Multimodal Emotion Recognition Based on a Dual-Stream MFCC-MobileNetV2 Architecture

Qiong Yang, Aodi Lv, Feng Liu

Abstract—**To tackle the challenge of efficiently extracting discriminative features in multimodal emotion recognition, this paper proposes a multimodal emotion recognition framework based on a dual-stream MFCC-MobileNetV2 architecture, aiming to enhance the expressive capacity of emotional features. The framework integrates both visual and audio modalities: the visual stream utilizes an improved MobileNetV2 model to extract spatial features of facial expressions, while the audio stream incorporates the Fractional Fourier Transform (FRFT) to enhance MFCC features, thereby better capturing the characteristics of non-stationary speech signals. For feature fusion, a residual network based on standard Conv1D is designed and combined with an attention mechanism to achieve dynamic cross-modal feature weighting. Experimental results show that the proposed fusion strategy achieves improvements of 10.62% and 10.99% on the RAVDESS and CREMA-D datasets, respectively, outperforming other methods in the same category. This provides an efficient technical approach and new research insights for multimodal emotion recognition.**

Index Terms—**emotion recognition, facial features, audio features, modality fusion.**

I. Introduction

EMOTION recognition refers to the identification of human emotional states through diverse sources of information, including facial expressions, speech, text, and physiological signals. In recent years, driven by rapid advances in artificial intelligence, emotion recognition has found widespread applications in fields such as human-computer interaction, healthcare, and education. Traditional approaches typically rely on a single modality—such as audio, visual, or textual data—limiting their ability to fully capture the complexity and nuance of human emotions. Since emotional experiences are inherently multimodal and context-dependent, unimodal methods often fall short in accuracy and robustness. To address these limitations, multimodal emotion recognition has emerged as a promising direction, integrating complementary information from multiple modalities—such as visual, auditory, and linguistic cues—to enable more accurate, reliable, and comprehensive emotion classification.

Emotion, as one of the most fundamental forms of human expression, exhibits significant variation across cultural contexts. These differences are primarily reflected in the norms and intensity of emotional display—some cultures encourage open and direct expression, while others value subtlety and emotional restraint. Furthermore, the conceptualization and categorization of emotions can vary across societies, influenced by cultural norms and social values. Such cross-cultural diversity in emotional expression presents a key challenge for the design and development of robust, generalizable emotion recognition systems, particularly in ensuring their cultural sensitivity and global applicability.

In the context of multimodal emotion recognition, this paper proposes a hybrid feature extraction and fusion framework built upon existing research. The framework integrates MobileNetV2 for visual feature extraction, standard Conv1D residual blocks for temporal modeling, an enhanced MFCC-based module for acoustic analysis, and an attention-based fusion mechanism to facilitate deep interaction and effective integration of multimodal information. This design promotes greater complementarity and efficient utilization of features across modalities, thereby enhancing both recognition accuracy and model generalization. The main contributions of this work are summarized as follows:

1. Multimodal Feature Extraction Framework Design: The visual modality utilizes MobileNetV2 to extract spatial features, combined with Conv1D residual blocks for temporal modeling. The audio modality incorporates the Fractional Fourier Transform (FRFT) to enhance the representation of time-frequency features. Finally, feature-level fusion is achieved through an attention mechanism for emotion classification.

2. Replacing Traditional Fourier with Fractional Fourier: The Fractional Fourier Transform is employed in place of the traditional Fourier Transform to enhance the ability to capture non-stationary speech features, thereby improving the expressiveness of acoustic features in emotion recognition.

3. Lightweight Visual Feature Extraction Module: A lightweight visual feature extraction module based on MobileNetV2 is introduced. By leveraging its optimized inverted residual structure and linear bottleneck design, the model achieves efficient inference with improved accuracy and speed.

4.Standard Conv1D Residual Blocks: Residual 1D convolutional structures are employed in both the visual and audio modalities to enhance feature reuse and cross-layer propagation. This design effectively improves the model's capability and stability in capturing temporal dependencies.

5.Feature-Level Fusion Strategy: In the multimodal emotion recognition process, a feature-level fusion strategy is adopted. By applying an attention mechanism to achieve

Manuscript received June 1, 2025; revised Aug 16, 2025. This research was funded by the 2024 Youth Innovation Team Project under the Scientific Research Program of Shaanxi Provincial Department of Education (Project Number: 24JP069).

Qiong Yang is a lecturer at the Computer Science School, Xi'an Polytechnic University, Shaanxi, 710048, P. R. China. (e-mail: yangqiong@xpu.edu.cn).

Aodi Lv is a postgraduate student at the Computer Science School, Xi'an Polytechnic University, Shaanxi, 710048, P. R. China. (corresponding author, e-mail: lv1195149724@163.com).

Feng Liu is a professor at the Computer Science School, Xi'an Polytechnic University, Shaanxi, 710048, P. R. China. (e-mail: liufeng@xpu.edu.cn).

weighted fusion of visual and audio modalities, this approach enhances efficient inter-modal interaction while avoiding redundant computation caused by separate branch inference. It significantly improves the model's ability to recognize complex emotional states.

II. Related work

In recent years, the rapid advancement of artificial intelligence has propelled human-computer interaction (HCI) to the forefront of research interest. As a key component of HCI, emotion analysis has evolved significantly—shifting from early unimodal approaches to more sophisticated multimodal strategies. These modern methods perform comprehensive analysis by integrating information from diverse sources, such as speech [1], text [2], visual cues [3], and even physiological signals like electroencephalography (EEG) [4]. Nevertheless, effectively processing and fusing multimodal information to achieve accurate emotion recognition and reliable decision-making remains a central challenge in the field.

In the field of emotion recognition, traditional machine learning approaches—such as Decision Trees (DT) and Support Vector Machines (SVM)—have been widely adopted. With the advancement of deep learning, more powerful models, including Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and their variants (e.g., GRU), have been extensively employed to capture complex patterns in emotional data. These deep architectures have significantly enhanced feature representation and classification performance, leading to substantial progress in unimodal emotion recognition.In the area of text-based emotion recognition, Xu et al. [5] proposed a CNN_Text_Word2vec model based on CNN, which improved the overall accuracy by 7% compared to mainstream methods.Yu et al. [6] propose a model averaging ensemble of Convolutional Neural Networks (CNNs) that consolidates multiple pre-trained CNN models. Transfer learning is first performed by replacing the classification layer with a Multilayer Perceptron (MLP). The models are then fine-tuned on the dataset to adapt to the facial expression recognition task and further optimized.Yuan et al. [7] combined LPCC and MFCC as text-independent speaker recognition features in their system. The experiment used Vector Quantization (VQ) and Dynamic Time Warping (DTW) for identity recognition, demonstrating that the combination of LPCC and MFCC achieved a higher recognition rate.Research has shown that multimodal emotion recognition—which integrates information from multiple modalities—offers greater advantages over unimodal approaches due to the richness and complementarity of multimodal data. Scholars at home and abroad have conducted in-depth studies on this topic. For example, Griol D et al. [8] proposed an emotion recognition method that evaluates transfer learning in speech recognition and adopts a dual-LSTM structure for facial emotion recognition, achieving improved performance through a late fusion strategy. Wang et al. [9] introduced a new Fourier Parameter (FP) model for speaker-independent speech emotion recognition, using perceptual quality features and first- and second-order differences, when combined with MFCC, the recognition rate increased by 10.5%. Yoon S et al. [10] proposed a novel deep dual recurrent encoder model that encodes both audio and text sequences using dual RNNs, outperforming previous state-of-the-art methods in emotion classification tasks by providing more accurate label assignment. Mittal T et al. [11] proposed the M3ER method, which fuses multiple co-occurring modalities using a novel, data-driven multiplicative fusion approach, achieving about a 5% improvement over previous methods. Although ongoing research continues to explore increasingly sophisticated models, several critical challenges persist in multimodal emotion recognition. In feature extraction, inadequate modeling of modality-specific characteristics can lead to suboptimal representations, resulting in inaccurate predictions and reduced robustness. With regard to feature fusion, despite the adoption of advanced fusion strategies in existing multimodal frameworks, many approaches still fail to fully capture and leverage the complementary nature of cross-modal information. This often leads to underutilization of discriminative features, thereby limiting the overall effectiveness and performance of the recognition system.

III. Methods

We propose a multimodal neural network designed for the joint processing and classification of audio and video data. The architecture employs a dual-stream heterogeneous structure, comprising two independent pathways that separately learn audio and visual features. In the audio branch, Mel-Frequency Cepstral Coefficients (MFCCs) [12] are used as input features, processed through a four-layer 1D convolutional block. Each block consists of a 1D convolution layer, batch normalization, ReLU activation, and max pooling, effectively capturing salient temporal features. In the visual branch, an enhanced MobileNetV2 serves as the feature extraction backbone, followed by a convolutional module structurally analogous to that in the audio pathway.The fusion module is positioned at the end of both streams, where audio and visual features are concatenated and further processed along the temporal dimension using four additional 1D convolutional blocks to capture cross-modal correlations. To enable adaptive and context-aware fusion, an attention mechanism is applied after the fusion layers, dynamically recalibrating the contribution of audio and visual features according to their relevance. Finally, the refined fused representation is passed to a fully connected (linear) classifier for emotion classification. A detailed schematic of the proposed framework is illustrated below:

A. Preprocessing

In multimodal emotion recognition, data preprocessing [13] plays a foundational role, significantly influencing the model's generalization ability and classification accuracy. Due to the inherently coupled nature of emotional expressions across modalities, the task involves integrating heterogeneous data sources—such as speech (spectral-temporal signals), vision (spatiotemporal dynamics), and text (semantic representations). These modalities exhibit substantial differences in data format, feature dimensionality, temporal resolution, and statistical

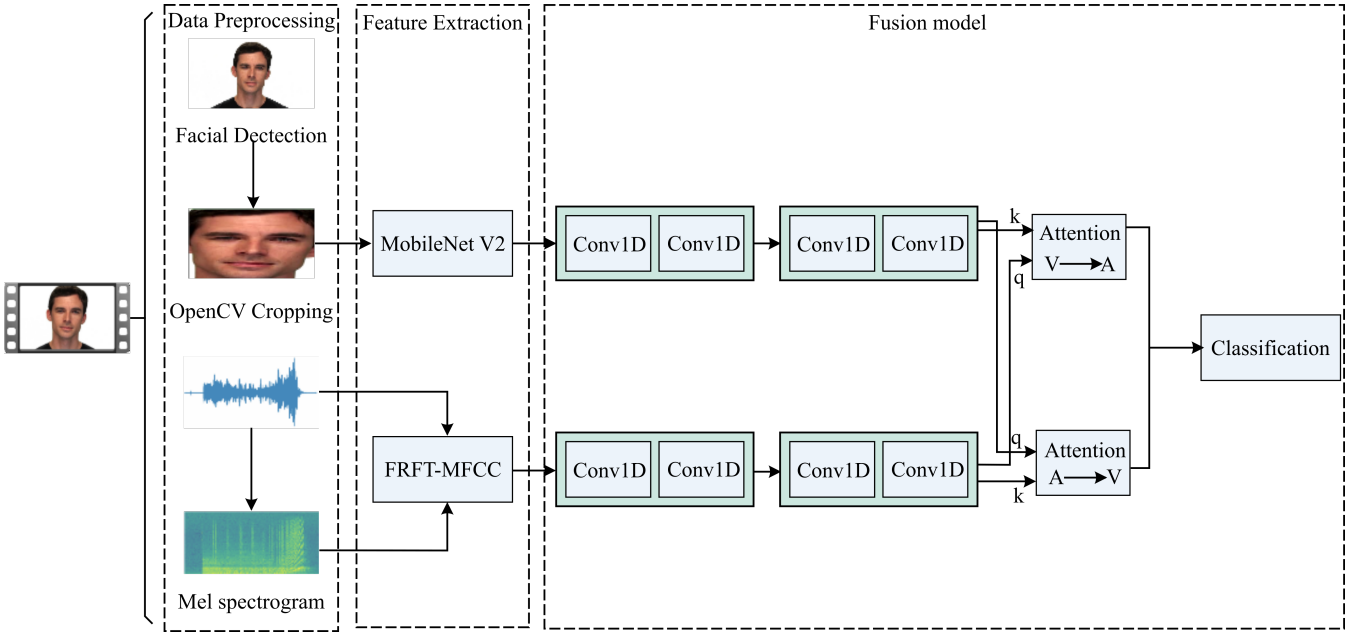

Fig. 1. Multimodal network model framework diagram.

properties, posing significant challenges for direct integration. Therefore, systematic and modality-specific preprocessing is crucial to align and normalize the data, ensuring effective fusion and robust model performance.To mitigate cross-modal inconsistencies and enhance the effectiveness of feature fusion, a comprehensive preprocessing pipeline is essential. This pipeline typically comprises three key components: feature decoupling, dimensional alignment, and distribution normalization. By systematically addressing modality-specific disparities, this preprocessing workflow constructs a normalized, temporally aligned, and structurally coherent multimodal feature space—laying a solid foundation for downstream tasks such as attention-based fusion, cross-modal representation learning, and emotion classification. Empirical studies have shown that well-designed preprocessing strategies can significantly boost performance in multimodal emotion recognition, particularly in challenging scenarios involving missing modalities, conflicting cross-modal signals, or imbalanced data distributions.

1) Facial expression preprocessing

First, the MTCNN model [14] is employed to detect faces in video frames, accurately localizing facial bounding boxes and extracting standardized regions of interest (ROIs). Subsequently, a fixed number of frames are uniformly sampled from the video sequence to ensure comprehensive coverage of dynamic facial expressions over time. Next, multimodal data normalization is performed: facial images are converted to grayscale to mitigate illumination variations, color channel ordering is standardized, and a secondary pass of MTCNN refinement is applied to enhance localization accuracy and spatial consistency. All images are uniformly resized to a resolution of 224×224 pixels using bilinear interpolation. The processed data is then serialized and stored either as NumPy arrays or AVI video streams to facilitate subsequent feature extraction and model inference. To address frame synchronization issues, a zero-padding strategy is applied to preserve temporal continuity across

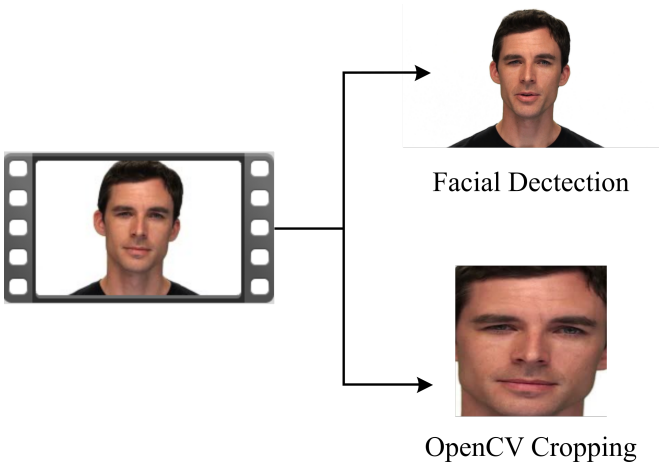

Fig. 2. Facial Preprocessing

sequences. Additionally, an exception handling mechanism is implemented to detect and log processing errors, such as video decoding failures, ensuring robustness and traceability in the data pipeline.

2) Speech preprocessing

This procedure is designed to preprocess audio files in the multi-speaker dataset by standardizing the duration of all clips to 3.6 seconds. First, audio samples are iterated within each speaker's directory to identify valid files. The Librosa library is then used to load the audio: clips shorter than 3.6 seconds are extended with zero-padding, while longer clips are center-cropped to retain the core speech content and preserve acoustic integrity. Finally, the processed files are saved in a standardized format. This uniform temporal alignment enhances the consistency of input data, thereby improving the accuracy and reliability of subsequent feature extraction and model training.

B. Feature extraction and fusion

In everyday life, humans express emotions through a rich variety of modalities—ranging from explicit verbal statements to subtle behavioral cues. In social interactions, emotions are often conveyed through emojis in instant messaging or implicitly expressed in carefully curated content such as sunset photos shared on social media platforms. As a core component of affective computing, multimodal emotion recognition utilizes deep learning and pattern recognition techniques to intelligently interpret human affective states. While emotions can be manifested across multiple dimensions, speech and facial expressions remain the most direct and effective channels for emotional communication. This paper focuses on bimodal emotion recognition by fusing audio (speech) and visual (facial) cues, proposing a unified framework that integrates both hierarchical feature extraction and adaptive feature fusion to enhance recognition performance.

In the proposed bimodal architecture, the visual and audio branches employ modality-specific module groups tailored to their respective input characteristics. The visual branch consists of four cascaded convolutional module groups. Each group integrates a standard 1D residual block with a 3×3 kernel, followed by batch normalization and a ReLU activation function, enabling effective spatial feature learning. The audio branch adopts a symmetric structure but includes an additional max-pooling layer with a stride of 2 in each module to enhance temporal modeling and progressively reduce feature dimensionality, which is critical for capturing long-range acoustic patterns.The system's key architectural parameters are summarized in Table I and Table II, where k denotes kernel size, d represents the number of filters, and s indicates stride. This design achieves a balance between modality-specific optimization and structural consistency, facilitating more coherent and effective feature fusion in subsequent stages.

TABLE I
EFFICIENTFACE BRANCH ARCHITECTURE.

	EfficientFace branch
Stage1	Conv1D[k=3,d=64,s=1] + BN1D + Relu Conv1D[k=3,d=64,s=1] + BN1D + Relu
Stage2	Conv1D[k=3,d=128,s=1] + BN1D + Relu Conv1D[k=3,d=128,s=1] + BN1D + Relu
Predict	Global Average Pooling + Linear

TABLE II
AUDIO BRANCH ARCHITECTURE.

	Audio branch
Stage1	Conv1D[k=3,d=64,s=1] + BN1D + Relu + MaxPool1d[K=2] Conv1D[k=3,d=64,s=1] + BN1D + Relu + MaxPool1d[K=2]
Stage2	Conv1D[k=3,d=64,s=1] + BN1D + Relu + MaxPool1d[K=2] Conv1D[k=3,d=64,s=1] + BN1D + Relu + MaxPool1d[K=2]
Predict	Global Average Pooling + Linear

The proposed residual block, built upon standard 1D convolutional layers, offers significant advantages over conventional Conv1D layers, particularly in mitigating common challenges such as vanishing gradients and network degradation in deep architectures. By introducing residual connections, the residual structure enables direct propagation of information across layers, enhancing feature flow, accelerating convergence, and improving training stability. Unlike simple stacked convolutional layers, this design preserves the model's representational capacity while facilitating the construction of deeper networks for temporal modeling. This capability is crucial for capturing complex temporal dynamics and cross-modal interactions inherent in multimodal emotion recognition tasks, as illustrated in Figure 3.

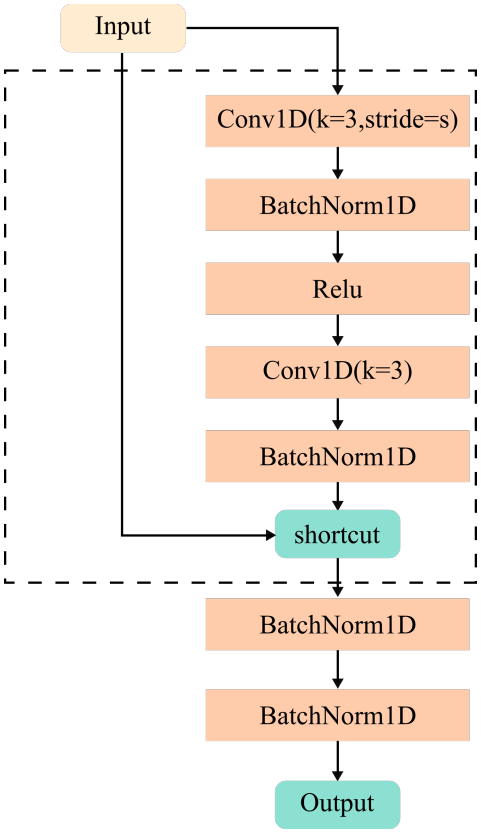

Fig. 3. Residual block based on standard Conv1D

Furthermore, standard 1D convolutions are inherently effective at modeling temporal dynamics in sequential data such as audio signals and facial expression variations. When integrated with residual mechanisms, they not only preserve strong local feature extraction capabilities but also enhance the network's capacity to capture long-range temporal dependencies. This combination yields a architecture with high generalizability and scalability, making it well-suited for multimodal emotion recognition tasks and providing a solid foundation for the development of deeper, more effective temporal fusion networks.

1) Video feature extraction

In the domain of video-based feature extraction, we propose a novel hybrid architecture that synergistically combines the spatial feature learning capability of MobileNetV2 with the temporal modeling strength of 1D convolutional networks. To meet the high demands of computational efficiency in video processing, MobileNetV2—a lightweight convolutional neural network—is employed as the spatial backbone. Its effectiveness stems from the integration of depthwise

separable convolutions and residual connections, which significantly reduce computational cost while preserving rich feature representation.Compared to MobileNetV1 [15], MobileNetV2 [16] introduces two key innovations: the inverted residual block and the linear bottleneck. The inverted residual structure adopts an ”expand–convolve–compress” strategy: input features are first expanded into a higher-dimensional space using pointwise convolutions, processed via depthwise convolutions, and then projected back to a lower-dimensional output. This enables richer feature learning within a compact computational footprint. Additionally, the linear bottleneck removes non-linear activation functions (e.g., ReLU6) in the final projection layer, preserving more discriminative information and alleviating representational collapse in low-dimensional spaces.This dual design not only ensures high efficiency and scalability but also enhances gradient flow and mitigates feature degradation in deep networks. As a result, MobileNetV2 delivers a robust and semantically rich spatial feature representation, forming a stable foundation for subsequent 1D convolutional modules that model the temporal dynamics of emotional expressions.

In conventional residual blocks, dimensionality is typically reduced via a 1×1 convolution, followed by a 3×3 standard convolution, and then expanded again using another 1×1 convolution. MobileNetV2 inverts this design: it first expands the channel dimension using a pointwise convolution, applies a 3×3 depthwise separable convolution to capture spatial features efficiently, and finally compresses the output back to a lower-dimensional representation. As illustrated in Figure 4, the shortcut (residual) connection is preserved only when the stride is 1, ensuring identity mapping. When the stride is 2 (indicating spatial downsampling), the residual path is omitted, and the block operates in a plain sequential manner. This architectural refinement enables MobileNetV2 to achieve a superior balance between representational power and computational efficiency. By decoupling feature transformation from channel mixing and minimizing redundant computations, the model maintains high expressiveness while significantly reducing parameter count and FLOPs. These advantages make it especially well-suited for real-time multimodal emotion recognition, where accurate modeling of both spatial details and temporal dynamics is required under resource-constrained conditions.

Depthwise separable convolution [17], a cornerstone innovation in lightweight convolutional neural networks, enables highly efficient feature learning by decoupling the spatial and channel-wise computations that are inherently entangled in standard convolutions. This operation employs a two-stage decomposition:Depthwise convolution processes each input channel independently, applying a single spatial filter per channel to extract localized spatial features. This drastically reduces spatial computation and parameter load. Pointwise convolution (a 1×1 convolution) then integrates information across channels, performing channel-wise feature combination and dimensionality transformation to preserve representational capacity. Compared to standard 3×3 convolutions, this factorized approach reduces both the number of parameters and computational cost by approximately N times, where N is the number of output channels. By significantly lowering

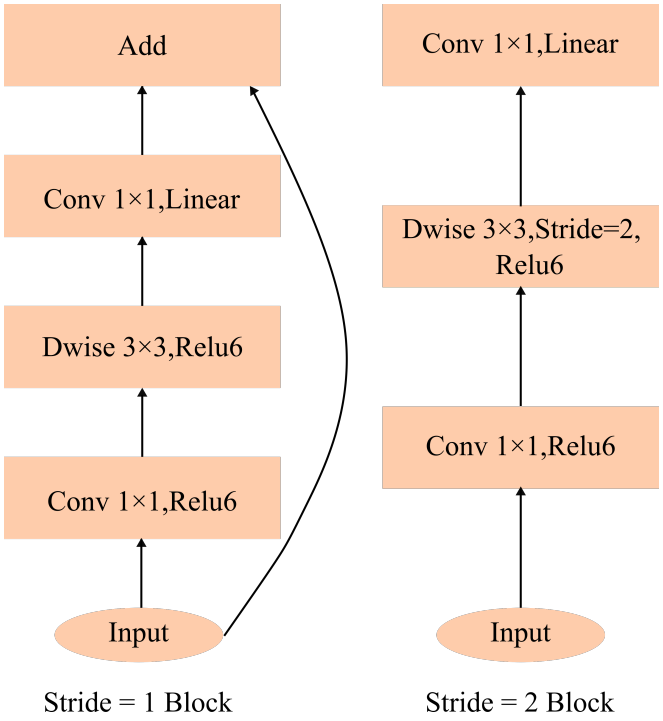

Fig. 4. MobileNetV2’s inverted residual structure

model complexity while maintaining competitive accuracy, depthwise separable convolution achieves an optimal trade-off between performance and efficiency—making it particularly well-suited for mobile vision tasks and edge computing applications. Figure 5 illustrates the architecture of the depthwise separable convolution module.

In the inverted residual structure, the final convolution layer removes the nonlinear activation function (e.g., ReLU), retaining only a linear transformation. This design performs nonlinear transformations in the high-dimensional expanded space while preserving linearity in the low-dimensional projection stage. By avoiding nonlinearities in the compressed representation, the model prevents potential distortion of critical features and retains more of the original information. This helps mitigate representational collapse and enhances gradient flow, ultimately contributing to improved feature fidelity and model accuracy.

2) Audio feature extraction

In speech emotion recognition, commonly used acoustic features are typically categorized into three main types: prosodic features (e.g., pitch, energy, and speaking rate), voice quality features (e.g., jitter, shimmer, and harmonics-to-noise ratio), and spectral features derived from frequency-domain analysis. In this study, Mel-Frequency Cepstral Coefficients (MFCCs) are adopted as the primary acoustic representation due to their effectiveness in capturing phonetically relevant information. MFCCs are extracted using established toolkits such as pyAudioAnalysis, Librosa, and openSMILE, ensuring robustness and compatibility with standard feature extraction pipelines.

The Fractional Fourier Transform (FRFT) [18], as a generalized mathematical extension of the classical Fourier Transform, overcomes the limitations of traditional frequency-domain analysis by enabling continuous representation of signals in the joint time–frequency

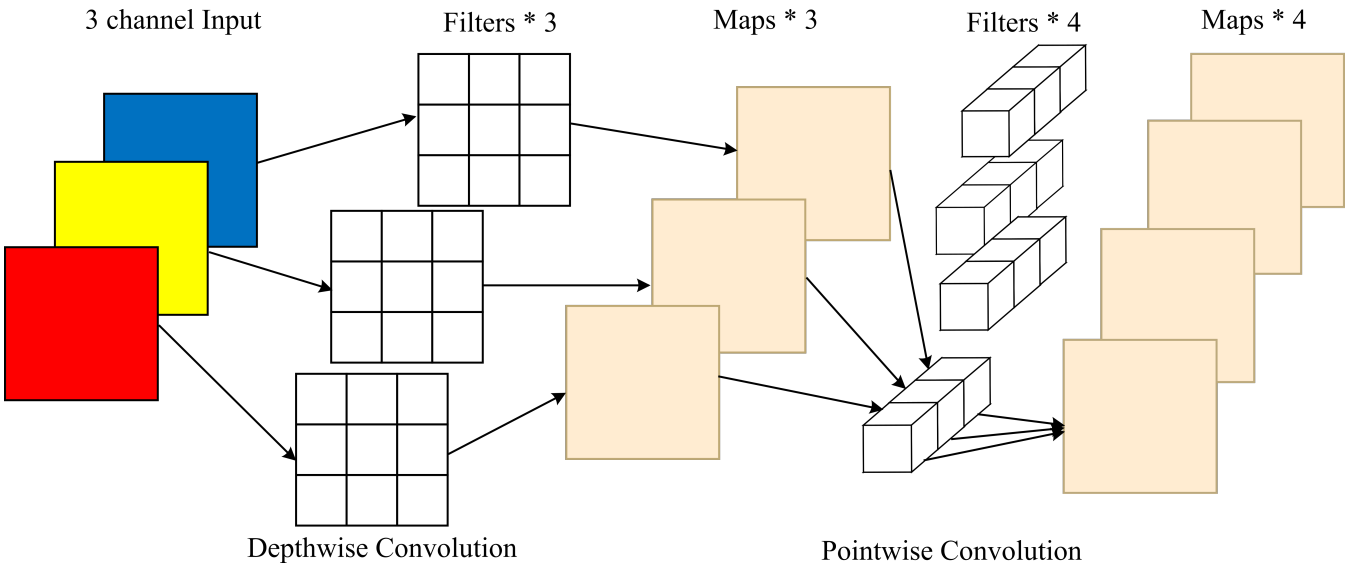

Fig. 5. Depthwise Separable Convolution Calculation Process

domain. Its core principle lies in introducing a fractional order parameter α, which defines a rotational coordinate system on the time–frequency plane. When $\alpha = 0$, the FRFT corresponds to the original time-domain signal; when $\alpha = 1$, it reduces to the classical Fourier transform; and for fractional values of α, it represents an intermediate domain between time and frequency. This property allows FRFT to project signal energy along any angle in the time–frequency plane, thereby establishing a unified mathematical framework for joint time–frequency analysis.Unlike the standard Fourier transform that provides only global frequency information, the FRFT demonstrates significant advantages in analyzing non-stationary signals, particularly those with time-varying frequency components, such as chirp signals, speech waveforms, and EEG signals. By optimizing the fractional order α, the FRFT adaptively identifies the optimal focusing direction of energy distribution in the time–frequency domain, enabling sparse representations of signals within specific fractional domains. This adaptive focusing capability enhances the resolution of time-varying feature extraction and provides more compact and discriminative signal representations for subsequent tasks such as classification or recognition.

The Fractional Fourier Transform (FRFT) [18], a generalized extension of the classical Fourier Transform, overcomes the inherent limitations of conventional frequency-domain analysis by enabling a continuous representation of signals in the joint time–frequency domain. At its core, FRFT introduces a fractional order parameter α, which effectively rotates the coordinate system on the time–frequency plane. When $\alpha = 0$, the transform yields the original time-domain signal; when $\alpha = 1$, it reduces to the standard Fourier Transform; and for intermediate values of $\alpha \in (0, 1)$,it produces a hybrid domain representation that interpolates between time and frequency.This rotational perspective allows the FRFT to project signal energy along arbitrary directions in the time–frequency plane, establishing a unified framework for analyzing transient and non-stationary signals. Unlike the traditional Fourier Transform—which provides only global spectral information—the FRFT excels in characterizing signals with time-varying frequency content, such as chirps, speech, and EEG signals. By optimizing the fractional order α, the transform can adaptively identify the domain in which the signal's energy is most concentrated, thereby achieving a sparse and highly focused representation.This adaptive energy-focusing capability significantly enhances the resolution of time-varying feature extraction, yielding more compact and discriminative representations that benefit downstream tasks such as classification, pattern recognition, and multimodal emotion analysis.

Moreover, the Fractional Fourier Transform (FRFT) inherently supports multi-scale analysis through the adjustment of its fractional order. Lower-order domains preserve fine temporal details, making them well-suited for capturing transient signal characteristics, whereas higher-order domains emphasize spectral features, enabling the extraction of stable frequency structures. By tuning the fractional order parameter α, FRFT can adaptively select the most appropriate time–frequency representation according to the nature of the signal and the specific task requirements. This flexibility establishes FRFT as a powerful and adaptive framework for time–frequency analysis. As a result, it has demonstrated significant theoretical value and practical potential in a range of complex signal processing applications, including speech emotion recognition, neural signal decoding, and image analysis.

The Fractional Fourier Transform (FRFT) enables multi-perspective time-frequency analysis by rotating the coordinate axes in the time-frequency plane. This rotation allows the signal to be viewed from different angles, thereby providing a more comprehensive representation of its time-varying characteristics. The p-th order FRFT of a signal $x(t)$ is defined as:

$$X_P(u) = F^p[x(t)] = \int_{-\infty}^{+\infty} K_p(t,u)x(t)dt, \qquad (1)$$

in which $K_p(t,u)$ is the kernel function of the FRFT, expressed as follows:

$$K_p(t,u) = \begin{cases} A_P exp\left(j\frac{u^2+t^2}{2}\cot\alpha - \frac{jut}{\sin\alpha}\right), \alpha \neq n\pi \\ \delta(u-t), \quad \alpha = 2n\pi \\ \delta(u+t), \quad \alpha = (2n+1)\pi, \end{cases} \tag{2}$$

$$A_p = \sqrt{\frac{1-j\cot\alpha}{2\pi}}, \tag{3}$$

where A_P is the amplitude factor, $\alpha = p\pi/2$ is the rotation angle, and is the Dirac delta function. The superior properties of the FRFT stem from the flexibility of the rotation angle α .

3) Feature-level fusion (based on attention mechanism)

Multimodal fusion is a key paradigm in intelligent systems that enhances perception and decision-making by integrating complementary information from heterogeneous modalities—such as images, audio, and text. Its primary advantage lies in overcoming the limitations of unimodal analysis through the exploitation of semantic correlations and complementary features across modalities, enabling the construction of a more expressive, robust, and generalizable joint representation. This approach has been widely adopted in complex tasks including emotion recognition, human-computer interaction, and medical diagnosis.Currently, mainstream multimodal fusion strategies can be broadly categorized into three technical pathways:Early fusion (feature-level fusion) [19]: This approach concatenates raw or low-level features from different modalities at the input or early processing stage. While simple and computationally efficient, it is often challenged by modality heterogeneity and misalignment in feature spaces. Intermediate fusion (representation-level fusion) [20]: In this strategy, high-level features are first extracted independently from each modality and then fused at intermediate layers of a deep network. This allows the model to capture rich cross-modal semantic interactions and structural complementarity, making it a dominant approach in recent research. Late fusion (decision-level fusion) [21]: Each modality is processed by a separate model, and their individual predictions are combined at the decision level—typically via weighted averaging, majority voting, or probabilistic fusion. Although flexible and robust to modality-specific noise, this method often fails to model deep inter-modal dependencies.

In this study, we introduce an attention-based fusion strategy to enhance the adaptability and discriminative capability of the model, particularly in audio-visual emotion recognition tasks. Specifically, audio and visual features are first preprocessed and independently encoded, then concatenated along the feature dimension to form an initial multimodal joint representation. This representation is fed into an attention module that dynamically learns modality-specific weighting coefficients based on the input context. The original features are subsequently re-weighted using these learned scores, enabling the model to selectively emphasize the modality that is more informative for the current emotional state. The re-weighted features are then combined—either through summation or concatenation—and passed to the classifier for final prediction.This fusion approach not only preserves the integrity of unimodal representations but also effectively captures context-dependent cross-modal interactions. By adaptively focusing on the most relevant sensory cues, the strategy achieves a favorable balance between performance and efficiency, significantly improving classification accuracy and robustness in multimodal emotion recognition without incurring substantial computational overhead.

The attention mechanism, as a cornerstone innovation in modern deep learning architectures, enables the dynamic and adaptive weighting of features through differentiable computation. It operates by projecting each element of the input sequence into three distinct vector spaces: Query (Q), Key (K), and Value (V). These vectors facilitate the modeling of long-range dependencies and contextual interactions across different positions in the sequence by computing weighted correlations—where the Query and Key vectors determine the attention scores (representing relevance or alignment), and the Value vectors are aggregated according to these scores to produce the output. The core computation can be formally described as follows:

$$Q = XW^Q, \tag{4}$$

$$K = XW^K, \tag{5}$$

$$V = XW^V, \tag{6}$$

where W^Q, W^K, W^V are learnable parameter matrices, and d denotes the dimensionality of the input features. Through these linear transformations, the model is able to dynamically capture the interactions among elements in the input sequence. This significantly enhances the model's capacity to represent complex dependency structures and contextual information.

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{7}$$

Q represents Query, K represents Key, V represents Value, and d denotes the dimension size.

This mechanism overcomes the limitations of traditional CNNs—such as restricted local receptive fields—and addresses the sequential processing bottlenecks inherent in RNNs. It enables the modeling of long-range or global dependencies between elements at arbitrary positions within the input sequence.

IV. Experimentation and Results

A. Datasets

1) RAVDESS

This dataset contains 7,356 audio files collected from 24 professional actors (12 males and 12 females). These actors expressed various emotional states—including happiness, sadness, fear, surprise, disgust, anger, and calmness—using a neutral accent.

2) CREMA-D

The dataset consists of 7,442 video and audio samples from 91 actors (48 male and 43 female). The actors perform 12 carefully selected sentences, each expressed with six different emotions and four levels of emotional intensity.

B. Evaluation metrics

In multimodal emotion recognition research, accuracy is a fundamental metric for evaluating model classification performance, reflecting the degree of consistency between predicted outputs and ground-truth labels. A widely adopted variant, Top-k accuracy, measures whether the true label is included among the k most probable predicted classes, ranked by their confidence scores. In particular, Top-1 accuracy—the most commonly reported metric—indicates that the class with the highest predicted probability matches the actual label. This hierarchical evaluation framework is applicable not only to binary classification tasks but is especially valuable in multi-class settings, where class ambiguity and emotional nuance are more pronounced. By providing a more nuanced and robust assessment of model performance, Top-k accuracy serves as a critical benchmark for guiding architectural design and hyperparameter optimization.

C. Experimental Results and Analysis

1) The RAVDESS dataset experiment

To evaluate the effectiveness of the proposed multimodal emotion recognition method, we use classification accuracy on the test set as the primary evaluation metric. To ensure the reliability and reproducibility of the results, all experiments are conducted on the RAVDESS dataset using consistent training procedures and hyperparameter settings. Experimental results demonstrate that the bimodal fusion approach achieves significantly higher accuracy compared to unimodal baselines. This improvement further underscores the benefits of integrating complementary information across modalities in emotion recognition tasks. As shown in Table III, the consistent performance gain validates the effectiveness of the proposed method and highlights the potential of multimodal fusion for enhancing affective state classification.

TABLE III
EMOTION RECOGNITION ACCURACY UNDER DIFFERENT MODALITIES.

Types of modalities	ACC(%)
Audio unimodal	62.50
Video unimodal	79.79
Multimodal	90.62

To evaluate the effectiveness of the proposed model, we conducted a comprehensive comparison with state-of-the-art methods reported in the literature. On the RAVDESS dataset, our approach achieves an emotion recognition accuracy of 90.62%, outperforming the methods presented in references [22], [23], [24], [25]. This performance gain demonstrates the superiority of the designed multimodal neural network architecture and fusion strategy, particularly in capturing cross-modal feature alignment and leveraging complementary information across modalities. The results further confirm the strong potential and practical effectiveness of the proposed method in real-world emotion recognition applications. A summary of the comparative experimental results on the RAVDESS dataset is provided in Table IV.

TABLE IV
COMPARISON WITH RESULTS OF OTHER METHODS.

Model	ACC(%)
CFN-SR[22]	75.56
STA-CNN[23]	76.39
WaDER[24]	81.45
MultiMAE-DER[25]	83.61
Ours	90.62

TABLE V
PERFORMANCE COMPARISON OF THE FEATURE FUSION MODEL WITH OTHER MODELS ON THE RAVDESS DATASET.

Feature extraction methods	Fusion method	ACC(%)
MFCC	Intermediate feature fusion	80.00
MFCC+MobileNetV2	Intermediate feature fusion	82.29
MFCC+MobileNetV2	Feature-level fusion	84.37
MFCC(FRFT)+MobileNetV2	Feature-level fusion	87.91
MFCC(FRFT)+Conv1D +MobileNetV2	Feature-level fusion	90.62

To validate the effectiveness of the proposed model components, we conducted a series of ablation and comparative experiments, with results summarized in Table V. The experimental results show that incorporating the improved MFCC—enhanced by the Fractional Fourier Transform (FRFT)—into the audio feature extraction module, combined with a lightweight visual feature extractor based on MobileNetV2, significantly boosts emotion recognition accuracy. Within each modality, residual blocks constructed from 1D convolutional layers (Conv1D) are employed to facilitate cross-layer feature propagation, effectively capturing the temporal dynamics and expressive variations of emotional signals. Regarding fusion strategies, feature-level fusion outperforms intermediate (representation-level) fusion. By enabling early cross-modal interaction through direct feature concatenation, this approach achieves efficient information integration, reduces computational overhead, and lowers model complexity—while preserving high recognition performance. Ultimately, the proposed architecture, which fuses enhanced MFCC and MobileNetV2 features at the feature level, achieves a recognition accuracy of 90.62% on the RAVDESS dataset. These results confirm the effectiveness and advancement of the integrated multimodal feature extraction and fusion framework.

2) The CREMA-D dataset experiment

Compared to the methods proposed in [26], [27], [28], [29], the proposed approach achieves higher accuracy, demonstrating superior performance in multimodal emotion recognition. This improvement highlights the effectiveness of the designed feature extraction and fusion strategy in capturing discriminative cross-modal patterns. A comprehensive performance comparison on the CREMA-D dataset, including the proposed method and other state-of-the-art models, is presented in Table VI.

Table VII systematically presents the performance evaluation results of the proposed multimodal emotion recognition method on the CREMA-D dataset.

V. CONCLUSIONS

To address the limitations in feature representation capacity in current multimodal emotion recognition, this

TABLE VI
COMPARISON WITH RESULTS OF OTHER METHODS.

Model	ACC(%)
DCNN[26]	75.56
CNN,Transformer[27]	78.70
CNN [28]	81.45
Attention[29]	83.61
Ours	90.62

TABLE VII
PERFORMANCE COMPARISON OF THE FEATURE FUSION MODEL WITH OTHER MODELS ON THE CREMA-D DATASET.

Feature extraction methods	Fusion method	ACC(%)
MFCC	Intermediate feature fusion	71.13
MFCC+MobileNetV2	Intermediate feature fusion	74.24
MFCC+MobileNetV2	Feature-level fusion	75.98
MFCC(FRFT)+MobileNetV2	Feature-level fusion	79.23
MFCC(FRFT)+Conv1D +MobileNetV2	Feature-level fusion	82.12

paper proposes a novel audio-visual emotion recognition framework. The method leverages deep learning to extract high-level features from both facial expressions and speech, and employs feature-level fusion to enhance recognition performance. For visual feature extraction, we improve the MobileNetV2 architecture to efficiently capture facial expression features. On the audio side, we incorporate the FRFT to enhance MFCCs, effectively addressing the challenges posed by non-stationary signals.

A cross-modal residual interaction network is constructed, utilizing Conv1D-based residual blocks to achieve multi-scale deep feature propagation. Furthermore, an attention mechanism is integrated to establish dynamic associative mappings between modalities, enabling the model to adaptively capture salient emotional features. Experiments on the RAVDESS and CREMA-D dataset demonstrate that the proposed framework achieves a significant improvement in seven-class emotion recognition tasks, with accuracy increased by 10.62% and 10.99%, validating the effectiveness of the proposed modules. In future work, we plan to explore the fusion of additional modalities and further optimize the model architecture to develop a more lightweight and efficient emotion recognition system.

REFERENCES

[1] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, pp. 99–117, 2012.

[2] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2937–2987, 2020.

[3] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, 2018.

[4] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[5] D. Xu, Z. Tian, R. Lai, X. Kong, Z. Tan, and W. Shi, "Deep learning based emotion analysis of microblog texts," *Information Fusion*, vol. 64, pp. 1–11, 2020.

[6] J. X. Yu, K. M. Lim, and C. P. Lee, "Move-cnns: Model averaging ensemble of convolutional neural networks for facial expression recognition." *IAENG International Journal of Computer Science*, vol. 48, no. 3, pp. 519–523, 2021.

[7] Y. Yujin, Z. Peihua, and Z. Qun, "Research of speaker recognition based on combination of lpcc and mfcc," in *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*, vol. 3. IEEE, 2010, pp. 765–767.

[8] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal emotion recognition on ravdess dataset using transfer learning," *Sensors*, vol. 21, no. 22, p. 7665, 2021.

[9] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015.

[10] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[11] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 02, 2020, pp. 1359–1367.

[12] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recognition using mfcc," in *International Conference on Computer Graphics, Simulation and Modeling*, vol. 9, 2012, p. 2012.

[13] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Analytics*, vol. 1, pp. 1–22, 2016.

[14] H. Ku and W. Dong, "Face recognition based on mtcnn and convolutional neural network," *Frontiers in Signal Processing*, vol. 4, no. 1, pp. 37–42, 2020.

[15] S.-H. Tsang, "Review: Mobilenetv1–depthwise separable convolution (lightweight model)," *Towards Data Science*, 2018.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[18] E. Sejdić, I. Djurović, and L. Stanković, "Fractional fourier transform as a signal processing tool: An overview of recent developments," *Signal Processing*, vol. 91, no. 6, pp. 1351–1369, 2011.

[19] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4. IEEE, 2005, pp. 3437–3443.

[20] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *ArXiv Preprint ArXiv:1707.07250*, 2017.

[21] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, 2020, pp. 1–6.

[22] Z. Fu, F. Liu, H. Wang, J. Qi, X. Fu, A. Zhou, and Z. Li, "A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition," *ArXiv Preprint ArXiv:2111.02172*, 2021.

[23] Y. Zhou and X. Xie, "Research on speech emotion recognition based on spatiotemporal attention cnn model." *IAENG International Journal of Computer Science*, vol. 52, no. 6, pp. 1852–1860, 2025.

[24] A. Dutt and P. Gader, "Wavelet multiresolution analysis based speech emotion recognition system using 1d cnn lstm networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2043–2054, 2023.

[25] P. Xiang, C. Lin, K. Wu, and O. Bai, "Multimae-der: Multimodal masked autoencoder for dynamic emotion recognition," in *2024 14th International Conference on Pattern Recognition Systems (ICPRS)*. IEEE, 2024, pp. 1–7.

[26] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, "Negative emotion recognition using deep learning for thai language," in *2020 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT & NCON)*. IEEE, 2020, pp. 71–74.

[27] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "Serab: A multi-lingual benchmark for speech emotion recognition," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 7697–7701.

[28] B. Mocanu and R. Tapu, "Speech emotion recognition using ghostvlad and sentiment metric learning," in *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2021, pp. 126–130.

[29] D. Li, Z. Yang, J. Liu, H. Yang, and Z. Wang, "Emotion embedding framework with emotional self-attention mechanism for speaker recognition," *Expert Systems with Applications*, vol. 238, p. 122244, 2024.