Lightweight Multimodal Emotion Recognition for Companion Robots: A Deep Learning Framework Integrating Facial and Speech Features

Cheng-Kai Lu, *Senior Member, IEEE*, Chien-Wei Lu and Guan Bo Lin

***Abstract*— This paper presents a lightweight multimodal deep learning framework for real-time emotion recognition on resource-constrained companion robots, exemplified by Zenbo Junior II. The framework integrates a customized GhostNet with Triplet Attention Modules (TAM) and a Frame Attention Network (FAN) for spatio-temporal facial feature encoding, and employs a depth-optimized one-dimensional convolutional neural network (1D-CNN) for compact speech representation. Decision-level fusion based on the geometric mean enhances robustness to noisy modality predictions. The proposed model comprises 0.92 million parameters and requires 0.77 billion floating-point operations (GFLOPs), achieving 97.56% accuracy on the RAVDESS dataset and 82.33% on CREMA-D. In contrast to existing approaches that optimize accuracy at the expense of computational efficiency, the proposed method demonstrates a balance of accuracy, efficiency, and deployability. These results highlight both the novelty and the feasibility of the framework for real-time emotion monitoring in healthcare and human–robot interaction.**

***Index Terms*— Lightweight, Multimodal Deep Learning, Emotion Recognition, Real-Time, Companion Robots**

I. INTRODUCTION

The rapid aging of the global population poses significant challenges to healthcare systems, particularly in resource-limited rural areas where a shortage of human caregivers has resulted in inadequate support for the elderly [1]. This demographic shift necessitates innovative solutions, such as companion robots, to alleviate the caregiving burden. Beyond aiding the elderly, these robots can provide valuable emotional detection services for individuals who may be experiencing depression or are reluctant to disclose their psychological struggles [2]. When equipped with emotion recognition capabilities, companion robots can monitor emotional well-being, facilitate cognitive activities, and offer essential emotional support. Accurate real-time perception and interpretation of human emotions are critical for effective companionship and intervention [3].

Recent advancements in deep learning have significantly improved emotion recognition from both facial and speech data through the development of complex neural network architectures [4]. Facial expressions serve as vital indicators of emotional states, conveying feelings such as happiness, sadness, anger, and surprise. Similarly, vocal characteristics—such as tone and volume—are crucial for expressing emotions in speech [5]. By analyzing these multimodal cues, companion robots can gain a deeper understanding of the emotional states of individuals they interact with, including those facing mental health challenges.

Despite these advancements, existing deep learning systems often require substantial computational resources, rendering them unsuitable for deployment on devices with limited processing power [6]. This study addresses this challenge by introducing a lightweight multimodal emotion recognition framework specifically designed for real-time use in resource-constrained companion robots. The proposed system employs a customized GhostNet-based architecture for facial emotion recognition alongside a depth-optimized 1D-Convolutional Neural Network (1D-CNN) for speech analysis. This bimodal approach not only achieves high accuracy in emotion recognition but also significantly reduces computational costs, making it suitable for real-time deployment in various contexts, including healthcare and emotional support services. Trained on the Emotional Speech and Song (RAVDESS) [7] and Crowd-sourced Emotional Multimodal Actors (CREMA-D) [8] datasets, the system effectively balances performance and resource efficiency, contributing to the advancement of companion robots in providing emotional detection and support. This study introduces, for the first time, the joint integration of a GhostNet enhanced with Triplet Attention Modules (TAM), a Frame Attention Network (FAN), and a depth-optimized 1D-CNN into a unified multimodal emotion recognition framework explicitly tailored for embedded companion robots. The main contributions of this study are:

1. Hardware-Amenable Visual Emotion Recognition: A customized GhostNet architecture enhanced with TAM

Manuscript received xx September 2024, revised xx xxx 2024, accepted xx 2025. Date of publication xx 2025, date of current version xx. This work was supported by the National Science and Technology Council, Taiwan, under Grant No. NSTC 111-2222-E-003 -001, NSTC 112-2221-E-003-008, and NSTC 113-2221-E-003 -006 -MY2. *(Corresponding author: Cheng-Kai Lu).*

Cheng-Kai Lu is with the Department of Electrical Engineering, National Taiwan Normal University, Taipei, Taiwan (e-mail: cklu@ntnu.edu.tw).

Chien-Wei Lu is with the Department of Electrical Engineering, National Taiwan Normal University, Taipei, Taiwan (e-mail: 61175053h@ntnu.edu.tw).

Guan Bo Lin is with the Department of Electrical Engineering, National Taiwan Normal University, Taipei, Taiwan (e-mail: 61375071h@ntnu.edu.tw).

and FAN to enable effective spatio-temporal feature extraction from facial sequences with minimal computational cost.
2. Compact Audio Emotion Modeling: A lightweight 1D-CNN that processes MFCCs with optimized depth and kernel size to preserve emotional cues under tight resource constraints.
3. Edge Deployability: The model is successfully deployed on the Zenbo Junior II companion robot, demonstrating its feasibility for real-time applications.

II. Related works

Emotion recognition has been extensively studied across various modalities, including facial expressions and speech. These modalities are crucial for developing systems that can accurately interpret human emotions, as demonstrated in the following sections:

A. Facial Emotion Recognition

Facial expressions are a vital channel for conveying emotional states and have become a prominent research area in emotion recognition, particularly within computer vision and human-computer interaction domains. Before emotional features can be extracted, the face must first be detected within the image.

Kansizoglou et al. [9] and Banskota et al. [10] utilized the Haar Feature-based Cascade Detector proposed by P. Viola and M. Jones [11]. This detector uses rectangular features to capture simple structures in images, such as edges, lines, and textures. The cascade classifier is a series of simple classifiers applied sequentially to improve detection performance. Furthermore, He et al. [12], Krishnani et al. [13], and Sun et al. [14] employed Multi-task Cascaded Convolutional Neural Networks (MTCNN), developed by Zhang et al. [15], which efficiently detect facial features such as the eyes and mouth, particularly in images with small faces.

Feature extraction is a critical step in emotion recognition systems, determining the extent of emotionally significant information derivable from facial data. Traditional methods [16] include Geometric Deformation Features, Local Binary Patterns (LBP), Local Gabor Binary Patterns, and Local Phase Quantization. However, with the advent of deep learning techniques, automatic extraction and learning of features directly from images have become dominant.

Deep learning models like Convolutional Neural Network (CNN) are foundational for extracting deep features [17]. For instance, Zhu et al. [18], Wu et al. [19], and Ghaleb et al. [17] employed VGG networks for feature extraction. Hajarolasvadi et al. [20] and Wei et al. [21] used ResNet architectures, while Zou et al. [22] and Lakshminarayana et al. [23] utilized DenseNet networks for emotion recognition tasks.

Despite significant advances in facial emotion recognition technology, challenges remain in improving accuracy and robustness under variable lighting conditions and diverse facial postures. Current systems often struggle with these variations, necessitating consideration of lighting changes and different face angles for improved performance. Moreover, processing complete image sequences to capture dynamic expressions requires considerable computational resources, posing a challenge for deployment on resource-constrained devices like companion robots.

B. Speech Emotion Recognition

The human voice is rich in emotional elements, making speech a valuable modality in emotion recognition systems. These systems analyze various acoustic features such as pitch, rhythm, volume, and timbre to capture the emotional state of the speaker. This section presents key technologies in speech emotion recognition.

Effectively capturing the correlation between acoustic features and emotions is crucial in emotion recognition systems. Various methods are used to express acoustic features, each with specific applications. Common methods [7] include Zero-Crossing Rate (ZCR), Energy, Entropy of Energy, Mel-Frequency Cepstral Coefficients (MFCCs), Chromagram, and Tonnetz. These techniques help capture the relationship between acoustic features and emotions.

In addition, deep learning neural networks are the most popular method for extracting higher-level features in recent years. Sharafi et al. [24] and Wang et al. [25] used CNNs to further extract features.

Zou et al. [26] employed Long Short-Term Memory (LSTM) networks, a specialized form of Recurrent Neural Network (RNN), to identify long-term dependencies in sequential data by introducing gates to regulate information storage, updating, and forgetting. Chen et al. [27] used Gated Recurrent Units (GRUs), another RNN variant designed to address gradient vanishing issues in long sequential data processing. GRUs simplify LSTM structures by incorporating Update and Reset Gates, enabling more effective capture of long-term dependencies in time series.

In a study by Braunschweiler et al. [28], CNNs were combined with bidirectional LSTMs (Bi-LSTM) to extract complex audio features for emotion recognition.

C. State-of-the-art Methods for the Emotion Recognition

In recent years, significant research has focused on emotion recognition, leading to the development of advanced methodologies. Hu et al. [29] proposed a method for speech emotion recognition that integrates visual information. This approach involves extracting emotion features from the log-mel spectrogram using a ResNet-based Attention Convolutional Neural Network (RACNN). Subsequently, facial emotion features are extracted using a CNN and Bidirectional Gated Recurrent Units (Bi-GRUs), and both modalities are fused.

Mocanu et al. [30] introduced a novel multimodal emotion recognition method based on the fusion of audio and visual data. This method incorporates spatial, channel, and temporal attention mechanisms within a three-dimensional convolutional neural network (3D-CNN) for vision and a temporal attention mechanism within a two-dimensional convolutional neural network (2D-CNN) for audio. The features from both

modalities are then integrated using a cross-attention fusion technique.

Sharafi et al. [24] combined visual data input into Deep Temporal Network (DTN), Deep Spatial Network (DSN), and Bidirectional Long Short-Term Memory (Bi-LSTM) networks. Emotional features from facial data were extracted using an LSTM network, while audio data was processed through a proposed 1D-CNN to obtain features from both modalities, which were then integrated through a fully connected layer.

In contrast, Ryumina et al. [31] proposed a feature extraction method for facial expression recognition based on the importance scores of distances between facial landmarks as emotion features, which were then input into a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN).

Most of these advanced techniques employ a bimodal approach to enhance emotion recognition systems. Unimodal inputs are limited in capturing emotion features and are susceptible to environmental disturbances. Bimodal inputs enhance the capture of emotional features by leveraging complementary information across modalities, and they improve robustness against environmental disturbances. Consequently, this study also adopts a bimodal approach for implementing the emotion recognition system.

While these advanced techniques have shown excellent results in emotion recognition—Sharafi et al., for instance, achieved a high accuracy rate—they do not consider deployment on computationally limited devices, potentially hindering their use on resource-limited platforms such as companion robots.

III. Methodology

This section introduces the overall architecture of the proposed bimodal emotion recognition framework, as illustrated in Fig. 1. The system comprises two parallel processing streams: one for facial emotion recognition (described in Section III-A) and the other for speech emotion recognition (described in Section III-B). Both streams are carefully designed to ensure computational efficiency and suitability for deployment on resource-constrained platforms. In addition, the decision-level fusion mechanism (outlined in Section III-C), which integrates outputs from both modalities, is presented to demonstrate how the framework achieves robust and reliable emotion classification.

A. Facial Emotion Recognition Architecture

Facial emotion recognition relies on the distribution of emotional cues across various regions of the face. Temporal dynamics, such as changes in facial expressions over time, are also essential for extracting emotional features. In this study, instead of using a single image frame, multiple frames from the same image sequence are utilized to capture the temporal variability of facial expressions. Consequently, the image sequences in the training set are divided into multiple frames, and each set is used as a single data input. This approach enhances the ability of the system to capture emotional states over time, reflecting the temporal dependencies inherent in human emotions.

1) Face Preprocessing: Face preprocessing employs the Multi-task Cascaded Convolutional Networks (MTCNN) framework [15] to extract 224×224 RGB face images. Each video is divided into 15 equal segments, and one central frame per segment is selected after skipping initial frames (default: zero), ensuring balanced temporal coverage of expression dynamics. Data augmentation with slight rotations (±5°), horizontal flips, and brightness/contrast adjustments is then applied. These augmentations help prevent overfitting by encouraging the model to learn more generalized features, thereby improving its performance across diverse facial expressions [32].

2) Facial Feature Extraction: CNNs, particularly 2D-CNNs, have proven highly effective for image processing tasks like classification and object detection. The 2D convolution operation can be defined as:

$$y_{i,j} = \sum_{m=0}^{K_h-1} \sum_{n=0}^{K_w-1} X_{i+m,j+n} \cdot W_{m,n} + b \tag{1}$$

where $X_{i+m,j+n}$ represents the input image, $W_{m,n}$ denotes the convolution kernel with a size of $K_h \times K_w$, and b is the bias term. However, traditional 2D-CNNs incur substantial computational costs, especially when processing large-scale video data with important temporal features. To address this, we adopt and customize GhostNet for facial emotion recognition. Unlike the standard GhostNet [33] , our model integrates Triplet Attention Modules (TAM) [34] within each Ghost Bottleneck to enhance inter-channel and spatial attention with minimal computational overhead, as illustrated in Fig. 2. GhostNet reduces computational burden by first generating a small number of base feature maps via conventional convolution, followed by linear transformations that generate "ghost" feature maps. This two-stage Ghost Module structure significantly lowers parameter count and FLOPs while preserving representational power. The first stage uses traditional convolution to generate base feature maps, as expressed by

$$y'_{h,w} = X_{h,w} \cdot W + b \tag{2}$$

where $X_{h,w}$ is the input eigenmap, $y'_{h,w}$ is the output base feature eigenmap Y', W represents the weight of the 1x1 convolution kernel, and b denotes the bias term.

In the second stage, each base feature map is transformed into multiple ghost feature maps ($g_{i,j}$) through linear operations:

$$g_{i,j} = \Phi_{i,j}(y'_i), \ \ \forall i = 1, \dots, p, \ \ j = 1, \dots, s \tag{3}$$

where y'_i represents the i-th base feature map in Y', Φ is the linear operation operator, and each of the p base feature maps is converted into s Ghost feature maps in the second stage.

Each Ghost Bottleneck (G-bneck) in our model is composed of two Ghost Modules with a TAM inserted between them,

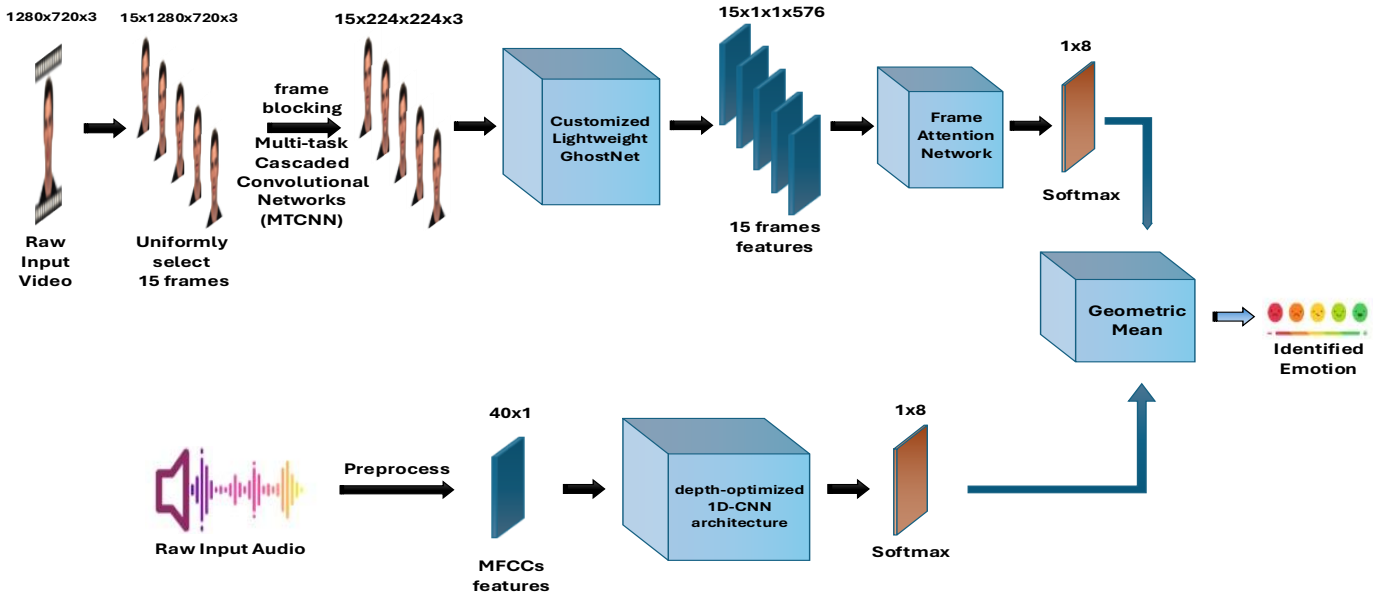

Fig. 1. Overview of the proposed lightweight multimodal emotion recognition framework. The system integrates facial and speech processing streams, with decision-level fusion for final classification. For the facial stream, 15 frames are uniformly sampled per video sequence as described in Section III-A, ensuring balanced temporal coverage of expressions.

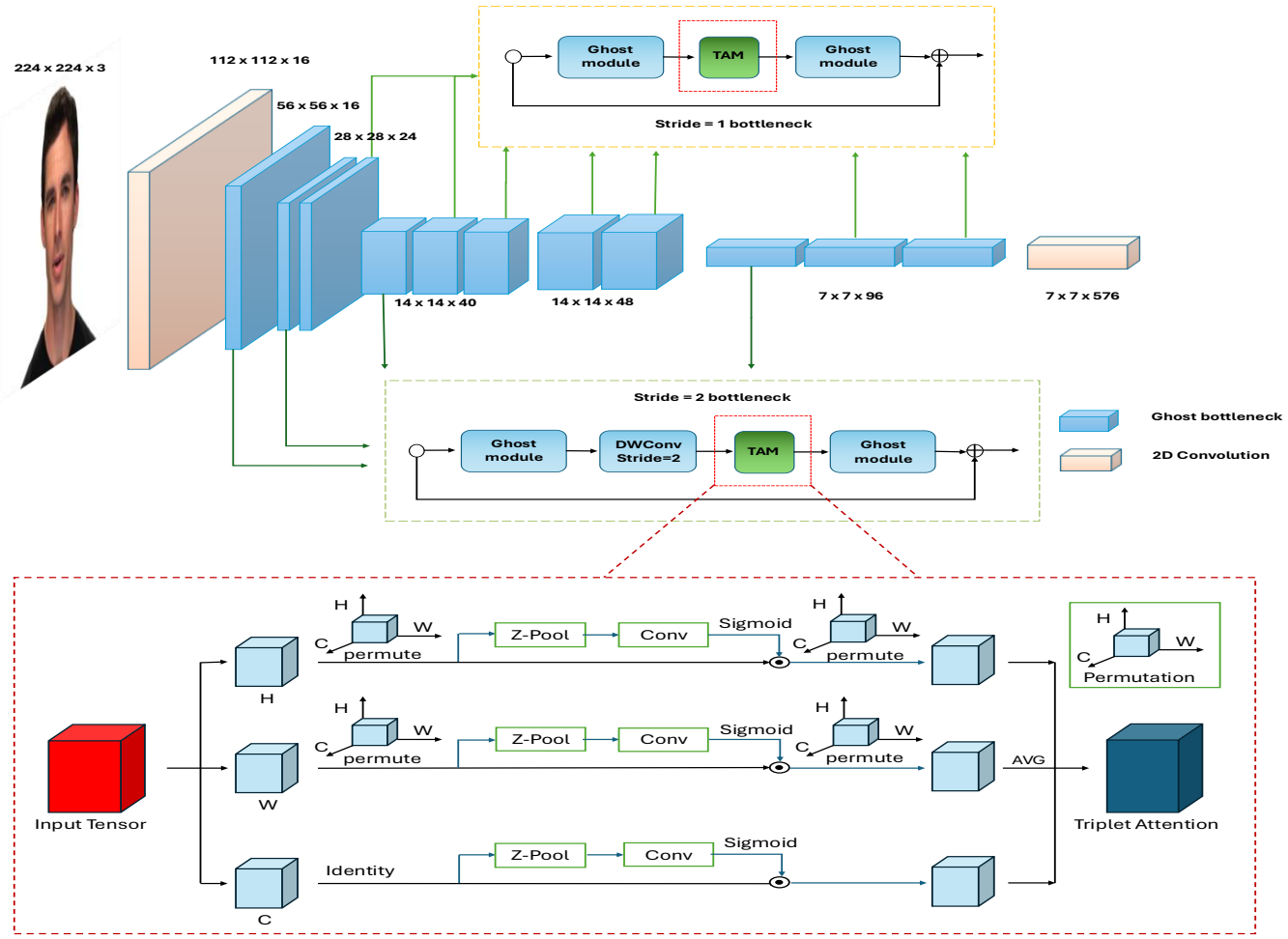

Fig. 2. Customized GhostNet architecture with Triplet Attention Modules (TAM) embedded in Ghost Bottlenecks. The TAM [34] (highlighted in green) comprises three rotated branches (H–C, W–C, H–W) that sequentially model inter-dimensional dependencies. The outputs of these branches are aggregated and fused with the input via a residual connection, delivering lightweight spatial–channel attention with negligible additional FLOPs.

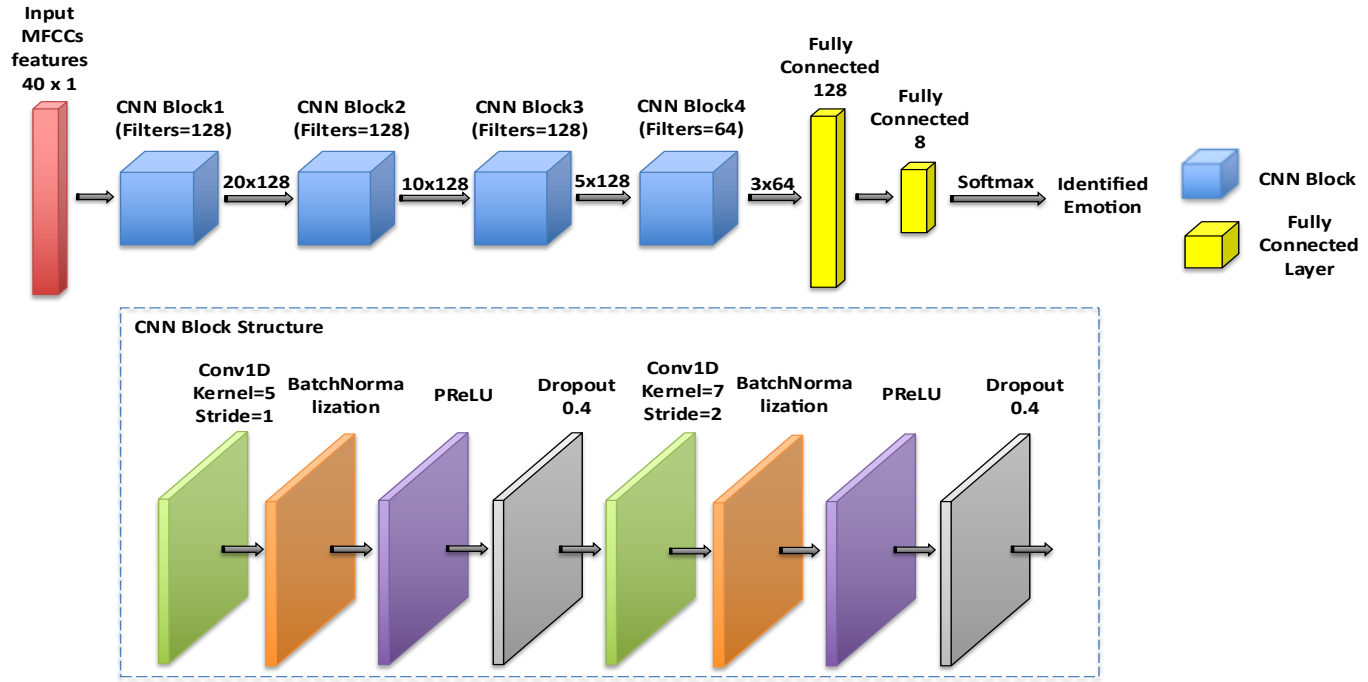

Fig. 3. Depth-optimized 1D-CNN architecture.

TABLE I GHOSTNET ARCHITECTURE INFORMATION

layer	K	Exp	TAM	stride	Output
Conv2D	3×3	-	-	2	112^2×16
G-bneck	3×3	16	1	2	56^2×16
G-bneck	3×3	72	-	2	28^2×24
G-bneck	3×3	88	-	1	28^2×24
G-bneck	3×3	96	1	2	14^2×40
G-bneck	5×5	240	1	1	14^2×40
G-bneck	5×5	240	1	1	14^2×40
G-bneck	5×5	120	1	1	14^2×48
G-bneck	5×5	144	1	1	14^2×48
G-bneck	5×5	288	1	2	7^2×96
G-bneck	3×3	576	1	1	7^2×96
G-bneck	3×3	576	1	1	7^2×96
Conv2D	1×1	-	-	1	7^2×576

Unlike earlier uses of TAM in large-scale CNNs, in this framework the Triplet Attention Module is embedded within each Ghost Bottleneck, enabling lightweight inter-channel and spatial attention without incurring extra FLOPs—a crucial advance for embedded emotion recognition tasks. Let x denote the input tensor with dimensions (H, C, W). The TAM processes x through three branches: the first two branches permute the tensor's dimensions and apply convolution, resulting in attention maps $\hat{x}_1$ (for the height–channel perspective) and $\hat{x}_2$ (for the width–channel perspective); the third branch utilizes Z-pooling and convolution to generate x_3^* (capturing the height–width relationship). The aggregated output of TAM can thus be expressed as:

$$y = \frac{1}{3}\left(\overline{\hat{x}_1\sigma(\psi_1(\hat{x}_1^*))} + \overline{\hat{x}_2\sigma(\psi_2(\hat{x}_2^*))} + x\sigma\left(\psi_3(x_3^*)\right)\right) \quad (4)$$

TABLE II 1D-CNN ARCHITECTURE INFORMATION

layer	Input	K	stride	output
Conv1D	40×1	5	1	40×128
Conv1D	40×128	7	2	20×128
Conv1D	20×128	5	1	20×128
Conv1D	20×128	7	2	10×128
Conv1D	10×128	5	1	10×128
Conv1D	10×128	7	2	5×128
Conv1D	5×128	5	1	5×64
Conv1D	5×64	7	2	3×64
FC	64	-	-	128
FC	128	-	-	8
Softmax	8	-	-	8

where σ is the activation function and ψ_1, ψ_2, ψ_3 are branch-specific convolutions. The overline ($\bar{}$) denotes restoring each branch output to the original (H, C, W) shape. This aggregated result is then fused with the original input x through a residual connection, ensuring that spatial–channel attention is integrated seamlessly and with negligible computational overhead.

Finally, customized G-bneck modules form a lightweight GhostNet architecture (Fig. 2), detailed layer-by-layer in Table I, specifying kernel sizes, expansion factors, TAM usage, stride settings, and output dimensions. This GhostNet efficiently captures essential spatial features from 15-frame facial sequences, balancing performance and computational efficiency. Kernel sizes (3×3 and 5×5) align with MobileNetV3-Small design principles to ensure optimal receptive field coverage under lightweight constraints. The extracted spatial features are subsequently processed by the FAN [35], which emphasizes critical frames through sequential self-attention (frame weighting using sigmoid activation) and relational attention (frame-to-global-context comparison). Together, these mechanisms prioritize the most discriminative emotional cues.

B. Speech Emotion Recognition Architecture

Speech emotion recognition relies on analyzing vocal characteristics such as pitch, tone, and rhythm, which convey emotional states. This section outlines the techniques and architectures used for processing speech data in the bimodal emotion recognition framework.

1) Audio Preprocessing: To enhance model generalizability and ensure robustness, data augmentation techniques are applied to the speech signals. These techniques include the addition of noise, alterations in pitch, and modifications to the speech rate. This expands the dataset and allows the model to learn more generalized representations of speech features, which is crucial for improved performance on unseen data.

Given the hardware constraints of the companion robot, Mel-Frequency Cepstral Coefficients (MFCCs) are selected for feature extraction. MFCCs are known for capturing the nonlinear perceptual characteristics of human hearing. The general frequency f is mapped to the Mel scale using the formula:

$$mel(f) = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \tag{5}$$

where f represents the frequency in Hertz, and $mel(f)$ is the Mel frequency corresponds to the frequency on the Mel scale. This transformation ensures that the extracted features align with human auditory perception, emphasizing key frequency components.

Conventionally, MFCCs are represented as a 2D matrix (n_{mfcc}, T), where n_{mfcc} is the number of coefficients and T is the number of time frames. In this study, the time axis T is averaged, yielding a feature vector of dimension $n_{\text{mfcc}} = 40$. The vector is then reshaped to (40, 1), where "1" denotes the channel dimension, allowing efficient input to the subsequent 1D-CNN. This transformation reduces computational complexity while preserving the spectral characteristics of MFCCs, enabling effective feature learning under resource constraints.

2) Feature Extraction from Speech

To maintain computational efficiency, a 1D-CNN is employed to process MFCC features. Let $x = [x_1, x_2, \ldots, x_T]$ denote the input sequence of length T. A convolution kernel with weights $W = [W_0, W_1, \ldots, W_{L-1}]$ of length L and bias b produces the output:

$$Y_i = \sum_{m=0}^{L-1} X_{i+m} \cdot W_m + b \tag{6}$$

where Y_i is the output feature at position i, X_{i+m} is the input element, W_m the kernel weight, and b is the bias term. This formulation explicitly describes how the kernel slides across the sequence to extract localized temporal dependencies.

The optimized 1D-CNN employs a depth-tapered structure to balance efficiency and accuracy, with eight convolutional blocks that progressively narrow channel width while preserving temporal information from 40-dimensional MFCCs. Kernel sizes of 5 and 7 capture critical temporal patterns, providing effective resolution with minimal computational overhead. Each block consists of convolution, Batch Normalization (for stabilized training), Parametric ReLU (for improved gradient flow), and Dropout (for regularization). This design achieves accurate emotional feature extraction with low FLOPs, supporting real-time deployment on resource-constrained platforms. Architecture details are summarized in Table II and visualized in Fig. 3. After feature extraction, the learned representations are passed through a fully connected (FC) layer and a final Softmax classifier, yielding an 8-dimensional probability distribution corresponding to the target emotion classes. The effectiveness of this compact design was validated through ablation studies (see Tables Ⅶ and Ⅷ), which confirm its ability to retain emotion-relevant features while significantly reducing model complexity.

C. Decision-Level Fusion Strategy

Combining features from different modalities (e.g., facial and speech) is challenging due to variations in feature scales and distributions, where improper integration may cause one modality to dominate and degrade overall performance. To address this, decision-level fusion is employed: the facial and speech streams are processed by separate networks, each producing an 8-dimensional softmax probability vector, Y_{F_i} and Y_{S_i}, where $i \in \{0, 1, \ldots, 7\}$ corresponds to the eight emotion classes in RAVDESS (neutral, calm, happy, sad, angry, fearful, disgust, and surprised). The final prediction is then obtained by combining these unimodal outputs using probability-level fusion. Two probability fusion strategies are examined:

1. Arithmetic Mean (AM):

$$Y_{A_i} = \frac{Y_{F_i} + Y_{S_i}}{2} \quad , \quad i = 0, 1, \ldots, 7 \tag{7}$$

This provides a simple average of probabilities and serves as a computationally efficient baseline.

2. Geometric Mean (GM):

$$Y_{G_i} = \sqrt{Y_{F_i} \times Y_{S_i}} \quad , \quad i = 0, 1, \ldots, 7 \tag{8}$$

This penalizes inconsistent predictions and amplifies cross-modal agreement, thereby improving robustness to noisy or imbalanced unimodal outputs.

Although AM and GM are mathematically simple, they were chosen for their negligible computational overhead, essential for real-time embedded deployment. GM is adopted in this study for its robustness in suppressing unreliable modality-specific predictions and reinforcing consistent cross-modal signals, thereby improving recognition of subtle or easily misclassified emotions.

D. Computing Cost Indicators for Neural Networks

In developing neural network models, computational cost is a crucial metric for evaluating both performance and resource

consumption. This section discusses the metrics used to measure the computational efficiency of the proposed neural network model and its suitability for deployment on resource-constrained devices, such as the Zenbo Junior II companion robot.

1) Parameters: The total number of trainable parameters in a neural network (i.e., weights and biases) directly affects its memory footprint and computational cost. A key objective of this study is to reduce the number of parameters while maintaining high accuracy.

In the equations of this section, C_{in} and C_{out} denote the number of input and output channels, respectively. In the facial emotion recognition system, the Ghost Module is employed to reduce the number of parameters. It achieves this by dividing the conventional 2D convolution operation into two stages. First, a smaller set of base feature maps is generated, followed by the creation of additional "ghost" feature maps through linear transformations. This strategy effectively reduces the parameter count compared to traditional 2D convolutions. The reduction in parameters is clearly demonstrated by comparing the results of equations (9) and (10), as detailed in equation (11). For a traditional convolution:

$$P_c = k^2 \times C_{in} \times C_{out} + C_{out} \tag{9}$$

where P_c signifies the number of parameters for traditional convolution, the first term accounts for convolutional weights, and the second term corresponds to the bias parameters, with one bias per output channel.

In the Ghost Module, parameter efficiency is enhanced through a two-stage process. In the first stage, a small set of C_1 base feature maps is produced via standard convolution. In the second stage, each base feature map undergoes several cheap linear transformations to generate the remaining "ghost" feature maps, as reflected in the total parameter count for Ghost Module （P_g）in equation (10):

$$P_g = (k^2 \times C_{in} \times C_1 + C_1) + (k^2 \times C_1 \times C_{out}) \tag{10}$$

Here, C_1 denotes the number of base features. To compare the efficiency of the Ghost Module versus traditional convolution, we can form their parameter ratio. If $(C_1 \ll C_{out})$ the dominant terms in both P_c and P_g give a ratio:

$$\frac{P_c}{P_g} = \frac{k^2 \times C_{in} \times C_{out} + C_{out}}{(k^2 \times C_{in} \times C_1 + C_1) + (k^2 \times C_1 \times C_{out})} \approx \frac{k^2 \times C_{in} \times C_{out}}{k^2 \times C_{in} \times C_1} = \frac{C_{out}}{C_1} \tag{11}$$

This shows the Ghost Module dramatically reduces parameter count, as summarized in equation (11).

For the 1D-CNN in the speech stream, further parameter savings arise because the convolutional kernel operates along only a single temporal dimension. Its parameter count, from equation (12), is:

$$P_s = k \times C_{in} \times C_{out} + C_{out} \tag{12}$$

Here, the smaller kernel size k further reduces parameter overhead, making 1D-CNNs especially suited to efficient speech modeling.

2) FLOPs (Floating-Point Operations): Floating-point operations (FLOPs) are an important measure of the computational complexity of a neural network. FLOPs quantify the arithmetic operations (multiplications and additions) required to perform a forward pass through the network. Minimizing FLOPs is essential for ensuring that the network can run efficiently on devices with limited computational power, such as the Zenbo Junior II.

For traditional 2D convolution, the FLOPs can be calculated as:

$$F_c = k^2 \times C_{in} \times H_{out} \times W_{out} \times C_{out} \tag{13}$$

where H_{out} and W_{out} denote the output feature map height and width, respectively. This indicates that FLOPs increase rapidly as the spatial dimensions or kernel size k become larger.

For the Ghost Module, FLOPs are split into two terms: one for generating base maps and another for generating ghost maps:

$$F_g = (k^2 \times C_{in} \times H_1 \times W_1 \times C_1) + (k^2 \times C_1 \times H_{out} \times W_{out} \times C_{out}) \tag{14}$$

The first part of the equation (14) reflects reduced base feature map generation (H_1, W_1 smaller than $H_{\text{out}}, W_{\text{out}}$) and the second term accounts for lightweight ghost transformations. Since C_1 is smaller than C_{out} the overall FLOPs are significantly reduced.

Similar to parameter reduction, the ratio of FLOPs between traditional convolution and GhostNet can also be approximated as $\frac{C_{out}}{C_1}$, under the assumption that C_1 is much smaller than C_{out}, as shown in equation (15).

$$\frac{F_c}{F_g} = \frac{k^2 \times C_{in} \times H_{out} \times W_{out} \times C_{out}}{(k^2 \times C_{in} \times H_1 \times W_1 \times C_1) + (k^2 \times C_1 \times W_{out} \times H_{out} \times C_{out})} \approx \frac{C_{out}}{C_1} \tag{15}$$

For the 1D-CNN in the speech stream, the FLOPs are expressed as:

$$F_s = k \times C_{in} \times L_{out} \times C_{out} \tag{16}$$

where L_{out} represents the length of the output feature map. Because convolution occurs only along the temporal dimension, FLOPs are significantly reduced, making 1D-CNN an ideal choice for MFCC-based speech processing under resource constraints.

IV. Experimental results

This section presents the experimental evaluation of the proposed bimodal emotion recognition framework across facial and speech modalities. The model is tested on two benchmark datasets—RAVDESS and CREMA-D. In addition to performance assessment, ablation studies examine the contributions of key architectural components. Comparative analysis with state-of-the-art (SOTA) methods is also included to validate the model's effectiveness and computational efficiency.

A. Benchmark datasets

To assess performance and generalizability, the RAVDESS and CREMA-D datasets—both offering high-quality, synchronized audio-visual data—were utilized.

RAVDESS comprises 2,452 clips (1,440 speech and 1,012 song) from 24 actors (12 male, 12 female), covering eight emotions: neutral, calm, happiness, sadness, anger, fear, disgust, and surprise—most presented at two intensity levels. The recordings (1280×720 resolution, 48 kHz audio) were augmented with Gaussian noise ($\sigma^2 = 0.01$) and subjected to 0.8× time-stretching to enhance robustness. Its structured and balanced design makes it well-suited for evaluation in controlled environments.

CREMA-D includes 7,437 utterances from 91 actors (48 male, 43 female), each expressing one of six emotions—anger, disgust, fear, happiness, neutral, or sadness—at a single intensity. Captured under diverse conditions and providing approximately 1,270 samples per class, it captures real-world expressive variability, making it suitable for testing model generalization.

B. Implementation Details

Model training utilized an Intel i7-12700H CPU, RTX 3090 Ti GPU, and 64 GB DDR5 RAM, while deployment testing employed Zenbo Junior II with ARM Cortex-A72/A53 processors, Mali-T860 GPU, and 4 GB RAM. Hyperparameters, including learning rate ([1e-4 to 1e-2]) and dropout (0.1–0.5), were tuned via grid search. Kernel configurations followed MobileNetV3 and GhostNet paradigms to balance emotional feature extraction and efficiency. Performance was validated using 5-fold stratified cross-validation with preserved class balance. Evaluation metrics included accuracy, precision, recall, and F1-score; complexity was assessed via parameter count, FLOPs, arithmetic intensity, and inference time.

C. Emotion classification with RAVDESS and CREMA-D

Performance was evaluated using class-wise precision, recall, and F1-score, computed via a one-vs-rest strategy within a multi-class classification framework—treating each emotion class as positive against all others. On the RAVDESS dataset, the proposed model achieved an average F1-score of 97.56%, with most classes demonstrating near-perfect recognition (see Tables Ⅲ).

TABLE III
PERFORMANCE OF THE PROPOSED METHOD ON THE RAVDESS DATASET

Class	Precision (%)	Recall (%)	F1-score (%)
Neutral	97.30	100.00	98.63
Calm	98.81	97.65	98.22
Happy	98.80	98.80	98.80
Sad	96.30	95.12	95.71
Angry	97.14	100.00	98.55
Fearful	95.52	96.97	96.24
Disgust	100.00	93.75	96.77
Surprised	97.44	97.44	97.44
Average	97.66	97.47	97.56

Neutral and angry reached 100% recall, while happy, calm, and sad exceeded 95% F1-scores. Even subtle emotions such as fearful, disgust, and surprise maintained scores above 96%, confirming the model's strong discriminative capability. These results underscore the effectiveness of the proposed architecture—particularly the integration of customized GhostNet with TAM, FAN, and a depth-optimized 1D-CNN—in capturing both spatial and temporal emotional cues.

TABLE Ⅳ PERFORMANCE OF THE PROPOSED METHOD ON THE CREMA-D DATASET

Class	Precision (%)	Recall (%)	F1-score (%)
Angry	88.75	84.52	86.59
Disgust	89.80	89.07	89.43
Fear	73.30	65.06	68.94
Happy	94.96	93.16	94.05
Neutral	80.77	85.14	82.89
Sad	67.59	76.86	71.93
Average	82.53	82.30	82.33

On the CREMA-D dataset, the model achieved an average F1-score of 82.33%, reflecting increased data variability and emotional ambiguity, as shown in Table Ⅳ. Emotions with clear expressive characteristics, such as happy and disgust, yielded the highest F1-scores (94.05% and 89.43%, respectively), while fear and sad were more prone to misclassification, with F1-scores of 68.94% and 71.93%. The observed performance degradation—particularly in subtle negative emotions—suggests sensitivity to inter-class overlap and recording inconsistencies. Nevertheless, the model demonstrated competitive generalization across domains, confirming its potential for real-time emotion recognition in both structured and unconstrained environments.

V. DISCUSSION

A. Comparison with State-of-the-Art Methods

Tables Ⅴ and Ⅵ compare the proposed bimodal framework with recent SOTA methods on the RAVDESS and CREMA-D datasets, respectively. Although Mocanu et al. [30] and Ryumina et al. [31] report the highest accuracies on CREMA-D (84.57 %) and RAVDESS (98.90 %), their models are too computationally intensive for edge deployment.

In contrast, the proposed framework—with only 0.92 M parameters and 0.77 GFLOPs—achieves a single-sample inference time of 58 ms. For comparison:

- Sharafi et al. [24] require 778 ms (13× slower),
- Mocanu et al. [30] need 6.2 seconds (107× slower),
- Hu et al. [29] require 12.5 seconds (215× slower).

These results highlight the proposed architecture's superior accuracy-to-efficiency balance, enabling real-time emotion recognition on resource-constrained hardware.

TABLE V COMPARISON WITH STATE-OF-THE-ART METHODS ON RAVDESS DATASET

Authors	Accuracy (%)	Params (M)	FLOPs (G)	Inference Time (ms)
Hu et al. [29]	87.92	5.61	390.73	12500
Mocanu et al. [30]	89.25	64.11	300.90	6200
Sharafi et al. [24]	94.99	50.80	12.31	778
Ryumina et al. [31]	**98.90**	63.17	-	-
Proposal	97.56	**0.92**	**0.77**	**58**

TABLE VI COMPARISON WITH STATE-OF-THE-ART METHODS ON CREMA-D DATASET

Authors	Accuracy (%)	Params (M)	FLOPs (G)	Inference Time (ms)
Ryumina et al. [31]	79.10	63.17	-	-
Mocanu et al. [30]	**84.57**	64.11	300.90	6200
Proposal	82.33	**0.92**	**0.77**	**58**

To better understand the model's decision patterns, we analyzed the Integrated Gradient (IG) attribution maps across both facial and acoustic modalities for each emotion, as visualized in Figs. 4-7. In the RAVDESS facial modality (Fig. 4), sadness was often confused with fear, disgust, or surprise, which can be explained by overlapping attribution regions—particularly the mouth corners (indicative of downward motion) and brow center (tightening). These areas also showed high IG responses in fear, disgust, and surprise. When the expressive intensity of sadness was weak (e.g., less visible brow tension), the features became ambiguous, increasing misclassification risk. Similarly, fear was frequently misidentified as disgust or surprise, as all three emotions activated similar facial regions around the mouth and eyes, especially under ambiguous conditions. These overlaps directly contribute to the model's comparatively lower F1-scores for disgust and surprise, as their boundaries with fear and sadness are less distinct in low-expression samples.

In the RAVDESS speech modality (Fig. 5), emotions like neutral, sadness, and disgust demonstrated similar IG distributions, characterized by low attribution values and weak contributions from higher-frequency bands. All three relied slightly more on low-frequency MFCC regions, but without dominant distinguishing features. This made them especially challenging to separate. Moreover, neutral and fear were frequently classified as sadness due to their comparable flat, low-energy spectral profiles. In contrast, anger and surprise showed stronger reliance on low-frequency bands, which also overlapped with disgust, explaining their tendency to be misclassified as such in the unimodal speech stream. For the CREMA-D facial modality (Fig. 6), the model struggled most with fear and sadness, which were often confused with each other. The IG maps show that fear had globally weak or negative attributions (predominantly blue), indicating a lack of salient facial features. Similarly, sadness lacked distinct red hotspots, further reducing class separability. Neutral was also difficult to classify and was frequently mistaken for anger or sadness. This can be attributed to its broad, evenly distributed facial reliance, which may lead the model to misclassify samples when local regions (such as the eyes or cheeks) show stronger activations matching patterns of anger or sadness.

In the CREMA-D speech modality (Fig. 7), only anger and sadness displayed consistently higher IG attributions, particularly in feature 0 (overall signal energy) and low-frequency MFCCs, which correlates with their relatively higher F1-scores. Other emotions lacked dominant IG regions and showed fragmented or low-magnitude activations, contributing to the notably poorer performance in the speech-only setting.

Overall, these attribution patterns confirm that emotions with strong localized facial features and distinct low-frequency acoustic cues (e.g., anger, happy) are more reliably classified, while subtly expressed or overlapping emotions (e.g., sadness, fear, neutral, disgust) present consistent challenges across both modalities.

B. Ablation Studies

Ablation studies were conducted to assess the contribution of different components in the proposed bimodal model. Table VII and Table VIII summarize the results on the RAVDESS and CREMA-D datasets.

TABLE VII ABLATION STUDIES ON THE RAVDESS DATASET

	S1		S2		S3	
	GM	**AM**	**GM**	**AM**	**GM**	**AM**
F1	95.72	92.67	95.93	92.46	94.91	91.45
F2	94.50	93.89	94.09	93.28	94.91	94.30
F3	95.32	94.91	95.52	93.89	95.52	94.09
F4	97.35	96.13	**97.56**	96.13	96.74	95.52

F1 = GhostNet, F2 = GhostNet + TAM, F3 = GhostNet + FAN, F4 = GhostNet + TAM + FAN, S1 = 1D-CNN which the output channels of last four layers are 64, S2 = 1D-CNN which the output channels of last two layers are 64, S3 = 1D-CNN which the output channels are 128, GM = Geometric Mean, AM = Arithmetic Mean.

TABLE VIII ABLATION STUDIES ON THE CREMA-D DATASET

	S1		S2		S3	
	GM	**AM**	**GM**	**AM**	**GM**	**AM**
F1	81.25	76.95	81.38	77.62	80.31	76.55
F2	80.44	77.08	80.38	77.96	80.65	77.82
F3	80.98	77.02	80.85	78.16	80.51	77.22
F4	81.59	78.97	**82.33**	79.44	81.45	79.10

F1 = GhostNet, F2 = GhostNet + TAM, F3 = GhostNet + FAN, F4 = GhostNet + TAM + FAN, S1 = 1D-CNN which the output channels of last four layers are 64, S2 = 1D-CNN which the output channels of last two layers are 64, S3 = 1D-CNN which the output channels are 128, GM = Geometric Mean, AM = Arithmetic Mean.

The highest performance was achieved by combining customized GhostNet with TAM and FAN for facial emotion recognition, and a depth-optimized 1D-CNN with 64 output channels for speech recognition. The GM fusion strategy consistently outperformed AM, providing greater accuracy and robustness. The studies underscore the significance of TAM and FAN, which improved accuracy by nearly 2% on the RAVDESS dataset, refining spatial and temporal feature extraction. The geometric mean fusion also outperformed arithmetic mean fusion, effectively balancing both modalities' contributions. These improvements are crucial for achieving high accuracy and efficiency in real-time applications.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3621086

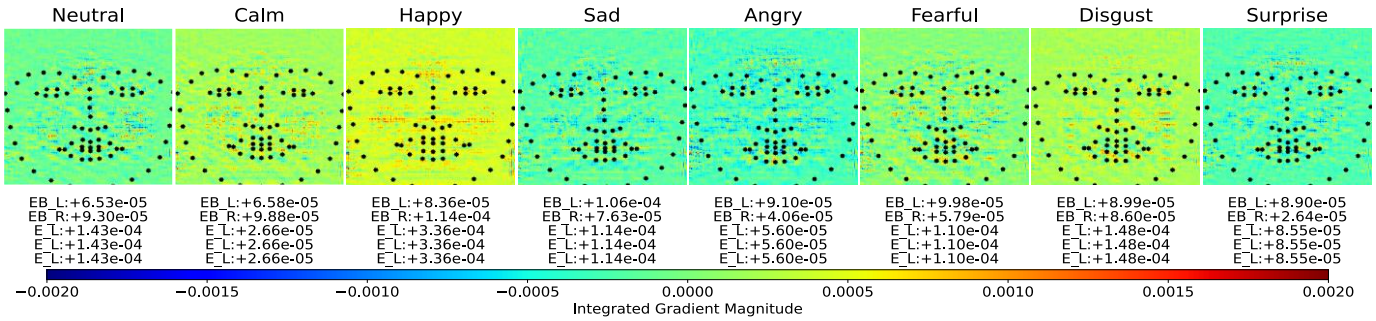

Fig. 4. Integrated-Gradients saliency maps for eight RAVDESS emotions. Warm hues (yellow–red) mark positive, cool hues (cyan–blue) negative contributions; dots indicate 68 landmarks. Numeric summaries (×10-4) for EB L/R and E L/R reveal concentrated brow-eye activity in high-arousal, high-valence states (Happy, Angry) and diffuse patterns in low-arousal emotions.

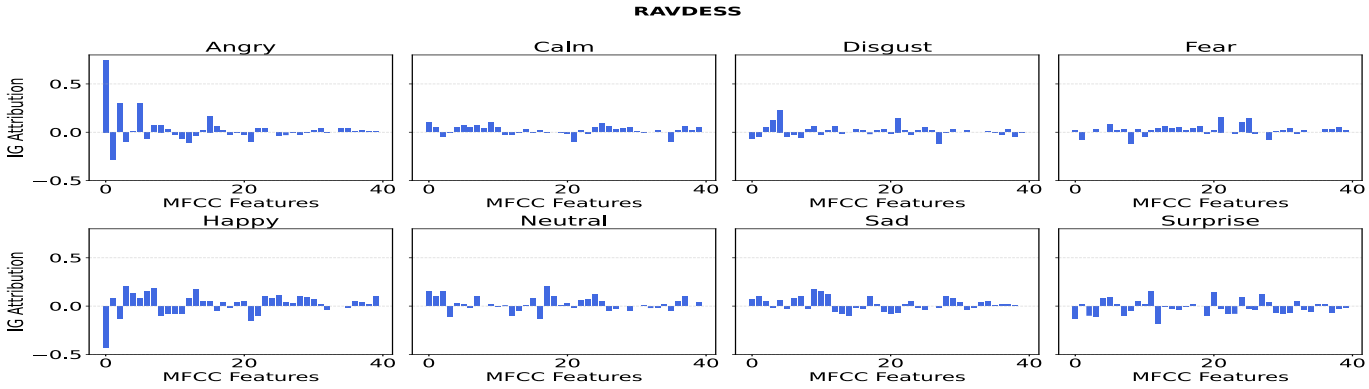

Fig. 5. Integrated-Gradients for 40-D MFCCs across eight RAVDESS emotions: high-arousal classes (Angry, Happy) focus on coefficients 1–5, whereas low-arousal states display flatter, diffuse patterns.

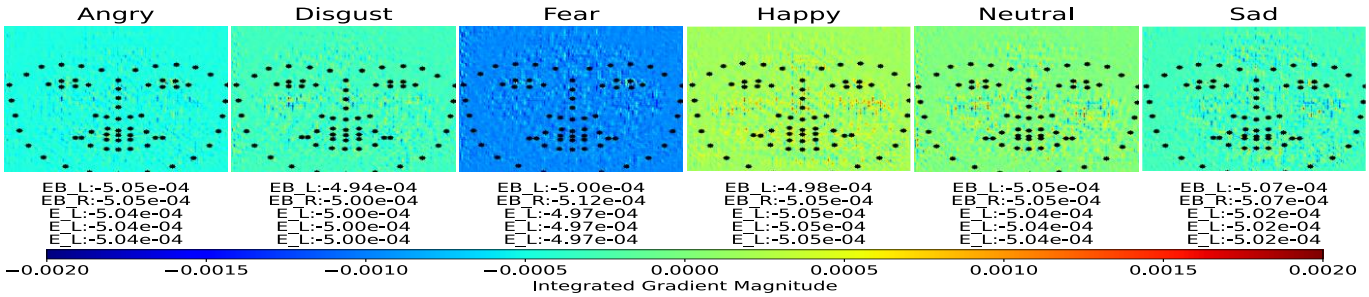

Fig. 6. Integrated-Gradients saliency maps of the facial-landmark branch for the six CREMA-D emotions. Warmer tones (yellow–red) indicate positive, cooler tones (cyan–blue) negative contributions; black dots locate the 68 landmarks. Aggregated scores (×10-4) for the left/right eyebrows (EB L/R) and eyes (E L/R) quantify region-specific importance.

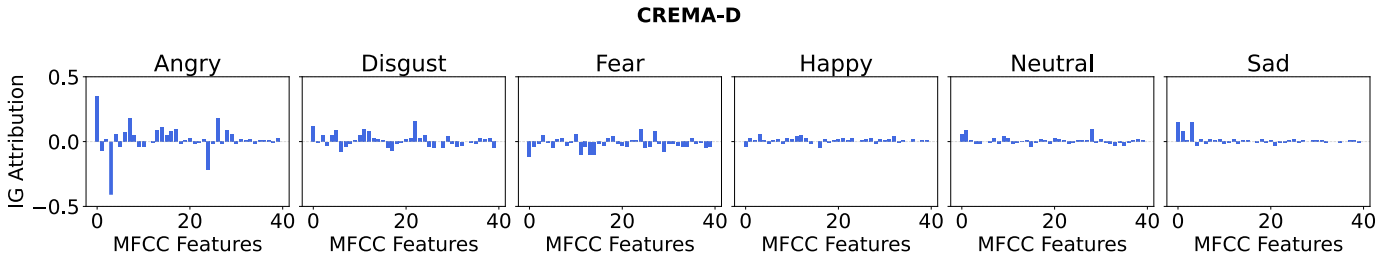

Fig. 7. Integrated-Gradients attribution profiles for the 40-coefficient MFCC vector across six CREMA-D emotions. Positive and negative bars denote supportive and suppressive contributions, respectively. High-arousal classes (e.g., Angry, Disgust) concentrate on the lowest-order coefficients (1 – 5), whereas lower-arousal states rely on broader mid-band features.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3621086

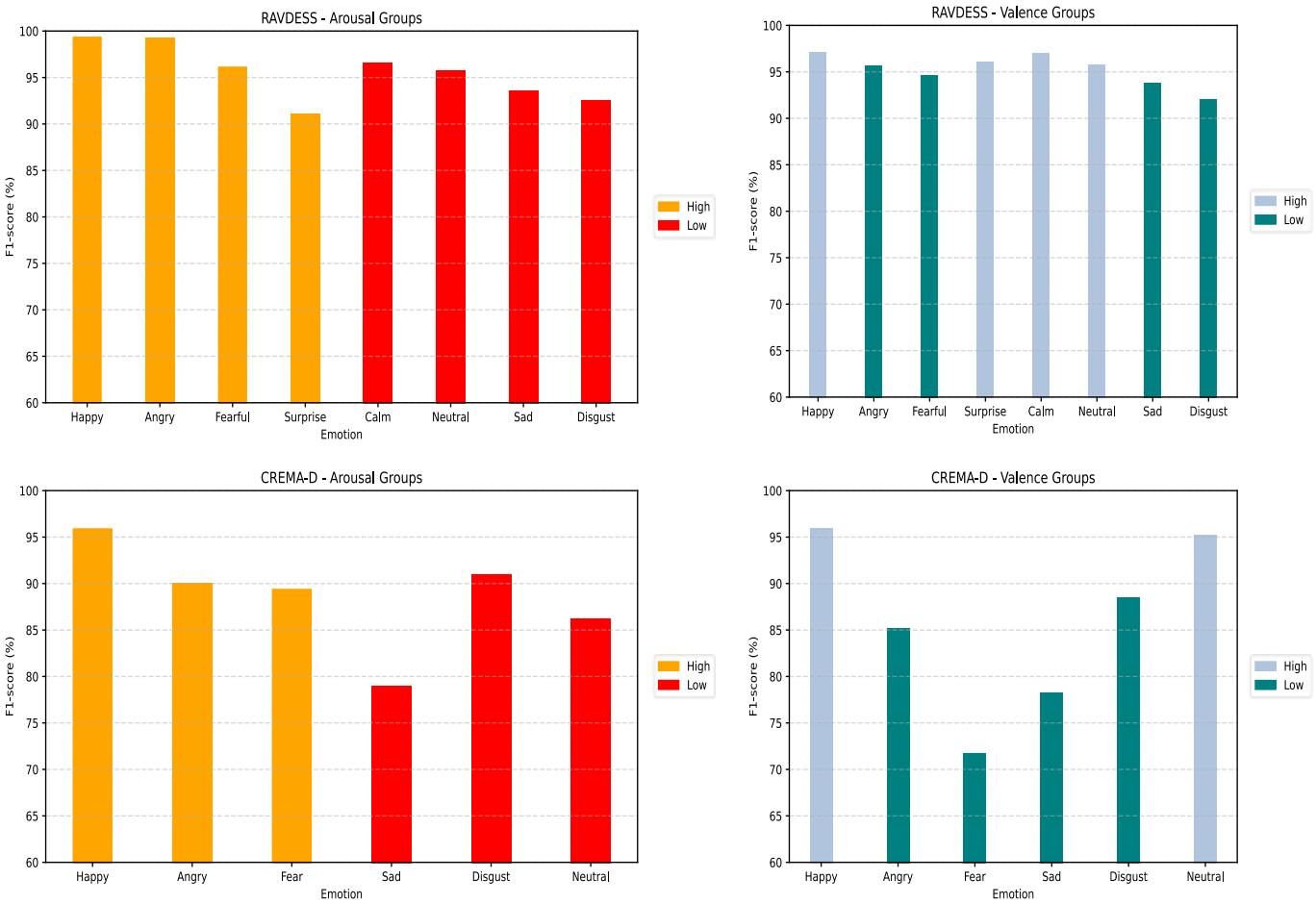

Fig. 8. Comprehensive Affective Grouping Analysis: Arousal and Valence (RAVDESS & CREMA-D).

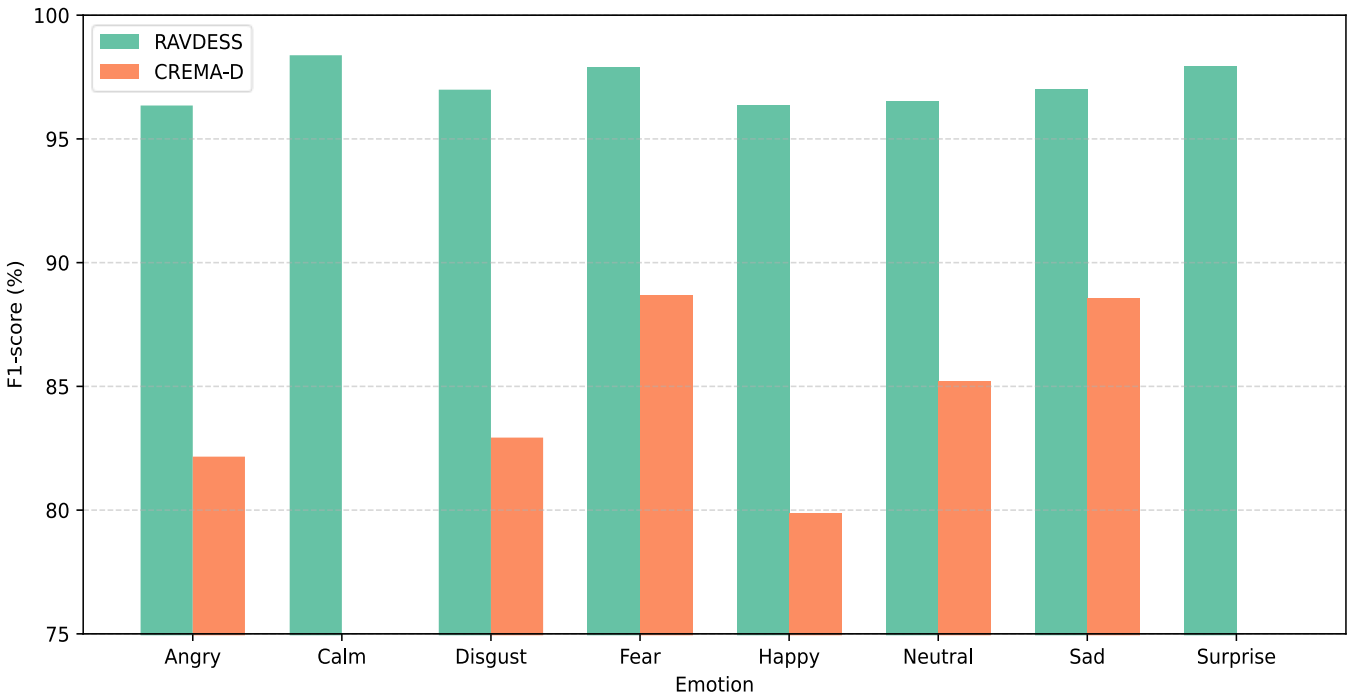

Fig. 9. Leave-One-Emotion-Out (LOEO) Performance Impact- RAVDESS vs. CREMA-D.

TABLE IX TARGETED EMBEDDED DEVICE WITH THEORETICAL COMPUTABILITY FOR OPTIMAL IMPLEMENTATION

Embedded computing board	Memory type	Memory bandwidth (GB/sec)	FP32 performance (GFLOPs/sec)	Arithmetic intensity (FLOPs/byte)	Implementation possibility			
					[29]	[30]	[24]	Proposed
RPi 4B	32-bit LPDDR4	12.80	30.00	2.34	✕	✕	✓	✓
RPi 5	32-bit LPDDR4X	17.10	120.00	7.01	✕	✕	✓	✓
Zenbo Junior II	-	25.60	25.76	1.01	✕	✕	/	✓

The tick (✓) indicates fulfilled criteria on the arithmetic intensity ratio of embedded computing boards (ECB) and proposed architecture is larger than 1 for optimal implementation. The cross (✕) indicates the GFLOPs of architecture exceeded the capped GFLOPs of ECB. The forward slash ('/') indicates possible implementation with more clock cycle due to arithmetic intensity overhead.

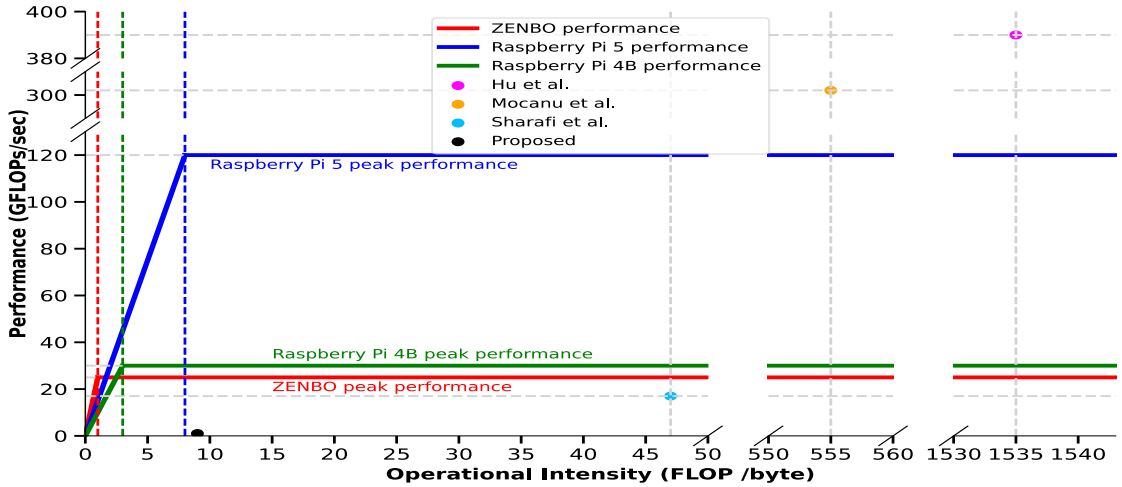

Fig. 10. Roof-line model of Raspberry Pi and Zenbo Junior II.

In addition, to holistically evaluate emotional model robustness, we analyzed classification performance from both affective and structural perspectives. Fig. 8 presents the F1-scores of emotions grouped by arousal and valence across RAVDESS and CREMA-D datasets. Results confirm that high-arousal and high-valence emotions such as Happy and Angry are consistently well-recognized, driven by their distinct and intense multimodal cues. In contrast, low-valence and low-arousal emotions—particularly Fear, Sad, and Disgust—exhibited reduced accuracy, especially under unconstrained conditions in CREMA-D, reflecting expressive ambiguity and inter-class overlap.

Further interpretation of the model's confusion patterns reveals that Disgust and Surprise often overlap with Fear in facial activation zones (e.g., eyes and mouth), particularly under low-expression conditions. This makes them susceptible to misclassification, a trend confirmed through both attribution maps and affective grouping. Similarly, Neutral and Sad expressions lack salient facial or acoustic cues, falling into the low-arousal/low-valence spectrum and resulting in less separable representations. IG attribution analysis shows dispersed or low-energy activations in these cases, while Leave-One-Emotion-Out (LOEO) results demonstrate that their exclusion causes minimal accuracy degradation—suggesting redundancy or overlap with more expressive classes.

Fig. 9 further quantifies individual emotion contributions using a LOEO ablation strategy. Omitting Happy or Angry from training significantly degraded performance—by over 2% in RAVDESS and up to 5% in CREMA-D—demonstrating their foundational role in affective representation learning. In contrast, excluding Calm, Neutral, or Disgust caused limited disruption, reaffirming that not all emotions carry equal structural weight in training.

These findings underscore two key insights:

(1) Expressiveness and arousal intensity are critical for robust multimodal affective modeling, and

(2) Certain emotions act as structural anchors—their presence facilitates generalization across varied emotional spectra and recording conditions.

C. Emotion Recognition System on Zenbo Junior II

The Android Studio development platform was utilized to develop an application for the Zenbo Junior II, into which the trained bimodal emotion recognition neural network model was successfully deployed. This deployment allows real-time emotion recognition on a resource-constrained companion robot. Table IX and Fig. 10 present a comprehensive analysis of deployment feasibility for emotion recognition models on embedded computing boards (ECBs), specifically Raspberry Pi 4B, Raspberry Pi 5, and Zenbo Junior II. This evaluation benchmarks each board's peak FP32 performance (GFLOPs/sec), memory bandwidth (GB/sec), and arithmetic intensity (FLOPs/byte) against the computational demands of the proposed model and selected SOTA methods ([24], [29], [30]).

In Table IX, the Implementation Possibility column categorizes feasibility using the following symbols:

✓: both FLOPs and arithmetic intensity fall within device limits (optimal deployment),

✕: model exceeds hardware constraints (non-deployable),

/: borderline feasibility—model is deployable but may incur latency due to memory bottlenecks.

The proposed model is marked ✓ across all ECBs, reflecting its low computational complexity (0.92M parameters, 0.77 GFLOPs) and favorable arithmetic profile, making it optimally suited for real-time embedded deployment.

Fig. 10 visualizes this relationship through a roofline performance model, where each board's peak compute capacity and memory bandwidth form the theoretical limits. Model positions in this space indicate whether execution is compute-bound or memory-bound. The proposed model consistently lies near the roofline ridge across all devices, confirming that it remains within deployable bounds while maintaining a practical trade-off between efficiency and performance.

This analysis substantiates the framework's real-world deployability, particularly on constrained platforms like Zenbo Junior II, affirming its utility for low-power, real-time affective computing applications.

VI. Conclusion

This study presents an efficient multimodal emotion recognition system designed for resource-constrained devices, such as companion robots. It integrates a GhostNet-based facial emotion recognition architecture—enhanced by the Triplet Attention Module (TAM) and Frame Attention Network (FAN)—with a 1D-CNN for speech analysis. This synergistic design achieves state-of-the-art accuracy with minimal computational overhead. TAM and FAN significantly improve facial emotion recognition by capturing spatial dynamics and emphasizing critical frames. With only 0.92 million parameters and 0.77 billion FLOPs, the system is well-suited for real-time applications on embedded platforms such as Zenbo Junior II, particularly in healthcare environments requiring continuous emotion monitoring. Future directions include extending the system's adaptability to a broader range of emotion categories using generative AI techniques, and enhancing speech-based emotion recognition through advanced signal processing or data augmentation, particularly to address challenges in recognizing subtle emotions such as disgust and surprise.

References

[1] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives," *Frontiers in Robotics and AI*, vol. 7, Dec. 2020, doi: 10.3389/frobt.2020.532279.

[2] Broekens, J., Heerink, M., and Rosendal, H, "Assistive social robots in elderly care: a review," *Gerontechnology*, *8*(2), 94-103, 2009.

[3] Ahuja, Saransh, and Amir Shabani, "Affective Computing for Social Companion Robots Using Fine-grained Speech Emotion Recognition," *2023 IEEE Conference on Artificial Intelligence (CAI)*, pp. 331-332, 2023.

[4] Geetha, A. V., Mala, T., Priyanka, D., and Uma, E., "Multimodal Emotion Recognition with deep learning: advancements, challenges, and future directions," *Information Fusion*, *105*, 102218, 2024.

[5] Raotole, A., Shirodkar, S. S., Shukla, R., Sisodia, J., and Devadkar, K., "WellBe: An Intelligent Elderly Care and Well-Being Monitoring System Using Deep Learning," In *2023 4th International Conference on Intelligent Technologies (CONIT)*, pp. 1-6, June 2024.

[6] Swathi, G., Kumar, R. P., and Elakkiya, R., "Ensemble Integration of Deep Learning Models for Gender-Based Speech Emotion Recognition," In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-7, July 2023.

[7] Livingstone, S. R., and Russo, F. A., "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, 13(5), e0196391, 2018.

[8] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R., "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, 5(4), 377-390, 2014.

[9] I. Kansizoglou, L. Bampis, and A. Gasteratos,,"An active learning paradigm for online audio-visual emotion recognition," *IEEE Trans. Affective Comput.*, vol. 13, no. 2, pp. 756-768, 2019.

[10] Banskota, N., Alsadoon, A., Prasad, P. W. C., Dawoud, A., Rashid, T. A., and Alsadoon, O. H. ,"A novel enhanced convolution neural network with extreme learning machine: facial emotional recognition in psychology practices," *Multimedia Tools and Applications*, 82(5), 6479-6503, 2023.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput Vis Pattern Recognit.*, 2001.

[12] He, J., Yu, X., Sun, B., and Yu, L.,"Facial expression and action unit recognition augmented by their dependencies on graph convolutional networks," *Journal on Multimodal User Interfaces*, pp. 1-12.

[13] Krishnani, D., Shivakumara, P., Lu, T., Pal, U., Lopresti, D., & Kumar, G. H.,"A new context-based feature for classification of emotions in photographs," *Multimedia Tools and Applications*, 80, pp. 15589-15618, 2021.

[14] Sun, Q., Liang, L., Dang, X., and Chen, Y.,"Deep learning-based dimensional emotion recognition combining the attention mechanism and global second-order feature representations," *Computers and Electrical Engineering*, 104, 108469, 2022.

[15] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y.,"Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, *23*(10), pp. 1499-1503, 2016.

[16] Kim, N., Cho, S., & Bae, B.,"SMaTE: A segment-level feature mixing and temporal encoding framework for facial expression recognition," *Sensors*, *22*(15), 5753, 2022.

[17] Ghaleb, E., Niehues, J., and Asteriadis, S.," Joint modelling of audio-visual cues using attention mechanisms for emotion recognition. Multimedia Tools and Applications," 82(8), pp. 11239-11264, 2023.

[18] Zhu, C., Ding, T., and Min, X.," Emotion Recognition of College Students Based on Audio and Video Image," *Traitement du Signal*, *39*(5), 2022.

[19] Wu, Z., Zhang, X., Zhi-Xuan, T., Zaki, J., and Ong, D. C.,"Attending to emotional narratives," In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 648-654, September 2019.

[20] Hajarolasvadi, N., Bashirov, E., and Demirel, H.,"Video-based person-dependent and person-independent facial emotion recognition," *Signal, Image and Video Processing*, *15*(5), pp. 1049-1056, 2021.

[21] Wei, J., Yang, X., and Dong, Y.,"User-generated video emotion recognition based on key frames," *Multimedia Tools and Applications*, *80*, pp. 14343-14361, 2021.

[22] Zou, S., Huang, X., Shen, X., & Liu, H. ,"Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation," *Knowledge-Based Systems*, *258*, 109978, 2022.

[23] Lakshminarayana, N. N., Sankaran, N., Setlur, S., and Govindaraju, V.,"Multimodal deep feature aggregation for facial action unit recognition using visible images and physiological signals," In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1-4, May 2019.

[24] Sharafi, M., Yazdchi, M., Rasti, R., & Nasimi, F.,"A novel spatio-temporal convolutional neural framework for multimodal emotion recognition," *Biomedical Signal Processing and Control*, *78*, 103970, 2022.

[25] Wang, S., Qu, J., Zhang, Y., & Zhang, Y.,"Multimodal emotion recognition from EEG signals and facial expressions," *IEEE Access*, *11*, pp. 33061-33068, 2023.

[26] Zou, S., Huang, X., Shen, X., and Liu, H. ,"Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation," *Knowledge-Based Systems*, *258*, 109978, 2022.

[27] Chen, F., Shao, J., Zhu, A., Ouyang, D., Liu, X., and Shen, H. T.," Modeling hierarchical uncertainty for multimodal emotion recognition in conversation," *IEEE Transactions on Cybernetics*, *54*(1), pp. 187-198, 2022.

[28] Braunschweiler, N., Doddipatla, R., Keizer, S., & Stoyanchev, S.,"Factors in emotion recognition with deep learning models using speech and text on multiple corpora," *IEEE Signal Processing Letters*, *29*, pp. 722-726, 2022.

[29] Hu, Z., Wang, L., Luo, Y., Xia, Y., and Xiao, H.,"Speech Emotion Recognition Model Based on Attention CNN Bi-GRU Fusing Visual Information," *Engineering Letters*, 30(2), 2022.

[30] Mocanu, B., Tapu, R., and Zaharia, T.,"Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning," *Image and Vision Computing*, 133, 104676, 2023.

[31] Ryumina, E., and Karpov, A.,"Facial expression recognition using distance importance scores between facial landmarks," In *CEUR Workshop Proceedings*, Vol. 2744, pp. 1-10, December 2020.

[32] Shorten, C., and Khoshgoftaar, T. M.,"A survey on image data augmentation for deep learning," *Journal of big data*, *6*(1), 1-48, 2019.

[33] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C.,"Ghostnet: More features from cheap operations," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1580-1589, 2020.

[34] Misra, D., Nalamada, T., Arasanipalai, A. U., and Hou, Q.,"Rotate to attend: Convolutional triplet attention module," In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3139-3148, 2021.

[35] Meng, D., Peng, X., Wang, K., & Qiao, Y.,"Frame attention networks for facial expression recognition in videos," In *2019 IEEE international conference on image processing (ICIP)*, pp. 3866-3870, September 2019.

Cheng-Kai Lu (Senior Member, IEEE) received his B.S. and M.S. degrees from Fu Jen Catholic University, Taiwan, and his Ph.D. in Engineering from the University of Edinburgh, UK, in 2012. From 2016 to 2021, he served as a Senior Lecturer (UK system) in the Department of Electrical and Electronic Engineering at Universiti Teknologi PETRONAS (UTP), Malaysia. In 2022, he joined the Department of Electrical Engineering at National Taiwan Normal University (NTNU) as an Assistant Professor and was reinstated as Associate Professor after two years. Dr. Lu's research interests include medical imaging, embedded systems, artificial intelligence, and clinical decision support. With over eight years of industrial experience, he holds several licensed patents, including technologies adopted by the Republic of China Air Force. He has also contributed to production line automation.

Chien-Wei Lu received the M.Sc. degree in Electrical and Electronics Engineering in 2024 from the Department of Electrical Engineering, National Taiwan Normal University. His research interests include emotion recognition and lightweight convolutional neural networks for real-time implementation.

Guan Bo Lin is currently pursuing his M.Sc. degree in the Department of Electrical Engineering at National Taiwan Normal University. His research interests include emotion recognition and lightweight generative AI architectures for real-time implementation.