	emotion2vec	BERT	<b>83.87</b>	<b>83.68</b>
<b>LPGNet(Utterance)</b>	emotion2vec	BERT	<b>87.99</b>	<b>87.96</b>

#### 4.5. Ablation Study

**Effectiveness of LPIA Components and Modal Interactions.** To evaluate the role of each component within the LPIA module, we perform a series of ablation studies, as summarized in Table 3.

We first examine the impact of removing the intra-modal and inter-modal attention blocks. The results show that eliminating either block leads to a clear performance drop. Specifically, removing inter-modal attention causes a 1.77% decrease in accuracy, indicating that modeling cross-modal emotional dependencies is crucial for effective fusion. On the other hand, removing intra-modal attention results in a 0.96% drop, showing that capturing temporal and contextual patterns within each modality also contributes meaningfully to performance. Together, these findings suggest that the two attention types play complementary roles—with intra-modal attention focusing on modality-specific context, and inter-modal attention capturing relationships across modalities.

In addition, we analyze the effect of removing the position-wise feedforward network (FFN), which follows the attention blocks. Although FFN does not directly model alignment, it applies non-linear transformations that refine and enrich the learned representations. Removing this component leads to a 1.12% drop in F1-score, confirming that the FFN is essential for final feature integration and improved discriminative power.

These results collectively validate the design of the LPIA module, showing that each sub-component contributes to the model’s ability to understand and fuse multimodal emotional cues effectively.

**Effect of Dual-Gated Fusion.** To further evaluate the effectiveness of our proposed fusion strategy, we conduct an ablation study by removing the Dual-Gated Fusion module from the model. As reported

in Table 3, this modification leads to a performance drop in F1-score from 87.96% to 86.59%. This confirms the positive impact of our gating mechanism on multimodal fusion.

The Dual-Gated Fusion module contains two key components: a unimodal-level gate and a multimodal-level gate. The unimodal gate filters out irrelevant or noisy features within each modality before fusion, helping the model focus on emotionally relevant content. The multimodal gate then adaptively assigns importance to each modality during the fusion process, allowing the model to balance their contributions based on context.

When this module is removed, the model loses its ability to regulate feature contributions both before and after fusion. This leads to less informative and more noisy multimodal representations. In contrast, the full LPGNet design benefits from dynamic feature calibration at both levels, which is especially important when modalities vary in quality or signal strength. These results demonstrate that combining attention-based interaction modeling (through LPIA) with gated control mechanisms leads to stronger and more robust emotional understanding. The two-stage gating design plays a crucial role in enhancing representation quality, particularly under real-world conditions involving modality imbalance or noise.

Table 3. Ablation results of LPGNet.

	ACC (%)	F1 (%)
<b>LPGNet</b>	<b>87.99</b>	<b>87.96</b>
Modality		
Text	81.39	81.19
Audio	84.53	84.55
LPIA Block		
w/o inter attention	86.22	86.26
w/o intra attention	87.03	86.95
w/o position-wise FFN	86.87	86.75
w/o dual-gated fusion	86.62	86.59

**Effectiveness of Feature Representation and Multimodal Fusion.** To assess the quality of unimodal feature representations in LPGNet, we compare its performance with existing unimodal baselines. As shown in Table 4 and Table 5, our model consistently outperforms other text-only and audio-only approaches when using the same feature extractors—Emotion2vec for audio and BERT for text. These results confirm that our chosen encoders are effective, and that LPGNet is able to fully leverage them through its attention-based and fusion-oriented architecture.

Importantly, LPGNet maintains strong performance even in unimodal settings. This demonstrates that its architecture can enhance emotion-relevant signals and suppress irrelevant noise, even when limited to a single modality. Compared with simpler baselines such as linear classifiers or prior multimodal models adapted for unimodal input, LPGNet achieves noticeable accuracy gains. This suggests that the model’s structural advantages go beyond feature quality and contribute directly to its discriminative power.

Furthermore, Table 3 illustrates the added value of integrating multiple modalities. Multimodal fusion significantly improves performance over each unimodal variant, indicating strong cross-modal complementarity. Acoustic and textual modalities contribute different but synergistic cues—such as prosody and semantics—which, when fused, offer a more holistic emotional representation. These findings highlight the necessity of both effective unimodal encoders and carefully designed multimodal integration mechanisms for robust emotion recognition in conversation.

**Table 4.** Comparison with prior unimodal (audio-only) methods.

Audio Model	Downstream	ACC (%)
Wav2Vec	EmotionNAS[46]	69.1
Hubert	MSTR[47]	71.6
emotion2vec	Linear[44]	74.68
emotion2vec	CFIA-Net[45]	76.42
<b>emotion2vec</b>	<b>LPGNet</b>	<b>84.53</b>

**Table 5.** Comparison with prior unimodal (text-only) methods.

Text Model	Downstream	ACC (%)
bert	MER-HAN[30]	69.54
bert	CwGHP[48]	79.01
bert	MemoCMT[43]	80.66
<b>bert (Ours)</b>	<b>LPGNet</b>	<b>81.39</b>

**Effect of Self-Distillation Coefficients.** To assess the quality of unimodal feature representations in LPGNet, we compare its performance with existing unimodal baselines. As shown in Table 4 and Table 5, our model consistently outperforms other text-only and audio-only approaches when using the same feature extractor, Emotion2vec for audio and BERT for text. These results confirm that our chosen encoders are effective and that LPGNet can leverage them through its attention-based and fusion-oriented architecture.

Meanwhile, we observe that a lower KL-divergence weight  $\lambda_{KL} = 0.3$  yields better generalization. Since LPGNet removes speaker embeddings and focuses on speaker-independent modeling, the soft label distributions may exhibit greater inter-sample variability. Down-weighting KL divergence mitigates the risk of overfitting to potentially noisy or over-smoothed soft labels. On the other hand, keeping  $\lambda_{CE} = 1.0$  ensures that unimodal branches receive sufficient intermediate supervision during distillation.

**Table 6.** Effect of distillation loss weights on performance.

$\lambda_{task}$	$\lambda_{CE}$	$\lambda_{KL}$	ACC (%)	F1 (%)
1.0	1.0	1.0	87.35	87.38
1.0	0.0	1.0	86.36	86.53
1.0	1.0	0.0	87.27	87.26
1.5	1.0	0.0	87.59	87.58
<b>1.5</b>	<b>1.0</b>	<b>0.3</b>	<b>87.99</b>	<b>87.96</b>

**Table 7.** Performance comparison with SDT (4-class and 6-class). SDT\* uses our feature extractors.

	Happy	Sad	Neutral	Angry	Excited	Frustrated	Overall
<b>LPGNet (4-class)</b>	<b>92.06</b>	85.38	<b>83.99</b>	86.75	–	–	<b>87.83</b>
<b>SDT* (4-class)</b>	89.72	<b>86.35</b>	82.32	<b>88.15</b>	–	–	86.62
<b>LPGNet (6-class)</b>	<b>51.3</b>	<b>79.1</b>	<b>73.99</b>	72.19	<b>77.83</b>	<b>68.59</b>	<b>71.78</b>
<b>SDT* (6-class)</b>	38.59	78.09	69.52	<b>73.53</b>	70.63	65.13	67.47

In addition, Table 3 shows the performance improvement from combining multiple modalities. The results indicate that multimodal fusion clearly outperforms unimodal variants, reflecting the complementary nature of acoustic and textual information. For example, speech can capture emotional tones and prosody, while text contributes semantic and contextual cues. LPGNet’s fusion design

enables these different signals to be integrated effectively, resulting in a more complete emotional understanding.

Together, these findings highlight the importance of both strong unimodal encoders and carefully designed multimodal integration. The success of LPGNet comes not only from its input features but also from its ability to combine them in a way that enhances emotional signal learning.

## 5. Conclusions

In this paper, we proposed LPGNet, a lightweight and speaker-independent model for multimodal emotion recognition in conversations (ERC). To reduce computational cost and avoid reliance on speaker identity, LPGNet introduces a Lightweight Parallel Interaction Attention (LPIA) module that efficiently models both intra- and inter-modal dependencies. A dual-gated fusion mechanism further refines unimodal features and adaptively integrates them. We also apply self-distillation to enhance unimodal learning by transferring knowledge from fused features. Experiments on the IEMOCAP dataset show that LPGNet achieves state-of-the-art performance with fewer parameters. Ablation studies confirm the importance of each component, and unimodal evaluations show strong feature representation.

For future work, we plan to extend LPGNet to handle missing modalities, making it more robust in real-world settings. In addition, we aim to explore sparse attention and low-rank approximations in the LPIA module to further reduce latency and support real-time deployment on resource-limited devices.

## References

1. Poria, S.; Majumder, N.; Mihalcea, R.; Hovy, E. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access* **2019**, *7*, 100943–100953.
2. Qin, X.; Wu, Z.; Zhang, T.; Li, Y.; Luan, J.; Wang, B.; Wang, L.; Cui, J. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2023, Vol. 37, pp. 13492–13500.
3. Ramakrishnan, S.; El Emery, I.M. Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems* **2013**, *52*, 1467–1478.
4. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine* **2001**, *18*, 32–80.
5. Ma, H.; Wang, J.; Lin, H.; Zhang, B.; Zhang, Y.; Xu, B. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia* **2023**, *26*, 776–788.
6. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. Dialoguecn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540* **2019**.
7. Chen, F.; Shao, J.; Zhu, S.; Shen, H.T. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10761–10770.
8. Song, R.; Giunchiglia, F.; Shi, L.; Shen, Q.; Xu, H. SUNET: Speaker-utterance interaction graph neural network for emotion recognition in conversations. *Engineering Applications of Artificial Intelligence* **2023**, *123*, 106315.
9. Tang, Y.; Hu, Y.; He, L.; Huang, H. A bimodal network based on audio–text–interactional-attention with arcfac loss for speech emotion recognition. *Speech Communication* **2022**, *143*, 21–32.
10. Zhao, X.; Curtis, A. Bayesian inversion, uncertainty analysis and interrogation using boosting variational inference. *arXiv preprint arXiv:2312.17646* **2023**.
11. Maji, B.; Swain, M.; Guha, R.; Routray, A. Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.
12. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779* **2021**.
13. Peng, T.; Xiao, Y. Dark experience for incremental keyword spotting. In Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.



14. Xiao, Y.; Peng, T.; Das, R.K.; Hu, Y.; Zhuang, H. AnalyticKWS: towards exemplar-free analytic class incremental learning for small-footprint keyword spotting. *arXiv preprint:2505.11817* **2025**.
15. Xiao, Y.; Yin, H.; Bai, J.; Das, R.K. Mixstyle based domain generalization for sound event detection with heterogeneous training data. *arXiv preprint:2407.03654* **2024**.
16. Xiao, Y.; Das, R.K. Where's That Voice Coming? Continual Learning for Sound Source Localization. *arXiv preprint:2407.03661* **2024**.
17. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* **2024**, *15*, 1–45.
18. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology* **2023**.
19. Xiao, Y.; Das, R.K. WildDESED: an LLM-powered dataset for wild domestic environment sound event detection system. *arXiv preprint:2407.03656* **2024**.
20. Wu, Z.; Gong, Z.; Ai, L.; Shi, P.; Donbekci, K.; Hirschberg, J. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances. *arXiv preprint arXiv:2407.21315* **2024**.
21. Zhang, Y.; Wang, M.; Wu, Y.; Tiwari, P.; Li, Q.; Wang, B.; Qin, J. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Neural Networks* **2025**, p. 107901.
22. Xue, J.; Nguyen, M.P.; Matheny, B.; Nguyen, L.M. Bioserc: Integrating biography speakers supported by llms for erc tasks. In Proceedings of the International Conference on Artificial Neural Networks. Springer, 2024, pp. 277–292.
23. Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; Ma, K. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 3713–3722.
24. Zhang, L.; Bao, C.; Ma, K. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**, *44*, 4388–4403.
25. Xiao, Y.; Das, R.K. Ucil: An unsupervised class incremental learning approach for sound event detection. *arXiv preprint:2407.03657* **2024**.
26. Li, L.; Liang, S.N.; Yang, Y.; Jin, Z. Teacher-free distillation via regularizing intermediate representation. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022, pp. 01–06.
27. Zhang, S.; Yang, Y.; Chen, C.; Zhang, X.; Leng, Q.; Zhao, X. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications* **2024**, *237*, 121692.
28. Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.P.; Zimmermann, R. Conversational memory network for emotion recognition in dyadic dialogue videos. In Proceedings of the Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting, 2018, Vol. 2018, p. 2122.
29. Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. Icon: Interactive conversational memory network for multimodal emotion detection. In Proceedings of the Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 2594–2604.
30. Zhang, S.; Yang, Y.; Chen, C.; Liu, R.; Tao, X.; Guo, W.; Xu, Y.; Zhao, X. Multimodal emotion recognition based on audio and text by using hybrid attention networks. *Biomedical Signal Processing and Control* **2023**, *85*, 105052.
31. Nguyen, D.K.; Lim, E.; Kim, S.H.; Yang, H.J.; Kim, S. Enhanced Emotion Recognition Through Dynamic Restrained Adaptive Loss and Extended Multimodal Bottleneck Transformer. *Applied Sciences* **2025**, *15*, 2862.
32. Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; Poria, S. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795* **2020**.
33. Shen, W.; Chen, J.; Quan, X.; Xie, Z. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2021, Vol. 35, pp. 13789–13797.
34. Kuhlen, A.K.; Rahman, R.A. Mental chronometry of speaking in dialogue: Semantic interference turns into facilitation. *Cognition* **2022**, *219*, 104962.
35. Yin, Y.; Zhu, B.; Chen, J.; Cheng, L.; Jiang, Y.G. Mix-dann and dynamic-modal-distillation for video domain adaptation. In Proceedings of the Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3224–3233.

36. Aslam, M.H.; Pedersoli, M.; Koerich, A.L.; Granger, E. Multi Teacher Privileged Knowledge Distillation for Multimodal Expression Recognition. *arXiv preprint arXiv:2408.09035* **2024**.
37. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* **2008**, *42*, 335–359.
38. Li, Y.; Sun, Q.; Murthy, S.M.K.; Alturki, E.; Schuller, B.W. GatedxLSTM: A Multimodal Affective Computing Approach for Emotion Recognition in Conversations. *arXiv preprint arXiv:2503.20919* **2025**.
39. Wang, J.; Li, N.; Zhang, L.; Shan, L. Emotion Recognition for Multimodal Information Interaction. In Proceedings of the 2025 International Conference on Intelligent Systems and Computational Networks (ICISCN). IEEE, 2025, pp. 1–7.
40. Wang, X.; Zhao, S.; Sun, H.; Wang, H.; Zhou, J.; Qin, Y. Enhancing multimodal emotion recognition through multi-granularity cross-modal alignment. In Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.
41. Shi, X.; Li, X.; Toda, T. Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation. In Proceedings of the Proc. interspeech, 2023, Vol. 2023, pp. 765–769.
42. Adeel, M.; Tao, Z.Y. Enhancing speech emotion recognition in urdu using bi-gru networks: An in-depth analysis of acoustic features and model interpretability. In Proceedings of the 2024 IEEE International Conference on Industrial Technology (ICIT). IEEE, 2024, pp. 1–6.
43. Khan, M.; Tran, P.N.; Pham, N.T.; El Saddik, A.; Othmani, A. MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific reports* **2025**, *15*, 5473.
44. Ma, Z.; Zheng, Z.; Ye, J.; Li, J.; Gao, Z.; Zhang, S.; Chen, X. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185* **2023**.
45. Hu, Y.; Yang, H.; Huang, H.; He, L. Cross-modal Features Interaction-and-Aggregation Network with Self-consistency Training for Speech Emotion Recognition [C]. In Proceedings of the Proc. Interspeech 2024, 2024, pp. 2335–2339.
46. Sun, H.; Lian, Z.; Liu, B.; Li, Y.; Sun, L.; Cai, C.; Tao, J.; Wang, M.; Cheng, Y. EmotionNAS: Two-stream neural architecture search for speech emotion recognition. *arXiv preprint arXiv:2203.13617* **2022**.
47. Li, Z.; Xing, X.; Fang, Y.; Zhang, W.; Fan, H.; Xu, X. Multi-scale temporal transformer for speech emotion recognition. *arXiv preprint arXiv:2410.00390* **2024**.
48. Chauhan, K.; Sharma, K.K.; Varma, T. Multimodal Emotion Recognition Using Contextualized Audio Information and Ground Transcripts on Multiple Datasets. *Arabian Journal for Science and Engineering* **2024**, *49*, 11871–11881.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.