Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TAC-Net:Triple Attention Contrastive Network for Speech Complex Emotion Recognition in Real-Scene

Hankiz Yilahun
XinJiang University
Chaobo Song
107552304131@stu.xju.edu.cn

XinJiang University
Askar Hamdulla
XinJiang University

Research Article

Keywords: Speech Emotion Recognition, Multi-Label Classification, Attention Mechanism, Contrastive Learning

Posted Date: February 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8807080/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TAC-Net:Triple Attention Contrastive Network for Speech Complex Emotion Recognition in Real-Scene

Hankiz Yilahun[1†], Chaobo Song[1*], Askar Hamdulla[2†]

[1*]School of Computer Science and Technology, XinJiang University, Urumuqi, 830046, XinJiang, China.
[2]School of Intelligence Science and Technology, XinJiang University, Urumuqi, 830017, XinJiang, China.

*Corresponding author(s). E-mail(s): 107552304131@stu.xju.edu.cn;
Contributing authors: hansumuruh@xju.edu.cn; askar@xju.edu.cn;
†These authors contributed equally to this work.

Abstract

Speech Emotion Recognition (SER) in real-world scenarios remains a challenging task due to uncontrolled acoustic interference and the co-occurrence of complex emotions. Existing single-label approaches often struggle to capture fine-grained emotional semantics while suppressing noise, leading to significant performance degradation in complex environments. To address this, we propose a Triple Attention Contrastive Network (TAC-Net), designed to enhance model robustness through multi-dimensional feature decoupling and reconstruction. First, we introduce a Label-wise Attention Module comprising three collaborative networks: Multi-label Attention (MLA) explicitly models the latent dependencies among emotion tags; Local-Global Attention (LGA) locates key emotional frames temporally while capturing long-range context; and Time-Frequency Attention (TFA) focuses on spectral energy distributions to mitigate channel distortion. Furthermore, a Contrastive Reconstruction-based Fusion Module is incorporated to align heterogeneous features in the latent space via contrastive learning, effectively filtering redundant noise while preserving critical emotional information. Experimental results on the large-scale M^{3}ED and CMU-MOSEI datasets demonstrate that TAC-Net not only significantly outperforms existing unimodal baselines but also achieves competitive performance comparable to multimodal methods in handling multi-label and noisy data.

Keywords: Speech Emotion Recognition; Multi-Label Classification; Attention Mechanism; Contrastive Learning

1 Introduction

As artificial intelligence progressively transitions from perceptual to cognitive intelligence, endowing machines with the capability to recognize, comprehend, and even empathize with human emotions has become a central vision in the field of Human-Computer Interaction (HCI) [16]. Within the myriad ways humans interact, speech serves not merely as a carrier of semantic information but as the most natural and intuitive channel for conveying emotional states, intent tendencies, and personality traits [20]. Consequently, Speech Emotion Recognition (SER) has demonstrated immense social value in real-world scenarios such as mental health monitoring [8], emotional soothing in intelligent customer service [26], and anthropomorphic dialogue systems [19].

Over the past decade, driven by the flourishing development of deep learning technologies, SER research has achieved significant breakthroughs in controlled laboratory environments (e.g., the IEMOCAP and EMO-DB datasets) [2, 3]. However, the academic community has increasingly recognized that when models trained in ideal settings are directly migrated to real-world scenarios ("in-the-wild"), their performance often suffers a precipitous decline [28]. This challenge of real-world deployment is primarily attributed to two fundamental obstacles inherent in realistic environments.

The primary challenge stems from the uncontrolled nature of the acoustic environment. In-the-wild speech data is replete with background noise, channel distortion, and drastic fluctuations in speaker style [21]. Although traditional Convolutional Neural Networks (CNNs) [1] or Transformers [25] excel at feature extraction, they often lack the necessary inductive bias for the fine-grained time-frequency textures specific to speech signals. Consequently, this leads to the loss of critical emotional formant information in noisy environments [23]. A more profound challenge lies in the ambiguity and complexity of human emotional expression. Psychological studies indicate that emotions in authentic dialogues are often fluid and blended [17]. Speakers frequently exhibit multi-label co-occurrence states, such as "anger mixed with surprise" or "sadness tinged with disgust" [32]. Forcing such states into a single-label classification not only contradicts psychological cognitive facts but also leads to the loss of critical emotional nuances.

Addressing such complex emotion recognition, the field of Multi-modal Multi-label Emotion Recognition (MMER) has witnessed tremendous success [7, 24]. Current MMER models, such as HHMPN [30] and CARAT [15], leverage complex heterogeneous graph networks or cross-modal attention mechanisms to fully utilize the semantic explicitness of text and the intuitiveness of vision to resolve ambiguity, thereby accurately approximating the ground truth of emotions.

Inspired by these advanced architectural designs in MMER [9, 11, 13–15, 29–31], this paper proposes a core scientific hypothesis: Even when the input is strictly confined to a single speech modality, is it possible to reconstruct the emotional reasoning

capability of unimodal models by drawing upon the successful concepts of "hierarchical interaction" and "label dependency modeling" from MMER? This is a highly challenging and counter-intuitive endeavor. It is generally believed that unimodal models cannot rival the performance of multimodal models. However, we argue that the underperformance of existing unimodal speech models stems from their predominant focus on feature extraction, while overlooking the mining of deep "label-feature" and "label-label" dependencies—the very essence of MMER models. If a similar logical reasoning mechanism can be constructed within a single modality, it becomes possible to approximate the "Ground Truth Emotion" (typically defined by full-modal information) relying solely on acoustic cues, even in the absence of semantic and cross-modal auxiliary information. Based on these inspirations and reflections, this paper proposes a Triple Attention Contrastive Network (TAC-Net) tailored for real-world scenarios. This framework successfully migrates the concepts of "multi-view feature decoupling" and "label dependency modeling" from MMER to the acoustic feature space, aiming to address the challenges of complex emotion recognition under unimodal input. The core architecture of TAC-Net consists of two synergistic key components: The first component is the Label-wise Attention Module. To excavate rich information capable of approximating full-modal perception from a single speech modality, we design three parallel feature extraction networks, each addressing distinct task challenges:

(a)Multi-label Attention (MLA) Network: Designed to effectively address the co-occurrence of complex emotions in real-world scenarios, the MLA network aims to explicitly model the alignment relationship between acoustic features and complex emotion labels.

(b)Local-Global Attention (LGA) Network: To mitigate noise interference and handle the localization of discrete key emotional information in authentic dialogue scenarios, the LGA network is employed to capture context at different granularities.

(b)Time-Frequency Attention (TFA) Network: Targeting the time-frequency dimensions of acoustic features, this network focuses on key energy regions within the spectrogram to counteract background noise and channel distortion in real-world environments.

The second component is the Contrastive Reconstruction-based Fusion Module. Distinct from simple feature concatenation, this module is designed to address the heterogeneity and redundancy inherent among the three feature branches. We introduce a supervised contrastive learning strategy [12] to bridge the distance in the latent space between samples sharing similar complex emotional patterns. Simultaneously, a feature reconstruction mechanism is employed to deeply integrate the three-way acoustic features extracted by TFA, LGA, and MLA. This process not only enhances the discriminative power of the features but also ensures that the final generated representation maximally reconstructs the distribution of the "Ground Truth Emotion".

The main contributions of this paper are summarized as follows:

1.Inspired by advanced models in the MMER domain, we propose the TAC-Net framework. This represents the first attempt to construct a dual-stage processing mechanism comprising "Label-wise Attention" and "Contrastive Reconstruction

Fusion", effectively mitigating the performance bottleneck caused by information deficiency in unimodal inputs.

2.We design three synergistic networks—TFA, LGA, and MLA—to achieve comprehensive decoupling and modeling of acoustic time-frequency details, dialogue context dependencies, and emotion label correlations, significantly enhancing the model's capability to parse complex emotions.

3.Extensive experiments on two large-scale real-world datasets, M^{3}ED and CMU-MOSEI, demonstrate that TAC-Net not only substantially outperforms existing unimodal baseline models but also rivals mature Multi-modal Multi-label Emotion Recognition methods on key metrics, setting new performance benchmarks for the field.

2 Methodology

The proposed Triple Attention Contrastive Network (TAC-Net) is designed to extract deep emotional representations from a single speech modality through multi-view feature decoupling and dependency modeling, specifically targeting complex emotion recognition tasks in real-world scenarios. The overall processing pipeline of the framework comprises four core stages: feature extraction and encoding, multi-branch attention modeling, feature fusion and reconstruction, and emotion classification.

First, the input raw speech signals undergo preprocessing and acoustic feature extraction to obtain low-level representations containing basic emotional cues. Subsequently, the extracted feature sequences are fed into three parallel processing networks. Each network initially employs an independent Transformer encoder to capture long-term dependencies, generating high-order semantic features. These features are then input into the Label-wise Attention Module—comprising the Multi-label Attention (MLA), Local-Global Attention (LGA), and Time-Frequency Attention (TFA) networks. This module aims to perform refined decoupling and enhancement of emotional features from three distinct dimensions: frequency-domain texture, spatiotemporal context, and label co-occurrence relationships.

To address the heterogeneity and redundancy among the multi-branch features, the feature vectors output by the aforementioned modules are further forwarded to the Contrastive Reconstruction-based Fusion Module. This module utilizes a contrastive learning mechanism to bridge the distance between samples belonging to the same emotion class within the latent space, while achieving deep interaction and complementarity of the three-way features through a feature reconstruction strategy. Finally, the three optimized features output by the fusion module are aggregated into a unified high-dimensional emotional representation and fed into a fully connected classifier layer, thereby achieving the final prediction and recognition of complex speech emotion states. The overall architecture of the TAC-Net model is illustrated in Figure 1.

2.1 Label-wise Attention Module

To achieve precise analysis of complex emotions within a single speech modality, we design the Label-wise Attention Module. This module accepts high-order semantic features from the upstream Transformer encoder as input and distributes them

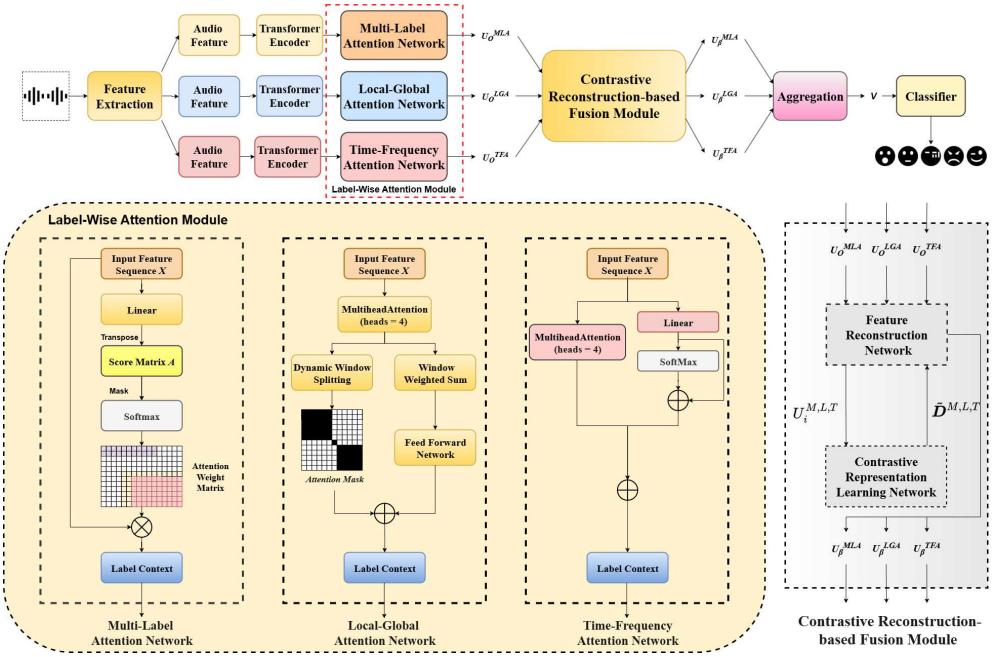


Fig. 1 The overall architecture of Triple Attention Contrastive Network (TAC-Net)

in parallel to three functionally distinct attention networks: the Multi-label Attention (MLA), Local-Global Attention (LGA), and Time-Frequency Attention (TFA) networks. These networks are designed to deeply decouple and reconstruct speech features from three respective dimensions—label dependency reasoning, perception of discrete emotional cues, and acoustic channel distortion—thereby compensating for the inherent information scarcity of unimodal input.

To intuitively illustrate the specific tasks emphasized by each attention branch, we randomly selected samples from the CMU-MOSEI test set to conduct a visualization analysis of attention weights. As shown in Figure 2, by mapping the second-by-second weight variations of the three attention branches to a normalized range, we can observe that the attentional regions of each network exhibit significant distinctiveness and complementarity in the temporal dimension. The weight curve of LGA exhibits distinct fluctuation patterns. It tends to keenly capture local abrupt changes and stressed segments within the speech signal, thereby precisely localizing the key frames of emotional outbursts. In contrast, the weight distribution of MLA presents a more steady and continuous trend. This indicates that this network primarily focuses on modeling the global co-occurrence relationships of emotion labels across the entire utterance, rather than being confined to isolated instantaneous features. TFA demonstrates selective persistence towards specific high-energy intervals, effectively suppressing interference from background silence segments and non-emotion-related noise.

This differentiated feature focusing mechanism compellingly demonstrates that TAC-Net can achieve multi-dimensional reconstruction of speech signals within the

latent space through multi-view synergy. This approach effectively mitigates the risks of feature loss or overfitting often associated with single attention mechanisms, thereby ensuring the robustness of emotion recognition in complex acoustic environments.

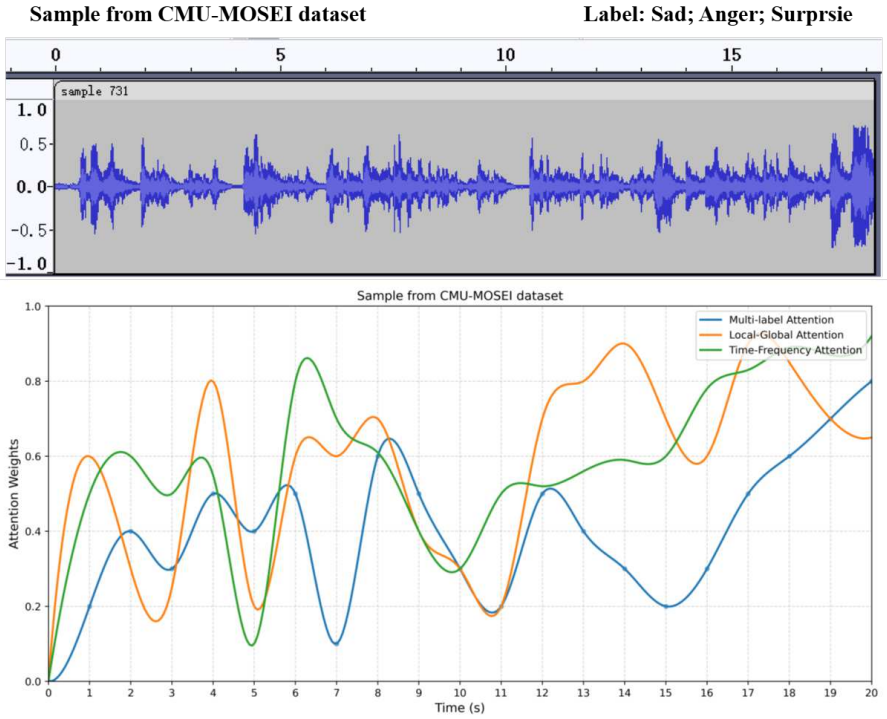


Fig. 2 Visualization of attention weights for a randomly selected sample from the CMU-MOSEI dataset. The horizontal axis represents time (in seconds), while the vertical axis denotes the attention weight values normalized via Softmax (ranging from 0 to 1). Curves of different colors depict the temporal response intensities of the Local-Global Attention (LGA), Multi-label Attention (MLA), and Time-Frequency Attention (TFA) networks, respectively, intuitively revealing the distinctness and complementarity of each network in feature extraction.

2.1.1 Multi-Label Attention Network

To effectively address the co-occurrence of complex emotions in real-world scenarios, the Multi-label Attention (MLA) network aims to explicitly model the alignment relationship between features and complex emotion labels. Distinct from traditional methods that solely extract global features, this network captures temporal segments that contribute significantly to specific emotion labels by calculating class-specific attention distributions.

Specifically, let the input feature sequence for this network be denoted as $X \in \mathbb{R}^{T \times d}$, where T represents the sequence length and d denotes the feature dimension. First, to map the acoustic features into the emotion label space, we feed X into a

learnable Linear Layer and perform a transpose operation on its output to obtain the initial label-time score matrix A. This process can be formally formulated as:

$$A = (XW_{att} + b)^\top \tag{1}$$

Here, $W_{att} \in \mathbb{R}^{d \times L}$ is the weight matrix of the linear transformation, $b \in \mathbb{R}^L$ is the bias term, L represents the total number of emotion labels (i.e., the number of classes), and $A \in \mathbb{R}^{L \times T}$ denotes the unnormalized label activation scores.

Subsequently, to eliminate interference from padding regions or irrelevant backgrounds during attention calculation, we introduce a masking operation. We then utilize the Softmax function to normalize the score matrix along the temporal dimension (T), thereby generating the final attention weight matrix α:

$$\alpha = \text{Softmax}(\text{Mask}(A)) \tag{2}$$

Here, the element $\alpha_{i,j}$ in $\alpha \in \mathbb{R}^{L \times T}$ characterizes the degree of contribution of the input feature at the j-th time step towards the i-th emotion category.

Next, we perform weighted aggregation on the original input features X using the computed attention weights α. This step is implemented via matrix multiplication, aiming to generate corresponding, highly discriminative feature representations for each emotion category. Finally, the aggregated features are passed through a Label Context processing module to generate the final output U_O^{MLA} of this network. The entire calculation process is defined as:

$$U_O^{MLA} = \Phi_{ctx}(\alpha \times X) \tag{3}$$

Where $\times$ denotes the matrix multiplication operation. The result of $\alpha \times X$ has a dimension of $\mathbb{R}^{L \times d}$, meaning a d-dimensional context vector is generated for each emotion label. $\Phi_{ctx}(\cdot)$ represents the Label Context processing function (composed of a Fully Connected layer and a non-linear activation function), designed to further refine the label-specific semantic information. The final output U_O^{MLA} is then utilized for subsequent feature fusion and classification prediction.

2.1.2 Local-Global Attention Network

To effectively address noise interference and the localization of discrete key emotional information in authentic dialogue scenarios, the Local-Global Attention (LGA) network aims to simultaneously capture short-term local abrupt changes and long-term global contextual dependencies within the speech signal through a parallel processing mechanism.

Given the input feature sequence $X \in \mathbb{R}^{T \times d}$, this network first employs a Multi-Head Self-Attention mechanism to capture preliminary correlations among features. First, the Query, Key, and Value matrices are computed as follows: $Q = XW_Q$, $K = XW_K$, $V = XW_V$, where W_Q, W_K, and W_V denote the learnable projection parameter matrices. Subsequently, the normalized attention score matrix S is obtained via the Scaled Dot-Product Attention calculation, formulated as follows:

$$S = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \tag{4}$$

where d_k is the scaling factor. Subsequently, this attention score matrix is fed into both the Local Branch and the Global Branch to extract features at different scales.

In the Local Branch, the model utilizes the attention scores to perform dynamic window segmentation, aiming to filter out local feature regions with high saliency. To focus on critical information, we introduce a threshold masking mechanism. Specifically, a threshold $\tau = 0.5$ is defined; regions within the score matrix exceeding this threshold are preserved, while those falling below it are suppressed via an Attention Mask. The masked weights then interact with the Value matrix V to generate the output of the local branch, denoted as X_{local}. This process can be formulated as:

$$M_{ij} = \begin{cases} 1, & \text{if } S_{ij} > \tau \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$X_{local} = (S \odot M) \times V \tag{6}$$

Where $\odot$ denotes the Hadamard Product, and M is the binary mask matrix generated based on the threshold.

Simultaneously, in the Global Branch, the model focuses on integrating holistic contextual information. This branch adopts a Window Weighted Sum strategy, utilizing the original attention scores S to perform global weighted aggregation on the Value matrix V, thereby capturing long-range contextual dependencies. The aggregated features are subsequently fed into a Feed Forward Network (FFN) for non-linear transformation and feature refinement, generating the global branch output X_{global}:

$$X_{global} = \text{FFN}(S \times V) \tag{7}$$

Where the FFN consists of two linear transformation layers interspersed with an activation function.

Finally, to integrate local details with global semantics, we perform an Element-wise Sum on the outputs of the Local and Global branches. The aggregated result is further refined via the Label Context module Φ_{ctx} to obtain the final output U_O^{LGA} of the LGA branch:

$$U_O^{LGA} = \Phi_{ctx}(X_{local} + X_{global}) \tag{8}$$

Through this design, the LGA branch effectively achieves unified modeling of local saliency and global coherence within the feature space.

2.1.3 Time-Frequency Attention Network

To overcome the prevalent background noise interference and channel distortion in real-world speech signals, the Time-Frequency Attention (TFA) network employs a dual-path parallel architecture. This design aims to perform refined modeling and enhancement of input features from two complementary dimensions: temporal dynamics and frequency-domain spectral characteristics.

Given the input feature sequence $X \in \mathbb{R}^{T \times d}$, the network first distributes it into the Time-domain Branch and the Frequency-domain Branch. In the Time-domain Branch, we utilize the Multi-Head Attention mechanism to capture the dynamic correlations of

the speech signal as it evolves over time. Specifically, the mechanism projects the input X into the Query matrix Q_t, Key matrix K_t, and Value matrix V_t. It then computes attention weights via the Scaled Dot-Product operation to aggregate global temporal information. The output of the Time-domain Branch, X_{time}, can be expressed as:

$$X_{time} = \text{MHA}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{9}$$

Where $\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V)$, and W^O denotes the output projection matrix. This process enables the model to focus on temporal segments exhibiting significant emotional expression.

Simultaneously, in the Frequency-domain Branch, the model focuses on excavating the distribution patterns of features along the frequency dimension. First, the input feature X undergoes feature projection via a learnable Linear Layer to extract latent frequency-domain semantic representations:

$$H_{freq} = XW_f + b_f \tag{10}$$

Where $W_f \in \mathbb{R}^{d\times d}$ and $b_f \in \mathbb{R}^d$ denote the weight and bias of the linear layer, respectively.

Subsequently, to obtain the normalized saliency distribution of the features, we feed H_{freq} into the Softmax function. To simultaneously preserve the intensity information of the original projected features and the relative importance after normalization, this branch adopts a feature enhancement strategy. Specifically, it performs an element-wise summation of the direct output of the linear layer and the output of the Softmax function, thereby obtaining the final result of the frequency-domain branch, X_{freq}:

$$X_{freq} = H_{freq} + \text{Softmax}(H_{freq}) \tag{11}$$

Via this summation operation, the model not only preserves the magnitude information of frequency-domain features but also enhances the saliency of key frequency components.

Finally, to achieve complementary fusion of time-frequency information, we aggregate the Time-domain Branch output X_{time} and the Frequency-domain Branch output X_{freq} via summation. The aggregated features are then input into the Label Context module Φ_{ctx} to generate the final output U_O^{TFA} of this network. The entire fusion process is formally defined as:

$$U_O^{TFA} = \Phi_{ctx}(X_{time} + X_{freq}) \tag{12}$$

This design effectively integrates temporal context with frequency-domain texture features, thereby elevating the model's emotional representation capability in complex acoustic environments.

2.2 Contrastive Reconstruction-based Fusion Module

Although the aforementioned TFA, LGA, and MLA networks effectively capture deep emotional representations from distinct perspectives, the simple concatenation of these heterogeneous features—which originate from different semantic spaces—often leads to feature distribution misalignment and redundancy, thereby limiting the model's final

discriminative performance.To this end, we propose the Contrastive Reconstruction-based Fusion Module. This module innovatively integrates a Feature Reconstruction Network with a Contrastive Representation Learning Network. It is designed to align and refine multi-view features within a unified latent space through a "reconstruction-contrast-re-reconstruction" closed-loop process. The framework of the Contrastive Reconstruction-based Fusion Module is illustrated in Figure 3.

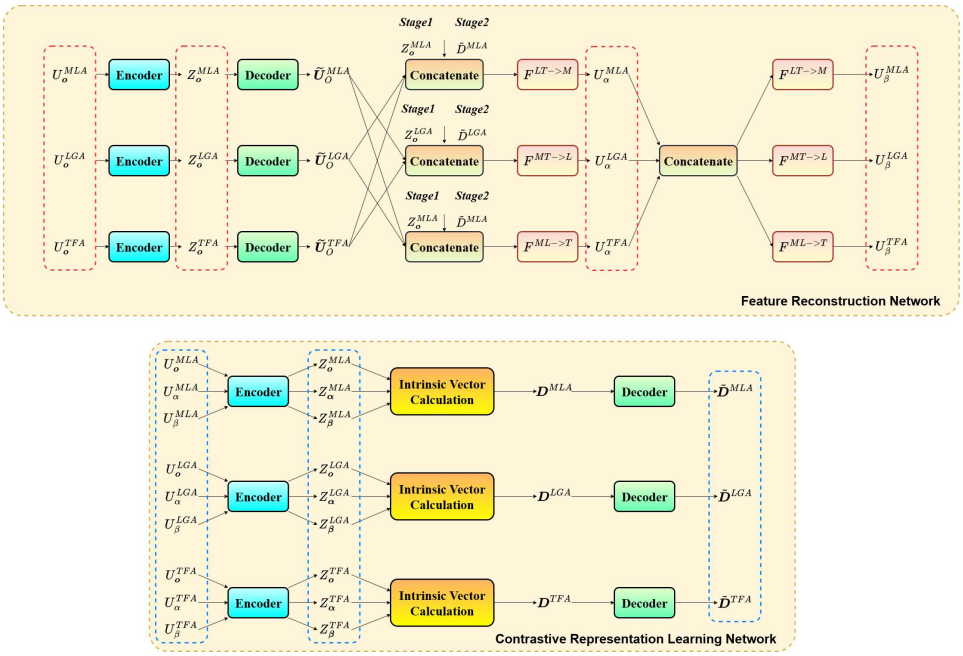


Fig. 3 The overall architecture of the Contrastive Reconstruction-based Fusion Module.

2.2.1 Feature Reconstruction Network

The Feature Reconstruction Network serves as the core component of the fusion module, designed to address the heterogeneity of multi-branch features regarding their distribution in the semantic space. Through an "Encoding-Decoding-Reorganization" mechanism, this network enforces interaction and alignment among features from different perspectives.

First, the feature representations $U_O^{MLA}, U_O^{LGA}, U_O^{TFA}$, output by the preceding modules, serve as the initial inputs. To excavate deep contextual associations, these three feature sequences are fed in parallel into a shared-weight Transformer Encoder. This encoder utilizes a self-attention mechanism to perform global dependency modeling on the original features, yielding enhanced feature representations. Let the enhanced features for each branch be denoted as Z_O^i, respectively. The process is formulated as follows:

$$Z_O^i = \text{Encoder}(U_O^i), \quad i \in \{MLA, LGA, TFA\} \tag{13}$$

Subsequently, to acquire high-order semantic information suitable for cross-branch interaction, the enhanced features Z_O^i are further mapped via a Decoder, yielding three complementary semantic features, denoted as $\tilde{U}_O^{MLA}$, $\tilde{U}_O^{LGA}$, and $\tilde{U}_O^{TFA}$.

Upon obtaining these foundational features, the network proceeds to the core reconstruction fusion process. To ensure informational integrity from specific perspectives, we design a set of specific permutation functions to effectively integrate "self" features (from the current branch) with "other" features (from the remaining branches). Taking the reconstruction process where the MLA branch serves as the primary view as an example, the procedure consists of the following two stages:

Stage 1: Feature Enhanced Reconstruction. This stage aims to utilize the decoded semantic features of the other two branches (LGA and TFA) to reconstruct the enhanced features of the MLA branch. Specifically, we combine the MLA enhanced feature Z_O^{MLA} with the decoded features $\tilde{U}_O^{LGA}$ and $\tilde{U}_O^{TFA}$.

Stage 2: Contrastive Consistency Reconstruction. This stage introduces intrinsic features (denoted as $\tilde{D}^{MLA}$) generated by the Contrastive Representation Learning network, aiming to incorporate class-discriminative information into the reconstruction process. Consequently, the input combination shifts to integrating $\tilde{D}^{MLA}$ with $\tilde{U}_O^{LGA}$ and $\tilde{U}_O^{LGA}$.

To formally describe this process, we define a Sequential Concatenation Function, denoted as $\mathcal{F}^{AB\to C}(\cdot)$. This function concatenates input vectors following a specific semantic order and subsequently applies a linear projection:

$$\mathcal{F}^{AB\to C}(Input_A, Input_B, Input_C) = \text{Linear}(\text{Concat}(Input_A, Input_B, Input_C)) \tag{14}$$

Consequently, the Level-1 Reconstructed Feature corresponding to the MLA branch, denoted as U_α^{MLA}, is derived by aggregating the features using the aforementioned function:

$$U_\alpha^{MLA} = \mathcal{F}^{LT\to M}(\tilde{U}_O^{LGA}, \tilde{U}_O^{TFA}, Z_O^{MLA}/\tilde{U}_O^{LGA}) \tag{15}$$

Similarly, by defining the corresponding functions for other branches, we can obtain the Level-1 reconstructed features U_α^{LGA} and U_α^{TFA} in parallel. Finally, to generate the final global representation employed for classification, we re-aggregate the three Level-1 reconstructed features to yield the Level-2 Reconstructed Feature. This process is similarly implemented via the concatenation function:

$$U_\beta^{MLA} = \mathcal{F}^{LT\to M}(U_\alpha^{LGA}, U_\alpha^{TFA}, U_\alpha^{MLA}) \tag{16}$$

$$U_\beta^{LGA} = \mathcal{F}^{MT\to L}(U_\alpha^{MLA}, U_\alpha^{TFA}, U_\alpha^{LGA}) \tag{17}$$

$$U_\beta^{TFA} = \mathcal{F}^{ML\to T}(U_\alpha^{MLA}, U_\alpha^{LGA}, U_\alpha^{TFA}) \tag{18}$$

By employing this hierarchical reconstruction strategy, the model not only preserves the unique strengths of each individual branch but also achieves deep complementarity

and fusion of information from the various attention branches within a unified feature space.

2.2.2 Contrastive Representation Learning Network

To ensure semantic consistency of features across different reconstruction levels and to further enhance the discriminative capability of features regarding emotion categories, we introduce the Contrastive Representation Learning Network. This network aims to extract "Intrinsic Representations" highly correlated with emotion labels from multi-level feature views and feeds them back into the Feature Reconstruction Network as a prior constraint.

Taking the MLA branch as an example, we first construct a multi-view feature triplet as the input for the network. This triplet comprises the feature states of the branch at distinct processing stages: Original Feature U_O^{MLA}, Level-1 Reconstructed Feature U_α^{MLA}, and Level-2 Reconstructed Feature U_β^{MLA}.

This set of feature inputs is first fed into a shared-weight encoder (Contrastive Encoder), designed to map features from different levels into a unified high-dimensional latent space. For each feature $Z \in S_{input}$ in the input set $S_{input} = \{U_O^{MLA}, U_\alpha^{MLA}, U_\beta^{MLA}\}$, its corresponding enhanced latent vector Z is calculated as follows:

$$Z = \text{Encoder}_{scl}(X) = \sigma(W_e X + b_e) \tag{19}$$

Where W_e and b_e are the learnable parameters of the encoder, and σ denotes the non-linear activation function. Through this step, we obtain three corresponding enhanced feature vectors: Z_O^{MLA}, Z_α^{MLA} and Z_β^{MLA}.

Subsequently, to extract the most essential and stable emotional semantics from representations across these three views, we feed these three enhanced features into the Intrinsic Vector Calculation Module. This module employs an adaptive aggregation mechanism to filter out noise introduced by hierarchical differences, calculating the intrinsic center vector D^{MLA} for this branch:

$$D^{MLA} = \Psi_{cal}(Z_o^{MLA}, Z_\alpha^{MLA}, Z_\beta^{MLA}) = \frac{1}{3} \sum_{k \in \{o,\alpha,\beta\}} \frac{Z_k^{MLA}}{\| Z_k^{MLA} \|_2} \tag{20}$$

In this formula, $\Psi_{cal}(\cdot)$ represents the intrinsic vector calculation function, where we employ normalized mean aggregation to ensure the directional consistency of the features. Consequently, D^{MLA} characterizes the core semantics that are shared across all reconstruction stages of the MLA branch and are most closely correlated with the emotion labels.

At this stage, to explicitly enhance the intra-class compactness and inter-class separability of D^{MLA}, we introduce the Supervised Contrastive Loss ($\mathcal{L}_{scl}$) during the training process. Specifically, for a training batch containing N samples, we treat the intrinsic vector D_i^{MLA} of the current sample i as the anchor. All samples in the batch sharing the same class label constitute the positive sample set $P(i)$, while the remaining samples are treated as negative samples. This loss function aims to maximize the mutual information between the anchor and the positive samples, defined as follows:

$$\mathcal{L}_{scl} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(D_i^{MLA} \cdot D_p^{MLA}/\tau\right)}{\sum_{a \in A(i) \setminus \{i\}} \exp\left(D_i^{MLA} \cdot D_a^{MLA}/\tau\right)} \tag{21}$$

Where I denotes the set of sample indices within the batch, $A(i)$ represents the set of all samples excluding i, τ is the Temperature Parameter regulating the smoothness of the distribution, and $(\cdot)$ denotes the dot product operation. By minimizing $\mathcal{L}_{scl}$, the network forces intrinsic vectors belonging to the same emotion category to cluster together on the hypersphere, while simultaneously repelling vectors of different categories.

Finally, to map this intrinsic representation back to a semantic space compatible with the Feature Reconstruction Network for subsequent calculations, we feed the intrinsic vector D^{MLA}, which has been optimized via contrastive constraints, into a Decoder. This decoder generates the final Contrastive Consistency Feature, denoted as $\widetilde{D}^{MLA}$, through a linear mapping:

$$\tilde{D}^{MLA} = \text{Decoder}_{scl}(D^{MLA}) = W_d D^{MLA} + b_d \tag{22}$$

Where W_d and b_d denote the parameters of the decoder. The output $\widetilde{D}^{MLA}$ not only encapsulates the inter-class discriminative power derived from contrastive learning but also integrates the stability of multi-level reconstruction fused via the intrinsic vector calculation.As previously stated, this feature $\widetilde{D}^{MLA}$ directly participates in the calculation process of Stage 2 (Contrastive Consistency Reconstruction) within the Feature Reconstruction Network. Serving as guidance information, it is combined with $\tilde{U}_o^{LGA}$ and $\tilde{U}_o^{TFA}$, thereby achieving a deep unification of physical feature reconstruction and semantic class constraints.

3 Experiments

3.1 Dataset selection and Evaluation metrics

To comprehensively evaluate the model's performance in recognizing complex emotions within real-world scenarios, this study selects two highly representative multimodal emotion datasets: CMU-MOSEI and M^{3}ED. These datasets represent single-person natural expressions in uncontrolled environments and multi-person dialogues in real interaction scenarios, respectively. This selection enables the verification of the model's ability to capture complex speech emotion features from different dimensions.

3.1.1 M^{3}ED dataset

M^{3}ED (Multi-modal Multi-scene Multi-label Emotional Dialogue Database) [32] is a Chinese multimodal, multi-scene, multi-label emotional dialogue dataset released by Zhao et al. in 2022. Unlike the monologue format of CMU-MOSEI, M^{3}ED focuses on emotional interaction in dyadic dialogue scenarios. Sourced from 56 TV series across various genres, the dataset comprises 990 dialogue fragments, consisting of 9,082 turns and 24,449 utterances. Since the data originates from diverse scenes such as family, romance, and modern dramas, the emotional expressions exhibit high spontaneity and

life-like characteristics, authentically reflecting the dynamic flow and mutual influence of emotions during conversation. The core advantage of the M^{3}ED dataset lies in its precise definition and annotation of "complex emotions." Considering that emotional expressions in real dialogues are rarely single-dimensional, this dataset adopts a multi-label annotation system, allowing a single utterance to be labeled with multiple emotion categories simultaneously. This accurately captures complex, mixed emotional states in humans. At the utterance level, the dataset covers seven emotion categories: Happy, Surprise, Sad, Disgust, Angry, Fear, and Neutral, while providing multimodal information including acoustic, visual, and textual data. In this study, M^{3}ED is utilized to specifically examine the model's capability in handling context-dependency in dialogues and recognizing mixed complex emotions, with a particular focus on generalization performance within the Chinese context.

3.1.2 CMU-MOSEI dataset

CMU-MOSEI (CMU Multimodal Opinion Sentiment and Emotion Intensity) [28], constructed by Zadeh et al. in 2018, stands as one of the largest and most richly annotated datasets for multimodal sentiment analysis and emotion recognition to date. Sourced from the YouTube video sharing platform, it contains 23,453 video clips from 1,000 distinct speakers, totaling over 65 hours. As a typical "in-the-wild" dataset, all samples in CMU-MOSEI are single-person monologues, covering 250 different topics such as movie reviews, product reviews, and opinion expression. This uncontrolled acquisition environment preserves the diversity of real-world lighting changes, background noise, and personalized speaker expressions, which is of significant value for verifying the model's robustness in complex acoustic environments.Regarding annotation, CMU-MOSEI differs from traditional discrete emotion classification tasks by offering more fine-grained emotion intensity annotations. The dataset not only includes sentiment polarity scores in the range of [-3, 3] but also provides quantitative intensity scores for six basic emotions: Happy, Sad, Angry, Fear, Disgust, and Surprise. Furthermore, it provides strictly aligned textual, visual, and acoustic data. Notably, the acoustic modality details paralinguistic features such as intonation and prosody, offering a rich data foundation for studying subtle variations in speech emotion. In this study, the use of this dataset aims primarily to verify the model's effectiveness in processing natural spoken language tasks.

3.1.3 Evalution Metrics

To objectively evaluate the model's performance in complex emotion recognition tasks, this study selects Accuracy, Precision, Recall, and the Micro-average F1 score (Micro-F1) as the core evaluation metrics. Accuracy intuitively reflects the correctness of the model's overall classification, while Precision and Recall assess the model's capability regarding exactness and completeness for specific emotion categories, respectively. The calculation formulas for these basic metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

Where TP, FP, TN and FN represent the numbers of True Positive, False Positive, True Negative, and False Negative samples, respectively.Furthermore, given that emotional data in real-world dialogue scenarios often exhibit significant class imbalance, single metrics struggle to comprehensively reflect model performance. Therefore, this experiment specifically adopts the Micro-F1 as a comprehensive measure. This metric calculates the global Precision (P_{micro}) and Recall (R_{micro}) by first aggregating prediction results across all categories. This approach assigns equal weight to each sample, thereby effectively mitigating the dominating influence of majority class samples on the evaluation results. Its calculation formula is as follows:

$$Micro - F1 = \frac{2 \times P_{micro} \times R_{micro}}{P_{micro} + R_{micro}} \tag{26}$$

3.2 Comparative Experiments

3.2.1 Selection of Comparative Models

To comprehensively evaluate the recognition performance of the proposed model on complex speech emotions in real-world dialogue scenarios, this section selects two categories of representative baseline models for comparison:Traditional Single-label Speech Emotion Recognition (SER) models, which focus on acoustic feature mining.Multimodal Multi-label Emotion Recognition (MMER) models, which possess more complex architectures. Through a cross-comparison with these two categories, we aim to verify the effectiveness of our method in feature extraction and its superiority in complex emotion modeling.

It is crucial to emphasize that the selection of MMER models as baselines is driven by a core motivation that transcends mere performance comparison. Traditional SER models often treat different emotions as mutually independent discrete categories, thereby ignoring the potential label correlations among complex emotions. In contrast, MMER models are inherently designed with architectures capable of modeling label dependencies (e.g., via Graph Convolutional Networks or Label Embeddings).

To evaluate the capacity of different architectures to capture the "acoustic feature - emotion label" mapping within the same dimension, we strictly restrict the input of these multi-modal baseline models to the single audio modality in our experiments. This setting aims to establish a rigorous evaluation benchmark: Firstly, following the principle of control variables, this strips away the information gain from textual and visual modalities, ensuring fairness in the comparison.Secondly, and more importantly, this allows us to investigate which architecture can more effectively utilize acoustic cues to reconstruct the intrinsic connections among emotion labels under constrained conditions (i.e., lacking the assistance of semantics and visual expressions).

By doing so, we aim to demonstrate that TAC-Net not only improves the precision of feature extraction but also fills the void in complex emotion label dependency modeling within the realm of uni-modal SER.

(A) Traditional Single-label Speech Emotion Recognition Models:

We selected four cutting-edge Transformer-based single-label SER models as baselines to evaluate the performance of existing architectures in feature extraction.

DWFormer (Dynamic Window Transformer) [4]:By introducing a dynamic window partitioning mechanism and a cross-window interaction strategy, DWFormer effectively addresses the limitations of traditional fixed windows in capturing multi-scale temporal features within speech signals.

DST (Deformable Speech Transformer) [6]: This model proposes a deformable attention architecture. By utilizing a lightweight decision network and data-driven offsets to adaptively adjust the position and size of attention windows, DST flexibly extracts multi-grained emotional cues.

ERT (Emotion in Real-scene Transformer) [22]: Targeting complex noise interference in real-world scenarios, ERT employs speech enhancement to improve feature quality and a dynamic window self-attention mechanism to precisely locate key emotional segments. It significantly enhances model robustness in uncontrolled environments through global information interaction.

CTMAM (CNN-Transformer with Multi-dimensional Attention Mechanism) [23]: CTMAM innovatively constructs a hybrid CNN-Transformer architecture. By integrating a multi-dimensional (time-channel-space) attention mechanism, it achieves joint modeling of local time-frequency details and global long-term dependencies of speech signals.

(B) Multi-modal Multi-label Emotion Recognition Models:

To address the complex feature interaction demands in multi-label emotion recognition tasks, we selected three representative multi-modal architectures as baselines and evaluated their feature modeling capabilities under single-modal input conditions.

HHMPN (Heterogeneous Hierarchical Message Passing Network) [30]: HHMPN innovatively constructs a heterogeneous graph structure. It explicitly models the multiple dependency relationships among features, modalities, and labels through a hierarchical message passing mechanism.

TAILOR (Versatile Multi-modal Learning) [31]: This model focuses on feature decoupling and refinement. It designs adversarial modules to distinguish between inter-modal commonalities and specificities, and utilizes a BERT-like cross-modal encoder to achieve deep cascading of label and modal features.

CARAT (Contrastive Feature Reconstruction and Aggregation) [15]: CARAT proposes a reconstruction-based fusion framework. It combines contrastive learning strategies to extract modality-separated and label-specific feature representations, thereby effectively capturing fine-grained modality-label mapping relationships.

3.2.2 Comparative Experimental Results and Analysis

Table 1 presents the detailed comparative results of TAC-Net against other baseline models on both the M^{3}ED and CMU-MOSEI datasets. The proposed TAC-Net achieves superior performance across all evaluation metrics on both datasets, robustly

demonstrating its effectiveness and robustness in handling complex speech emotion tasks.

Table 1 Performance comparison of different models on M^{3}ED and CMU-MOSEI datasets. The best results are highlighted in bold.

Dataset	Model	Accuracy	Precision	Micro-F1	Recall
M^{3}ED	TAILOR (only audio)	0.502	0.816	0.650	0.540
	CARAT (only audio)	0.564	0.829	0.665	0.555
	HHMPN (only audio)	0.503	0.818	0.653	0.543
	ERT	0.492	0.735	0.578	0.476
	DWFormer	0.493	0.677	0.488	0.381
	DST	0.494	0.687	0.515	0.412
	CTMAM	0.531	0.805	0.616	0.499
	Ours	**0.603**	**0.947**	**0.752**	**0.623**
CMU-MOSEI (Aligned)	TAILOR (only audio)	0.388	0.541	0.468	0.410
	CARAT (only audio)	0.393	0.569	0.489	0.428
	HHMPN (only audio)	0.384	0.551	0.464	0.401
	ERT	0.301	0.491	0.405	0.345
	DWFormer	0.298	0.473	0.452	0.432
	DST	0.315	0.461	0.434	0.411
	CTMAM	0.376	0.511	0.447	0.398
	Ours	**0.410**	**0.592**	**0.501**	**0.435**

First, on the highly challenging Chinese dialogue dataset M^{3}ED, TAC-Net exhibits significant performance advantages. Specifically, TAC-Net achieved an Accuracy of 0.603 and a Micro-F1 of 0.752. Compared to the second-best performing baseline, CARAT (Acc 0.564, Micro-F1 0.665), these represent substantial improvements of 3.9% and 8.7%, respectively. It is worth noting that in terms of Precision, TAC-Net attained a remarkable score of 0.947, significantly outperforming multi-modal multi-label emotion recognition architectures such as CARAT (0.829) and TAILOR (0.816). This result indicates that TAC-Net's unique emotion decoupling and feature interaction mechanisms effectively reduce the false positive rate, demonstrating high discriminative capability when dealing with complex and easily confused emotion labels in dialogues.

Furthermore, an inspection of the data reveals that traditional single-label SER models (such as DWFormer and DST) generally exhibit lower Micro-F1 scores (approximately 0.49–0.52) on the M^{3}ED dataset, falling far behind multi-modal architectures (even when the latter are restricted to audio-only input). This further confirms that traditional methods relying solely on temporal modeling struggle to address the issue of multi-label co-occurrence in real dialogue scenarios. In contrast, TAC-Net successfully bridges this gap by introducing multiple attention mechanisms and deeper feature interaction fusion, boosting the Recall to 0.623 from the average level of baseline models.

Secondly, on the CMU-MOSEI dataset, although overall metrics for all models are lower due to the high noise levels in the uncontrolled environment and the subtlety of emotional expression, TAC-Net maintains a stable lead. TAC-Net achieved

an Accuracy of 0.410 and a Micro-F1 of 0.501 on this dataset, both surpassing current state-of-the-art baseline models. Compared to the strong-performing CTMAM model (Acc 0.376, Micro-F1 0.447), TAC-Net improved Accuracy by 3.4% and the F1 score by 5.4%. Although multi-modal models like CARAT and HHMPN also exhibited strong competitiveness when using only the audio modality (Micro-F1 of 0.489 and 0.464, respectively), TAC-Net, by virtue of its ability to capture fine-grained acoustic features, reached a Recall of 0.435, slightly outperforming CARAT's 0.428. This marginal but consistent advantage indicates that even in single-person monologue scenarios characterized by drastic changes in emotional intensity and complex backgrounds, TAC-Net extracts key emotional features more effectively than existing complex architectures. This validates the model's strong generalization capability across different linguistic environments and emotional task settings.

3.3 Ablation Studies

To comprehensively verify the specific contributions of each core component of TAC-Net in addressing diverse task challenges, and to investigate the effectiveness of the feature fusion strategy, this section designs detailed ablation experiments. The model proposed in this study comprises not only three key networks responsible for feature decoupling—namely, the Time-Frequency Attention (TFA) network aimed at capturing fine-grained time-frequency information of acoustic features, the Local-Global Attention (LGA) network designed to extract discrete emotional information in real dialogue scenarios, and the Multi-Label Attention (MLA) network specialized in handling complex emotion co-occurrence and dependency relationships—but also a Contrastive Reconstruction Fusion Module utilized for feature alignment and enhancement. In the experiments, we constructed various variant models, including single-network configurations, dual-network combinations, and variants with specific fusion components removed (denoted as w/o FRN and w/o CRLN). Comparative tests were conducted on the M^{3}ED and CMU-MOSEI datasets to quantitatively evaluate the specific impact of each module on Accuracy, Precision, Recall, and Micro-F1. The experimental results of the ablation studies on the M^{3}ED and CMU-MOSEI datasets are presented in Table 2.

Based on the analysis of the experimental results in Table 2, the complete model (Ours), which integrates all components, achieves optimal comprehensive performance on both datasets. Specifically, on the M^{3}ED dataset, it attained an Accuracy of 0.603 and a Micro-F1 of 0.752. This result primarily confirms the functional complementarity of the three feature extraction modules: single-dimensional feature modeling fails to meet the high demands of complex emotion recognition in real-world scenarios. Performance maximization can only be achieved by synergistically utilizing time-frequency details, key emotional information, and label dependency relationships.

An in-depth analysis of single-module performance reveals distinct characteristics:

TFA (Only TFA) exhibited the strongest foundational performance among single networks, achieving an Accuracy of 0.573 and a Micro-F1 of 0.735 on M^{3}ED, effectively outperforming standalone LGA and MLA. This indicates that in Speech Emotion Recognition (SER) tasks, extracting high-quality time-frequency features from acoustic signals is the cornerstone of model performance. TFA effectively captures subtle

Table 2 Ablation study of different components on M^{3}ED and CMU-MOSEI datasets. The best results are highlighted in bold.

Dataset	Approaches	Accuracy	Precision	Micro-F1	Recall
M^{3}ED	Only MLA	0.507	0.871	0.663	0.535
	Only LGA	0.552	0.899	0.698	0.570
	Only TFA	0.573	0.912	0.735	0.616
	MLA+LGA	0.564	0.938	0.735	0.605
	MLA+TFA	0.590	0.925	0.755	0.638
	LGA+TFA	0.568	0.941	**0.763**	**0.641**
	w/o FRN	0.501	0.918	0.733	0.610
	w/o CRLN	0.499	0.855	0.667	0.547
	Ours	**0.603**	**0.947**	0.752	0.623
CMU-MOSEI (Aligned)	Only MLA	0.384	0.548	0.456	0.391
	Only LGA	0.387	0.552	0.481	0.426
	Only TFA	0.389	0.560	0.477	0.415
	MLA+LGA	0.391	0.583	**0.501**	0.439
	MLA+TFA	0.406	0.546	0.477	0.424
	LGA+TFA	0.407	0.555	0.498	**0.452**
	w/o FRN	0.373	0.568	0.471	0.402
	w/o CRLN	0.361	0.508	0.438	0.385
	Ours	**0.410**	**0.592**	**0.501**	0.435

emotional fluctuations, ensuring the model's foundational perception capability for emotional events.

LGA (Only LGA), while slightly inferior to TFA in Accuracy (0.552 on M^{3}ED), demonstrated competitive potential. On the CMU-MOSEI dataset, it achieved a Micro-F1 of 0.481, surpassing the single TFA model's 0.477. This suggests that in real-world scenarios characterized by high noise or complex contexts, the global and local fields of view provided by LGA effectively filter background interference and enhance feature robustness.

MLA (Only MLA) showed relatively limited performance when used alone (Accuracy 0.507 on M^{3}ED). This aligns with expectations, as modeling label dependencies relies on high-quality underlying feature inputs. Lacking the rich acoustic and contextual representations provided by TFA and LGA, MLA struggles to exert its full advantage in decoupling complex emotions.

At the level of module combinations, the synergistic effects among components are particularly significant. When MLA is introduced into dual-network combinations (e.g., MLA+LGA), the model's Precision achieves a substantial boost. For instance, on the M^{3}ED dataset, the Precision of the MLA+LGA combination reached 0.938, far exceeding that of the single LGA (0.899). This phenomenon powerfully demonstrates the core role of the MLA network in handling "complex emotions": by modeling label co-occurrence relationships, MLA effectively suppresses ambiguous erroneous predictions, thereby greatly enhancing classification precision. Notably, the LGA+TFA combination achieved the highest Recall (0.641) and Micro-F1 (0.763) on M^{3}ED, slightly surpassing even the full model in these specific metrics. This indicates that the

strong alliance of TFA and LGA maximizes the coverage of emotional feature capture, reducing missed detections, although the full model (Ours) achieves the best balance with the highest Accuracy and Precision (0.947). On CMU-MOSEI, the MLA+LGA combination achieved the highest Micro-F1 (0.501), matching the full model, further validating the importance of label-context interaction in noisy environments.

Furthermore, ablation experiments on the fusion strategy highlight the critical roles of the Contrastive Reconstruction-based Fusion Module (CRFM) and its internal Feature Reconstruction Network (FRN).

w/o CRLN: When the entire fusion module is removed (i.e., using a simple feature concatenation strategy), model performance suffers a drastic decline. On M^{3}ED, Accuracy dropped to 0.499 and Micro-F1 to 0.667—lower than even the single-branch TFA model. This counter-intuitive result profoundly reveals the difficulty of heterogeneous feature fusion: since TFA, LGA, and MLA focus on time-frequency, emotional information, and label semantics respectively, their generated feature distributions exhibit significant heterogeneity. Direct concatenation not only fails to achieve complementary advantages but also introduces redundancy and noise due to conflicts in the feature space. In contrast, TAC-Net effectively resolves this by mapping heterogeneous features into a unified latent space via CRLN.

w/o FRN: Further, when retaining the fusion module but removing the FRN component, although performance surpasses simple concatenation, there remains a distinct gap compared to the full model. For example, on M^{3}ED, removing FRN resulted in a 10.2% drop in Accuracy (from 0.603 to 0.501). This indicates that feature reconstruction, serving as an auxiliary task, acts as a crucial regularization mechanism. It forces the fused features to restore the original information of each branch, preventing the loss of key information during compression and fusion.

3.4 Visualization of Prediction Results

To intuitively demonstrate the discriminative capacity and prediction confidence of TAC-Net across different emotion distributions, this section constructs an emotion prediction visualization matrix.

We randomly selected four representative samples from the test sets of both the CMU-MOSEI and M^{3}ED datasets. A comparative display is presented for each sample, encompassing the Ground Truth, the prediction probability scores output by the model, and the final binarized prediction results.

As illustrated in Figure 4, within the CMU-MOSEI dataset—despite the samples originating from uncontrolled natural monologue environments characterized by significant background noise and non-verbal acoustic interference—TAC-Net demonstrates exceptional prediction certainty. Observing Sample 731 reveals that for the ground truth emotion category "Sad", the model assigns a high confidence score approaching 0.82. Conversely, prediction scores for other irrelevant emotion categories are effectively suppressed to extremely low levels (generally below 0.1).This distinct "Confidence Margin" indicates that, even under the constraints of single audio modality input and complex acoustic emotions, the model is capable of precisely capturing decisive acoustic emotion features, thereby effectively avoiding ambiguous judgments.

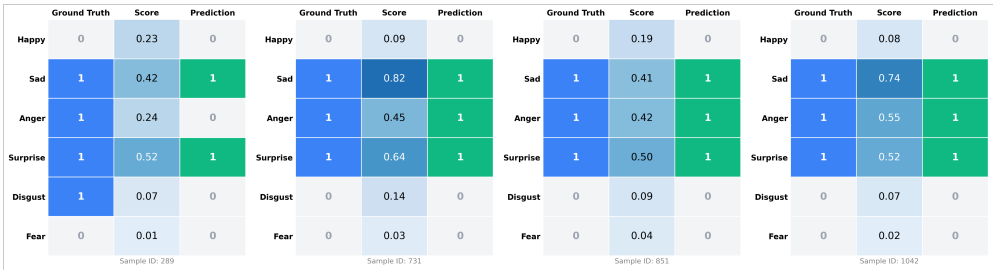


Fig. 4 Visualization results on the CMU-MOSEI dataset.

As illustrated in Figure 5, the visualization results for the M^{3}ED dataset further highlight the model's superiority in handling complex multi-label tasks. In real-world dialogue scenarios, a speaker's emotions are often mixed and dynamic.Taking Sample 125 as an example, its ground truth labels simultaneously encompass five emotions: Neutral, Sad, Disgust, Anger, and Surprise. TAC-Net successfully identified this emotion co-occurrence pattern, outputting high prediction scores exceeding the threshold across all these categories, which aligns perfectly with the ground truth. This success is attributed to the Multi-Label Attention (MLA) network incorporated into the model. By not only focusing on single-dimensional feature intensity but also implicitly modeling the latent semantic correlations among emotion labels, MLA ensures that the model maintains exceptional decoupling and recall capabilities when confronting complex entangled emotions such as "surprise with anger" or "sadness with disgust."

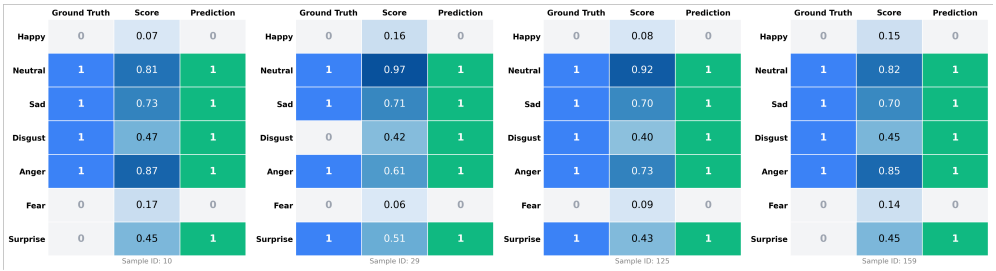


Fig. 5 Visualization results on the M^{3}ED dataset.

4 Conclusion

This paper proposes an innovative Triple Attention Contrastive Network (TAC-Net) to address the challenges of environmental noise interference and complex emotion co-occurrence in real-world speech emotion recognition scenarios. By integrating Multi-Label Attention (MLA), Local-Global Attention (LGA), and Time-Frequency Attention (TFA), the model effectively achieves multi-dimensional decoupling and precise extraction of emotional features at the levels of label correlation, temporal

context, and frequency-domain energy. Furthermore, the introduction of a contrastive reconstruction-based fusion mechanism significantly enhances the discriminability and noise robustness of features within the latent space. Extensive experiments on the large-scale M^{3}ED and CMU-MOSEI datasets demonstrate that TAC-Net not only significantly outperforms existing SER models across various evaluation metrics but also exhibits superior robustness when dealing with complex real-world acoustic environments, validating the immense potential of TAC-Net in complex emotion recognition tasks.

5 Future Work

Despite the satisfactory results achieved by TAC-Net in recognition accuracy, its multi-branch attention architecture still presents room for optimization regarding computational overhead. Future work will primarily focus on the lightweight design and engineering deployment of the model. We plan to explore model compression techniques such as knowledge distillation, model pruning, and low-rank decomposition. The aim is to drastically reduce the parameter count and computational complexity (FLOPs) while maintaining high model performance. Additionally, we will further investigate inference acceleration strategies tailored for mobile and edge devices to address operational efficiency issues in resource-constrained environments. Through these efforts, we aim to promote the practical application deployment of the model in scenarios such as intelligent customer service terminals, real-time mental health monitoring, and human-computer interaction (HCI) systems.

References

[1] Abdel-Hamid O, Mohamed A, Jiang H, et al (2014) Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 22(10):1533–1545

[2] Burkhardt F, Paeschke A, Rolfes M, et al (2005) A database of German emotional speech. In: Interspeech, vol 5. pp 1517–1520

[3] Busso C, Bulut M, Lee CC, et al (2008) IEMOCAP: Interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335–359

[4] Chen S, Xing X, Zhang W, et al (2023a) Dwformer: Dynamic window transformer for speech emotion recognition. In: ICASSP 2023. IEEE, pp 1–5

[5] Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PmLR, pp 1597–1607

[6] Chen W, Xing X, Xu X, et al (2023b) DST: Deformable speech transformer for emotion recognition. In: ICASSP 2023. IEEE, pp 1–5

[7] Chen ZM, Wei XS, Wang P, Guo Y (2019) Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF CVPR. pp 5177–5186

[8] Cummins N, Scherer S, Krajewski J, et al (2015) A review of depression and suicide risk assessment using speech analysis. Speech Commun 71:10–49

[9] Firdaus M, Chauhan H, Ekbal A, et al (2020) MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In: Proceedings of COLING. pp 4441–4453

[10] Gulati A, Qin J, Chiu CC, et al (2020) Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100

[11] Ju X, Zhang D, Li J, et al (2020) Transformer-based label set generation for multi-modal multi-label emotion detection. In: Proceedings of ACM Multimedia. pp 512–520

[12] Khosla P, Teterwak P, Wang C, et al (2020) Supervised contrastive learning. Adv Neural Inf Process Syst 33:18661–18673

[13] Le HD, Lee GS, Kim SH, et al (2023) Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. IEEE Access 11:14742–14751

[14] Li X, Liu J, Xie Y, et al (2024) Magdra: a multi-modal attention graph network with dynamic routing-by-agreement for multi-label emotion recognition. Knowl-Based Syst 283:111126

[15] Peng C, Chen K, Shou L, Chen G (2024) CARAT: Contrastive feature reconstruction and aggregation for multi-modal multi-label emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 38. pp 14581–14589

[16] Picard RW (2000) Affective computing. MIT press

[17] Plutchik R (1980) A general psychoevolutionary theory of emotion. In: Theories of emotion. Elsevier, pp 3–33

[18] Poria S, Cambria E, Gelbukh A (2015) Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of EMNLP. pp 2539–2544

[19] Poria S, Majumder N, Mihalcea R, Hovy E (2019) Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE Access 7:100943–100953

[20] Scherer KR (2003) Vocal communication of emotion: A review of research paradigms. Speech Commun 40(1-2):227–256

[21] Schuller B, Steidl S, Batliner A, et al (2013) The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. Proceedings of the INTERSPEECH 2013

[22] Song C, Anwer S, Yilahun H, et al (2024) ERT: Emotion in Real-Scene Transformer for Speech Emotion Recognition. In: International Conference on Intelligent Multilingual Information Processing. Springer, pp 189–204

[23] Tang X, Huang J, Lin Y, et al (2025) Speech emotion recognition via CNN-transformer and multidimensional attention mechanism. Speech Commun 103242

[24] Tsai YHH, Liang PP, Zadeh A, et al (2018) Learning factorized multimodal representations. arXiv preprint arXiv:1806.06176

[25] Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30

[26] Vidrascu L, Devillers L (2005) Annotation and detection of blended emotions in real human-human dialogs recorded in a call center. In: 2005 IEEE International Conference on Multimedia and Expo. IEEE, pp 4–pp

[27] Wang W, Huang M, Xu XS, et al (2018) Chat more: Deepening and widening the chatting topic via a deep model. In: The 41st International ACM SIGIR Conference. pp 255–264

[28] Zadeh AB, Liang PP, Poria S, et al (2018) Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the ACL. pp 2236–2246

[29] Zhang D, Ju X, Li J, et al (2020) Multi-modal multi-label emotion detection with modality and label dependence. In: Proceedings of EMNLP. pp 3584–3593

[30] Zhang D, Ju X, Zhang W, et al (2021) Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 35. pp 14338–14346

[31] Zhang Y, Chen M, Shen J, Wang C (2022) Tailor versatile multi-modal learning for multi-label emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 36. pp 9100–9108

[32] Zhao J, Zhang T, Hu J, et al (2022) M3ED: Multi-modal multi-scene multi-label emotional dialogue database. arXiv preprint arXiv:2205.10237

Declarations

- **Funding**: This work was supported by the Science and Technology Plan Project of the Autonomous Region - Natural Science Foundation (General Project) under Grant No. 2023D01C19.

- **Conflict of interest**: The authors declare that they have no conflict of interest.
- **Ethics approval**: Not applicable.
- **Consent to participate**: Not applicable.
- **Consent for publication**: Not applicable.
- **Availability of data and materials**: The datasets used during the current study are available from Github.
- **Code availability**: The code is available at https://github.com/Songsong0824/TAC-Net.git
- **Authors' contributions**: Chaobo Song wrote the main manuscript text and Hankiz Yilahun, Askar Hammdulla prepared figures 1-5 and Table 1-2. All authors reviewed the manuscript.