Improved Attention-Enhanced Efficient Face-Transformer Model for Multimodal Elderly Emotion Recognition in Smart Homes

Peiyao Li
School of Architecture and Design, Yangtze University College of Arts and Sciences, Jingzhou, 434020, China
E-mail: peiyaolip@outlook.com

Keywords: home environment, elderly people, emotion recognition, attention, transformer

Received: April 9, 2025

Recognizing the emotions of the elderly is key to achieving personalized services in smart home environments. Traditional methods have difficulty capturing the correlation and temporal information of multimodal features. To this end, the study proposes a multimodal emotion recognition model that integrates the EfficientFace and Transformer structures to construct an improved attention mechanism. A modal interaction compensation term is introduced into the similarity calculation to improve the ability to model dynamic dependencies between modalities. Meanwhile, the dynamic importance factor is used to adaptively adjust feature weights. The model was tested on IEMOCAP and self-constructed EMED datasets. The emotion recognition precision reached up to 94.37%, the recall rate reached 93.64%, the F1 value was 94.21, and the specificity reached 94.85%. Additionally, the model achieved 96.24% classification accuracy and 94.13% emotional intensity on easily confused categories, such as "disgust" and "contempt," with a minimum detection latency of 0.55 seconds. The results show that the model exhibits excellent performance in multimodal fusion and emotion recognition for the elderly, and is suitable for the task of smart home emotion monitoring.

Povzetek: Članek uvaja izboljšan EfficientFace-Transformer z nadgrajenim mehanizmom pozornosti za multimodalno prepoznavo čustev starejših, ki dosega boljšo točnost, stabilnost in nizko zakasnitev v pametnih domovih.

1 Introduction

With the acceleration of global aging process, smart home as an emerging technology can provide convenient life services for the elderly. At the same time, the special needs of the elderly population in terms of physiology, psychology and cognition make emotion recognition (ER) an important research direction in smart home systems. Emotion is an important part of human psychological activities. Changes in the emotions of the elderly are often influenced by health conditions, environmental factors, and social interactions. Moreover, ER can provide personalized services for smart home systems to optimize the living experience of the elderly [1-3]. To solve the issues of elderly people having trouble recognizing emotions and traditional machine learning models' incapacity to adequately capture the nonlinear relationship between physiological signal data, Feng G et al. proposed a recursive mapping method of ER in conjunction with a visual transformer. According to experimental results, this method's recognition accuracy was up to 94.35%, which was higher than that of the conventional approach [4]. Using an inverted neural network as the main body and training data from audio and video modalities, Sreevidya et al. created an automated emergency room system designed especially for the elderly. Experimental results showed that the system showed a minimum relative improvement of 6.5% for happy emotions and a maximum relative improvement of 46% for sad emotions compared to the baseline model [5]. Park H et al. used MobileNet-V2 to recognize six emotions for different age groups. The results showed that the ER of teenagers was more obvious than that of older people, and the model recognition accuracy was up to 83.7% [6]. To investigate the emotional experience of the elderly when using smart terminals, Lu et al. The researchers proposed a novel accessibility ER model for the elderly after combining with the support vector machine network. The outcomes displayed that the model had the highest accuracy in detecting the negative emotions of the elderly when using the terminal. Moreover, it could help reflect the design defects of the terminal application technology [7].

Speech recognition, image processing, natural language processing, and other domains have made extensive use of the attention mechanism (AM) in recent years. Multimodal learning and time-series data analysis have seen particularly impressive outcomes [8, 9]. To investigate the hidden emotional states from human motion, Zhao et al. proposed a human multi-site ER method that combined a fuzzy algorithm, AM, and multiple inertial measurement units. The outcomes revealed that the method had the highest accuracy of 94.02% for six common types of emergencies [10]. Saganowski developed a two-channel recognition technique in an attempt to improve the robustness of the existing ER technique by introducing the AM to optimize the attention of the traditional convolutional recognition network. Experimental results indicated that the effectiveness of this technique for human ER was stronger than before the improvement [11]. According to Zhou et al., the majority of ER models now in use overlook inter-feature interactions and fall short of capturing the crucial complimentary benefits between contextual and facial information in video clips. According to the findings, the network was quite successful in predicting the emotions associated with work-related stress based on visual emotional evidence [12]. Lin et al. came to the conclusion that emotional shifts brought on by outside factors were accompanied by measurable and identifiable alterations in physiological signals. For this reason, this study proposed a dynamic ER model by combining EEG signals, ECG signals, EMG signals, and AM. The experimental results indicated that the model was robust to ER in application scenarios with different physiological signals and was suitable for different occasions of ER work [13]. The review of the above literature is summarized in Table 1.

Table 1: Literature review summary comparison table

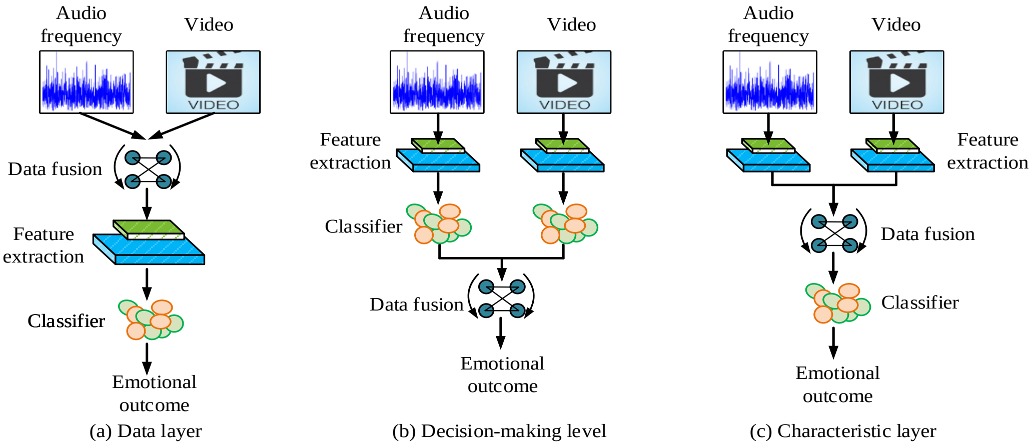
Authors	Method/Model	Advantages (Metrics)	Limitations
Feng G et al. [3]	Visual Transformer+recurrent mapping	Accuracy up to 94.35%	Dependent on physiological signals, weak generalization
Sreevidya P et al. [4]	Audio-visual Fusion+recurrent neural network	46% improvement for sadness; 6.5% for happiness	No refined cross-modal interaction mechanism
Park H et al. [5]	MobileNet-V2	Accuracy up to 83.7%	Limited performance on elderly users
Lu J et al. [6]	SVM-based emotion recognition model	High accuracy in detecting negative emotions	Limited to mobile terminal usage scenarios
Zhao Y et al. [8]	IMU+Attention+Fuzzy Algorithm	Accuracy up to 94.02% for six emotions	Single modality based on IMU only
Saganowski S [9]	Dual-channel CNN with Attention	Improved performance over standard CNN	Insufficient modeling of inter-modal interaction
Zhou S et al. [10]	Cross-attention+hybrid feature weighting	Effective fusion of facial and contextual cues	Complex modeling, lacks real-time performance
Lin W et al. [11]	EEG/ECG/EMG+dynamic attention	High robustness across physiological modalities	Heavy reliance on sensors, low adaptability

In summary, although existing research has produced some results in the field of ER, significant deficiencies remain in addressing the complexity of emotional expression and the unique multimodal feature interactions of the elderly population. These deficiencies are evident in the rough modal information fusion strategy and limited cross-modal dynamic modeling capability. This makes it difficult to meet the demand for high-precision, low-latency ER in smart homes. To this end, the study focuses on the following core research questions: (1) Can the introduction of dynamic importance factors significantly improve the accuracy of modal feature weight adjustment? (2) Can the modal interaction compensation term enhance the co-modeling capability between audio and video? (3) Can the combination of the above mechanisms improve the F1 value by at least three percentage points over the baseline model? To validate the above questions, the study focuses on two key challenges of multimodal ER in older adults. First, it proposes an end-to-end recognition model that incorporates the EfficientFace-Transformer structure with an improved AM. The model introduces a dynamic importance factor and a modal interaction compensation term. These terms strengthen the deep interaction relationship between the audio and video modalities, as well as the ability to regulate weights. This allows for more efficient information fusion and robust modeling while maintaining the integrity of the modal representation. The core innovation of this study is the application of an optimized AM to ER scenarios for the elderly for the first time. A generalizable, multimodal emotion fusion framework is systematically constructed and verified for actual smart home requirements in terms of performance and practicability. This provides a feasible path and theoretical support for constructing a personalized emotion service system.

2. Methods and materials

2.1 Multimodal emotion recognition based on EfficientFace-Transformer

Due to the special physiological, psychological, and cognitive needs of the elderly, their emotional expressions are often different from those of young people and are affected by a variety of factors, such as health status, living environment, and social interactions [14-15]. Therefore, a single ER method may not be able to comprehensively and accurately reflect the emotional state of older adults. The multimodal fusion approach is a useful technique for enhancing ER's accuracy and robustness. By integrating data from multiple modalities, such as facial expression, speech, movement, heart rate, etc., it can make up for the possible limitations of a single modality and provide a more comprehensive and accurate emotion analysis [16-17]. Whereas the effectiveness of the multimodal fusion method mainly relies on the close collaboration of data layer (DL), feature layer (FL), and decision layer (DeL). The schematic diagram of DL, FL, and DeL is shown in Figure 1 [18].



(a) Data layer (b) Decision-making level (c) Characteristic layer

Figure 1: Diagram of data layer, decision layer, and feature layer of multimodal fusion

Figure 1(a), 1(b), and 1(c) shows the schematic diagram of DL, DeL, and FL of multimodal fusion. Data from different modalities such as audio and video in the DL are directly fed into the system and fused to form a joint multimodal data representation. Subsequently, these data are passed through a unified feature extraction module to extract key features and a classifier to classify the emotions. The DeL and FL processes are roughly similar, differing only in the order of feature classification and data fusion. It can be concluded that the multimodal fusion method can capture the emotion information more comprehensively and improve the robustness and accuracy of ER compared to the single modality recognition method. However, traditional multimodal fusion recognition methods often use separate feature extraction and simple fusion strategies when dealing with data from different modalities, which makes it difficult to fully explore the inter-modal correlations and temporal information [19]. For this reason, the study relies on the complementary information between different modalities and makes full use of this feature to propose an end-to-end ER model for the elderly. Figure 2 displays the model's structure.

In Figure 2, in the audio modality, the features of the audio signal are first extracted by Mel-frequency cepstral coefficients (MFCC) feature extraction algorithm, which generates audio features reflecting the rhythm, pitch, and emotion intensity of speech. In video modality, features of facial expressions in video frames are extracted using EfficientFace network structure. Next, one-dimensional convolutional coding is applied to the audio and video features to improve their temporal modeling capabilities. Then, they are fed into the Transformer module, which uses its self-AM to capture intra- and inter-modal feature relationships and generate fused multimodal feature representations. The fused features are further dimensionality reduction processed through the fully connected layer to extract the key emotion information features. Finally, the Softmax classifier is used to complete the emotion classification task, outputting the recognition outcomes of the emotion categories including pleasure, anger, sadness, surprise, and so on. To ensure reproducibility and stable training of the model, the study implements the following configuration: The Transformer module is set as a stacked structure of three layers, each of which contains eight heads of multi-head attention. The embedding dimension is 256. The width of the feedforward layer is 512. Moreover, the activation function is GELU. A dropout operation with a rate of 0.1 is added in the fusion stage to prevent overfitting. The model is trained using the Adam optimizer with an initial learning rate of 1×$10^{-4,}$ a batch size of 32, and 120 training rounds. The Early Stopping strategy is introduced to stabilize convergence performance. Figure 3 displays the schematic diagram of the EfficientFace-Transformer, which serves as the backbone network of the entire recognition model.

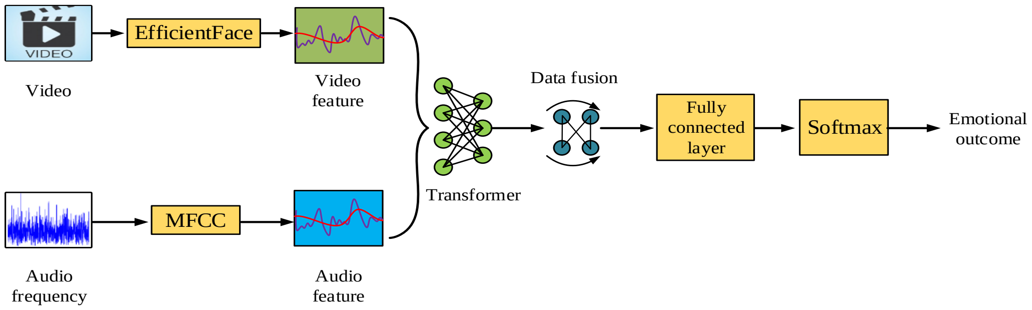


Figure 2: Multimodal end-to-end emotion recognition model for the elderly

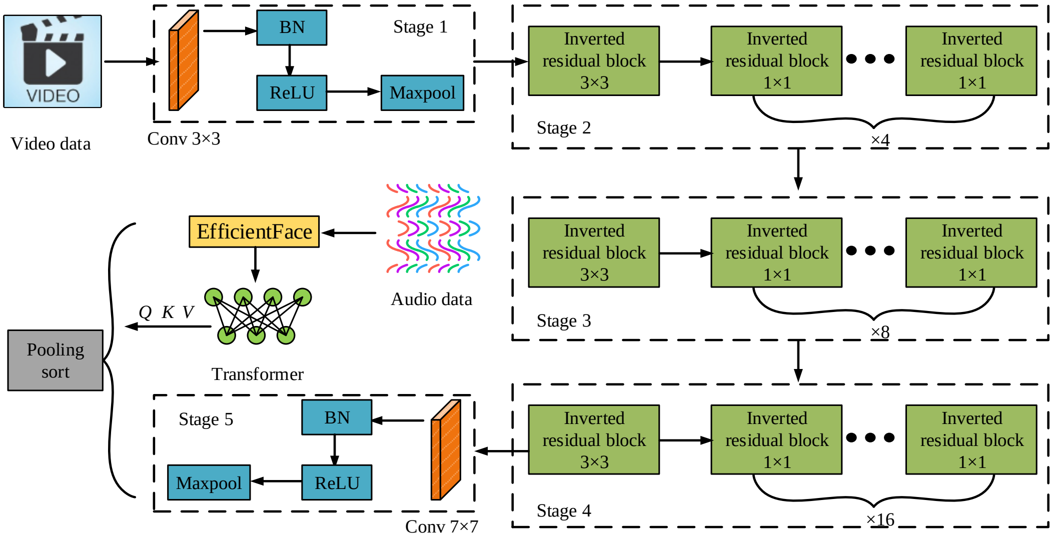


Figure 3: EfficientFace-Transformer diagram

In Figure 3, first, the data of video modality is fed into the EfficientFace network for multi-stage feature extraction. Among them, Stage 1 completes the generation of preliminary feature representations through convolutional and maximum pooling layers. Stages 2 to 4 use stacked inverted residual blocks to extract high-level semantic features of the video modality. In terms of processing the temporal dimension of video frames, the study employs a strategy combining frame-level feature splicing and average pooling. First, feature vectors are extracted for each frame independently. Next, frame sequence feature representations are obtained by splicing the frames one by one in the temporal dimension. Finally, frame sequence features are integrated via time-averaged pooling to generate compact feature vectors representing the entire video clip. Stage 5 generates compact feature vectors by global average pooling dimensionality reduction. Meanwhile, the audio modality is processed by a 1D convolutional layer on the input MFCC features to extract the key features of the audio modality. Next, the features of the video modality and the features of the audio modality are used as inputs to the Transformer module, respectively. The Transformer computes the relationship between Q, K, and V through its self-AM to dynamically capture the intra- and inter-modal interaction features. Subsequently, after pooling operation, the fused features of audio and video modalities are further compressed and integrated, and finally the task of classifying emotion categories is accomplished through the fully connected layer. Since the audio is a one-dimensional temporal data, it needs to be first framed and windowed to maintain its temporal structure, and the computation process is shown in Equation (1) [20].

$$x_n(m) = x(a + nR) \cdot \omega(m) \tag{1}$$

In Equation (1), $x_n(m)$ denotes the signal of n frame. m denotes the time index within the frame. R denotes the interval between the upper and lower frames. $\omega(m)$ denotes the window function. The video data consists of a series of consecutive frames, and the frames need to be sampled, grayscaled and normalized in preprocessing to extract the key expression information related to emotions. The computational formula is shown in Equation (2).

$$I'(x, y) = \frac{I(x, y) - \mu}{\sigma} \tag{2}$$

In Equation (2), $I(x, y)$ and $I'(x, y)$ are the pixel values before and after normalization, respectively. $I(x, y)$ is the mean of the pixel values. $I(x, y)$ is the standard deviation of the pixel values. After preprocessing, the formula for calculating the filtered energy on the Mel scale is shown in Equation (3).

$$E_i = \sum_{k=f_{i-1}}^{f_i} \left| X(k) \right|^2 H_i(k) \tag{3}$$

In Equation (3), E_i is the energy of the i th Mel filter.

$H_i(k)$ is the transfer function of the i th filter. f_i and f_{i-1} are the upper and lower limits of the Mel filter frequency, respectively. $|X(k)|^2$ denotes the frequency domain power spectrum. Equation (4) illustrates how the Transformer module determines the attention weights.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

In Equation (4), Q, K, and V distributions denote the query, key, and value matrices. d_k denotes the dimension of the key vector for mapping the fused features to the sentiment category probability distribution. Equation (5) displays the computational formula.

$$P(y=i|z) = \frac{e^{w_i^T z+b_i}}{\sum_{j=1}^{C} e^{w_j^T z+b_j}} \tag{5}$$

In Equation (5), $P(y=i|z)$ denotes the probability that the sample belongs to emotion category i. z denotes the fusion feature. w_i and b_i denote the weight and bias of the i th emotion category, respectively. C denotes the total number of emotion categories.

2.2 Optimization of a multimodal emotion fusion recognition model for the elderly incorporating the attention mechanism

After constructing the completed EfficientFace-Transformer-based multimodal emotion fusion recognition model, it is found that the EfficientFace-Transformer is a multimodal information fusion of audio and video at a later stage. This, while better able to preserve the independent feature representation of each modality, still has some limitations in capturing the deep inter-modal interactions and joint dynamic properties. Specifically, the late fusion strategy mainly relies on a single fusion layer before classification for the integration of modal features, which may lead to insufficient inter-modal information interactions, thus affecting the robustness and accuracy of ER [21]. To improve the ability to model multimodal feature interactions, the study optimized the self-AM of the standard Transformer module. The Transformer still serves as the modal fusion backbone network in the overall design, but the part that computes internal attention is replaced with an improved AM. This mechanism introduces three terms to the standard dot product attention: a modal interaction compensation term, a dynamic importance factor, and a cross-modal enhancement term. These terms enhance the inter-modal collaborative modeling capability. Figure 4 shows the structure of the traditional dot product AM for comparison purposes. Figure 5 shows the structure of the improved attention module embedded in the Transformer of the final model, as proposed in the study. Figure 4 shows the structure of the standard dot product AM for comparison with the improved mechanism. The improved mechanism has not been directly applied to the final model [22].

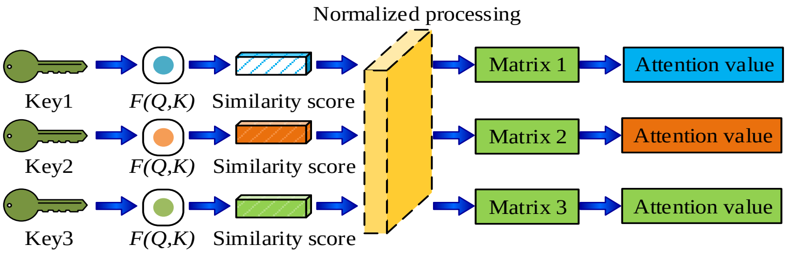


Figure 4: Attention mechanism diagram

Figure 4 shows that the AM is divided into three stages. These stages gradually complete feature weighting, strengthen the interaction between audio and video, and highlight key emotional information. The initial step involves mapping the input characteristics into query, key, and value vectors. Then, using the dot product, the similarity score between each query and all keys is calculated. The weight matrix is obtained in the second stage by normalizing the similarity scores produced in the first stage using the Softmax function. Equation (6) displays the calculating formula.

$$S1_{ij} = soft\max\left(\frac{\exp(F(Q_i,K_j)+W_m)}{\sum_j \exp(F(Q_i,K_j)+W_m)}\right) \tag{6}$$

In Equation (6), $S1_{ij}$ denotes the i th row and j th column element of the normalized weight matrix. $F(Q_i,K_j)$ denotes the similarity score matrix obtained from the first stage by Equation (4). W_m denotes the modal association weight vector. In the third stage, the attention value is also combined with the residual connections between modalities to preserve the original features of the modalities and enhance the training stability of the model.

Equation (7) displays the computational formula.

$$AttentionValue = S1 \cdot V + \lambda U \tag{7}$$

In Equation (7), λ denotes the residual scale factor. U denotes the initial input features. V denotes the initial eigenvalue matrix. In order to capture the deep interactions between audio and video more effectively, the AM is improved in this study. The improved AM is shown in Fig. 5. Its structure is used to replace the standard self-attention module in Transformer and embedded into the final model.

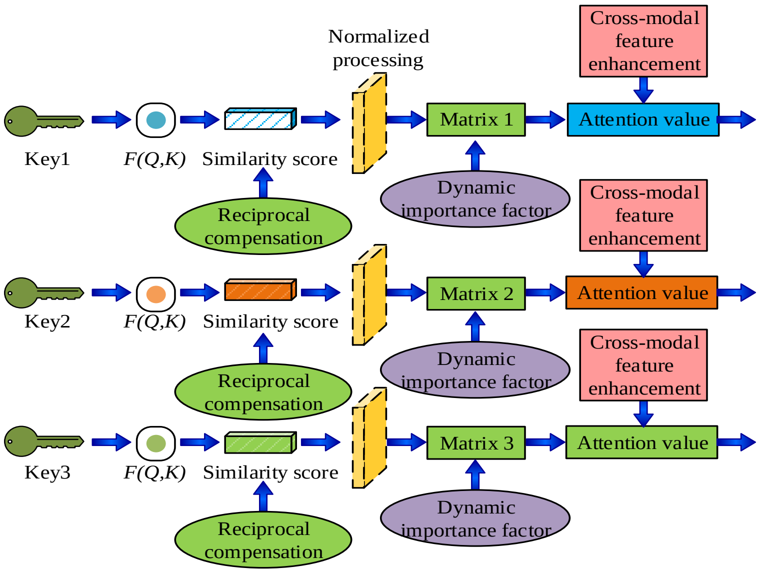


Figure 5: Improved attention mechanism diagram

In Figure 5, the improved AM is mainly generated by the similarity computation improvement, the weight normalization dynamic adjustment, and the enhancement of the final attention value. In the similarity calculation stage, the modal interaction compensation term is introduced. By combining the dot product similarity with the modal interaction compensation term, the dynamic dependency modeling between modes is realized. In the weight normalization stage, a dynamic importance factor is added. The generated dynamic importance factor is able to adaptively adjust the weights of modal features by computing the differences between the global features of audio and video modalities. In the attention value generation stage, a cross-modal feature (CMF) enhancement term is added. The fusion effect of audio and video modalities is further enhanced by feature splicing and weight adjustment. In this process, the similarity score is calculated as shown in Equation (8).

$$\begin{cases} F'(Q,K) = \dfrac{QK^T}{\sqrt{d_k}} + B_m \\ B_m = W_b \cdot (ReLU(Q \cdot K^T) - \gamma), \end{cases} \tag{8}$$

In Equation (8), B_m denotes the modal interaction compensation term. W_b denotes the learnable modal compensation weight matrix. γ denotes the modal balance bias. d_k denotes the dimension of the key vector. $F'(Q,K)$ denotes the similarity score after considering the modal compensation term. Then, the dynamic importance factor is added in the weight normalization stage. The D_m calculation formula is shown in Equation (9).

$$D_m = \eta(W_d \cdot (G_a - G_v)) \tag{9}$$

In Equation (9), D_m denotes the dynamic moderator based on modal significance. G_a and G_v denote the global features of audio and video modalities, respectively. W_d denotes the importance adjustment weights. η denotes the Sigmoid function. Finally, in the attention value generation stage, the feature representation is further optimized by introducing CMF enhancement terms in combination with inter-modal contextual information. The attentions value calculation at this time is shown in Equation (10).

$$AttentionValue' = S1 \cdot V + \lambda U + C_m \tag{10}$$

In Equation (10), $AttentionValue'$ denotes the final stage attentional output result. $S1 \cdot V$ denotes the first stage attentional output. C_m denotes the CMF enhancement term, which is calculated as shown in Equation (11).

$$C_m = W_c \cdot (Concat(G_a, G_v)) \tag{11}$$

In Equation (11), W_c denotes the cross-modal enhancement weight matrix. $Concat(\_)$ denotes the feature splicing operation. In summary, the study finally proposes a novel multimodal emotion fusion recognition model for the elderly that combines EfficientFace-Transformer and improved attentions. Figure 6 displays the model's structure.

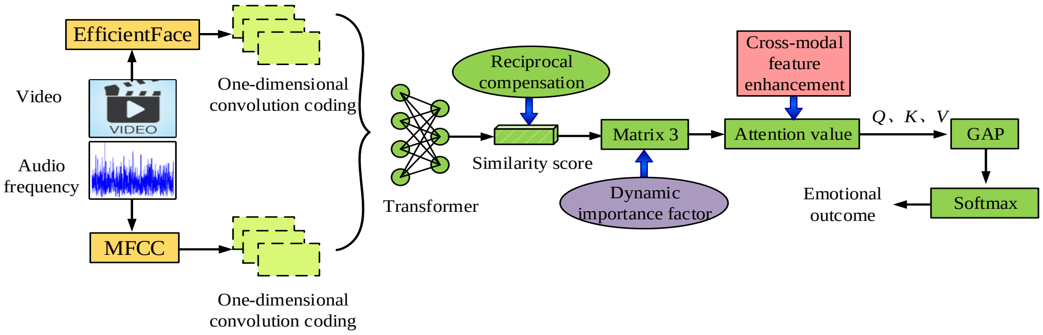
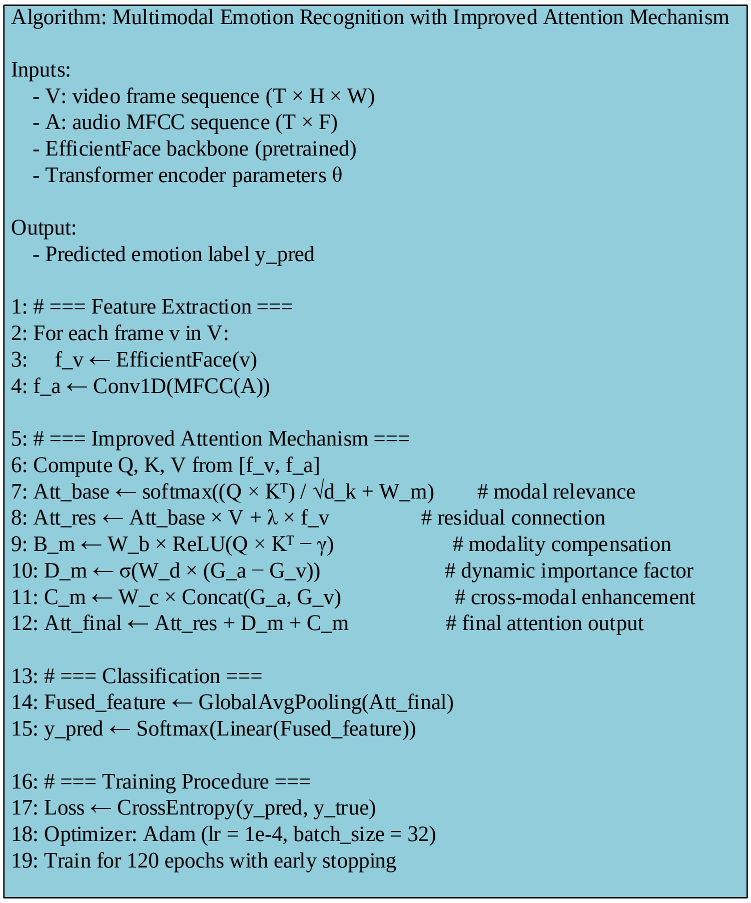


Figure 6: New multimodal emotion fusion recognition model for the elderly

Figure 6 shows that the video modalities' features are first extracted through the EfficientFace network for multilayer representation. Meanwhile, the audio modalities' features are encoded through a one-dimensional convolutional layer to preserve their time-series information. During the intermediate fusion phase, the Transformer module uses an improved AM, as shown in Fig. 5. This mechanism replaces the standard self-attention operation to improve the model's ability to represent global relationships between modalities. The modal feature weights are adaptively adjusted through an improved AM combined with a dynamic importance factor, which significantly enhances the deep inter-modal interaction capability. Specifically, the AM dynamically introduces a modal compensation term and a global feature enhancement term on top of Q, K, and V. The model is designed to capture emotionally relevant CMFs more accurately. This enables the model to capture emotionally relevant CMFs more accurately. The fused features are further processed through a 1D convolutional layer to enhance feature compactness. It is also combined with a global pooling layer to generate the final emotion feature representation, and finally the recognition of emotion categories is completed by a classification layer. Finally, the identification of emotion categories is accomplished through the classification layer. This is the model architecture that represents the final model proposed in the results and conclusions section. The algorithmic pseudo-code of the final model is shown in Figure 7.



```
Algorithm: Multimodal Emotion Recognition with Improved Attention Mechanism

Inputs:
   - V: video frame sequence ($T × H × W$)
   - A: audio MFCC sequence ($T × F$)
   - EfficientFace backbone (pretrained)
   - Transformer encoder parameters $\theta$

Output:
   - Predicted emotion label y_pred

# === Feature Extraction ===
For each frame v in V:
    f_v ← EfficientFace(v)
f_a ← Conv1D(MFCC(A))

# === Improved Attention Mechanism ===
Compute Q, K, V from [f_v, f_a]
Att_base ← softmax($(Q × K^T) / \sqrt{d\_k} + W\_m$)      # modal relevance
Att_res ← Att_base × V + $\lambda$ × f_v                  # residual connection
B_m ← W_b × ReLU($Q × K^T − \gamma$)                       # modality compensation
D_m ← $\sigma$(W_d × (G_a − G_v))                          # dynamic importance factor
C_m ← W_c × Concat(G_a, G_v)                               # cross-modal enhancement
Att_final ← Att_res + D_m + C_m                            # final attention output

# === Classification ===
Fused_feature ← GlobalAvgPooling(Att_final)
y_pred ← Softmax(Linear(Fused_feature))

# === Training Procedure ===
Loss ← CrossEntropy(y_pred, y_true)
Optimizer: Adam (lr = 1e-4, batch_size = 32)
Train for 120 epochs with early stopping
```

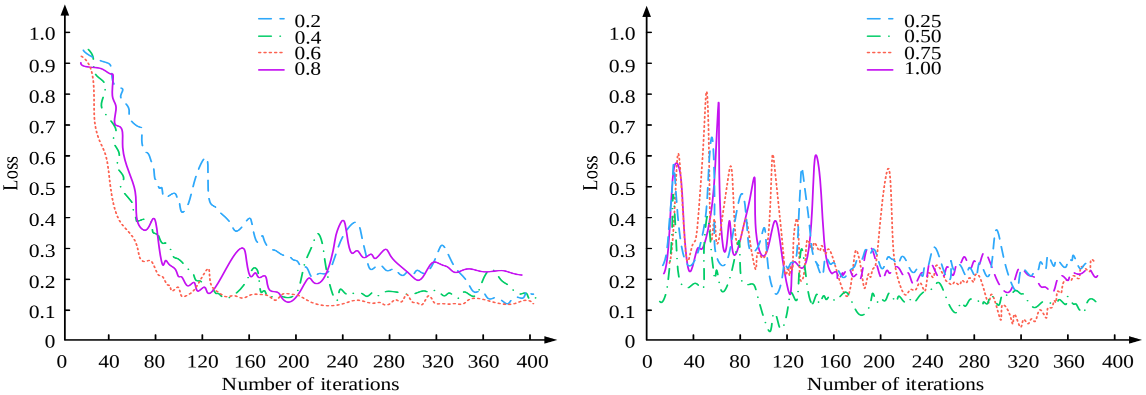
Figure 7: Algorithmic pseudo-code for the final model

3. Results

3.1 Performance test of multimodal emotion fusion recognition model for the elderly

The study is set up with AMD Ryzen 9 5950X for CPU, NVIDIA GeForce RTX 3090 for GPU, 64GB of RAM, and Ubuntu 20.04 for the operating system. The PyTorch framework is also used for model implementation. Two multimodal ER datasets are used for model validation: IEMOCAP and the self-constructed EMED dataset. IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) is a publicly available multimodal emotion dataset containing audio and video data of 10 actors expressing multiple emotion categories such as anger, happiness, sadness, surprise, etc. The EMED (Elderly Multimodal Emotion Dataset) is a dataset designed for this study. It is intended for use with an elderly population and is designed to elicit real emotional responses through natural conversations, image stimulation, and situational recollection. It also synchronizes facial video and voice signals. Facial videos and voice recordings are collected simultaneously. The dataset includes the five most common emotional states of the elderly: pleasure, sadness, anger, anxiety, and neutrality. It covers multiple home environments and interaction scenarios to simulate the actual perceptual needs of smart home applications. The study first tests the selected values of two types of hyperparameters that have a large impact on the pre-signal feature extraction and mid-signal feature fusion. The test results are shown in Figure 8.

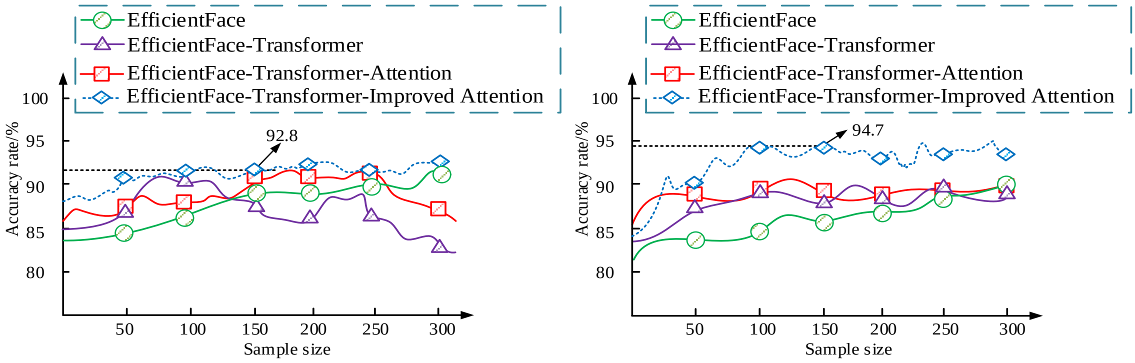
The test results for the emotion weights are displayed in Figure 8(a). The test results of calculating the importance regulatory weights are displayed in Figure 8(b). Figure 8 shows the average of five independent training runs performed with the corresponding weight settings, represented by each loss curve. While the standard deviation is not shown directly in the figure, the average trend of all configurations remains consistent across multiple runs, and the convergence interval fluctuates minimally, with a maximum standard deviation of less than ±0.025. This verifies the reproducibility and controllability of the hyperparameters on the model's convergence. In Figure 8(a), the loss function of the model decreases the fastest when the emotion weight is 0.6, and the loss value after final convergence is the lowest. It indicates that the model can more effectively balance the significance of features from various emotion categories with this weight value, increasing feature extraction accuracy and efficiency. When the weights take the value of 0.2 or 0.8, the convergence speed of the model becomes slower and the final loss value is higher. It indicates that too low or too high emotion weights can lead to bias in the model’s capture of emotion information. In Figure 8(b), the model has the least fluctuation at the beginning of the iteration when the weights take the value of 0.75. The downward trend of the loss function is the most stable, and the final loss value is also optimal. This suggests that the model may distribute the weight percentage of various modal features more efficiently at this weight value. Thus, it can capture important emotional information more precisely during the feature fusion process. However, the model’s iterative process exhibits greater volatility and worse convergence when the weights are set to 0.25 or 1.00. It implies that incorrect weighting could result in a drop in feature fusion effectiveness. The study tested the model for ablation and the results are shown in Figure 9. Among them, attention denotes the use of the standard dot product AM (as shown in Figure 4) and improved attention denotes the introduction of the optimization structure of modal interaction compensation with importance factors (as shown in Figure 5).



(a) Emotional weight

(b) Importance adjustment weight

Figure 8: Hyperparameter value test



(a) IEMOCAP

(b) EMED

Figure 9: Ablation test results

Figure 9(a) displays the outcomes of the model ablation test in the IEMOCAP and EMED dataset. In Figure 9(a), the EfficientFace-Transformer-Improved attention model has the highest accuracy of 92.8% for a sample size of 300, which is significantly better than the other models. In contrast, the EfficientFace model has the lowest accuracy of about 86.4%. The introduction of the Transformer and improved AM significantly improves the modeling ability and recognition performance of intermodal interactions. Figure 9(b) shows that the Improved attention model also performs best on the EMED dataset, with an accuracy of 94.7% when the sample size reaches 300. In contrast, the EfficientFace model performs the worst, indicating the limited effectiveness of unimodality in the recognition of emotional features in the elderly. Improved AM and multimodal fusion strategy are more advantageous. Additionally, the accuracy curves in Fig. 8 are plotted based on the mean results of each group of model configurations that are trained independently five times with the same sample size. In the experiments, the standard deviation of the accuracy of each model under different sample sizes is calculated. The standard deviation of EfficientFace-Transformer-Improved Attention is controlled to be within ±1.2% when the number of samples is greater than 150. This significantly outperforms the other compared models. This result indicates that the model not only has high accuracy, but also performs stably with good statistical consistency under different training subsets. The study continues to test the effects of different modular incremental optimizations on the final model's performance. This includes the effects of independently introducing modal interaction compensation, a dynamic importance factor, and CMF enhancement terms. The results are shown in Table 2.

Table 2: Modal interaction compensation and dynamic significant factor ablation test results

Model Variant	Emotional intensity/%	mAP/%	Detection delay/s
Baseline (EfficientFace-Transformer)	89.14	91.78	0.62
+ Modality Interaction Compensation	91.03	93.4	0.61
+ Dynamic Importance Factor	90.84	93.18	0.61
+ CMF Enhancement (Final)	93.67	95.24	0.55

In Table 2, after adding only the modal interaction compensation term, the affective strength is increased from 89.14% to 91.03%, and the mean average precision (mAP) from 91.78% to 93.40%. Meanwhile, the detection latency is slightly decreased to 0.61 s. The addition of the dynamic importance factor alone improves the affective strength to 90.84%, the mAP to 93.18%, and the latency remains the same. Further incorporating CMF enhancement optimizes the model's performance, increasing the sentiment intensity and mAP to 93.67% and 95.24%, respectively, while reducing the detection latency to 0.55 seconds. This demonstrates significant improvements in accuracy and real-time performance through the synergistic optimization of multiple mechanisms. This ablation validation clearly shows that each component contributes to performance gains in different ways. The final improved model achieves systematic improvements in fusion performance while maintaining recognition efficiency. Visual geometry group face (VGGFace), face embedding network (FaceNet), and efficient neural network (EfficientNet) are introduced in the study. The precision, recall, F1 value, and specificity are used as metrics for comparison testing. Table 3 displays the findings.

In Table 3, in the IEMOCAP dataset, the F1 value of the proposed model is studied to reach 92.44%, which is about 3.29% higher than that of EfficientNet, and the specificity reaches 93.12%, reflecting its significant advantage in reducing misidentification. The suggested model’s performance is considerably better in the EMED dataset. Compared to the other models under comparison, the F1 value, precision, and recall are all substantially higher at 94.21%, 94.37%, and 93.64%, respectively. This indicates that the proposed model under study has stronger generalization ability and robustness in specific scenarios targeting elderly ER. In contrast, VGGFace has relatively insufficient feature extraction capability due to its shallow network structure, which leads to its performance being lower than FaceNet and EfficientNet on both types of datasets. FaceNet and EfficientNet show significant improvement in ER performance through deeper network structure and optimization strategies. However, it still performs less well than the the proposed model in the complex multimodal emotion fusion task.

3.2 Simulation test of multimodal emotion fusion recognition model for the elderly

The study extracts seven types of emotional states (anger, disgust, fear, happiness, surprise, sadness, and contempt) from the IEMOCAP dataset. These states are used to test the fine-grained classification performance of multimodal ER. The study excludes the interference of other modal information, such as audio. It also compares the differences in emoji emotion classification performance between models in a unimodal manner. The study plotted the confusion matrix results are shown in Figure 10.

Figure 10(a), 10(b), 10(c), and 10(d) shows the expression recognition results of VGGFace model, FaceNet model, EfficientNet model, and proposed model. The recognition accuracy of the research proposed model for the seven types of expressions, namely, angers, detest, fear, happy, amazed, sad, and comtempt, are 99%, 95%, 93%, 92%, 95%, 97%, and 96%, respectively, which surpass those of the other models by a considerable margin. Especially, it shows higher accuracy on the difficult-to-distinguish emoji categories such as detest and comtempt. In contrast, the VGGFace model, while better at recognizing ananger and comtempt, performs the worst with a recognition accuracy of only 69% on the fear category. In addition, the FaceNet model outperforms VGGFace overall, but has 76% and 72% on the amazed and fear categories, respectively, indicating its limited ability to distinguish complex expression features. The EfficientNet model recognizes the anger and sad categories better, but the accuracy on the amazed category is only 65%, which is also deficient. Receiver Operating characteristic curve (ROC) is used as an indicator to validate the most prominent expression emotions of anger and sadness, and the results are shown in Figure 11.

Figure 11(a) shows the ROC curves of different models for the detection of anger emotion, and Figure 11(a) shows the ROC curves of different models for the detection of sadness emotion. Among them, the horizontal axis represents the false-positive rate, and the vertical axis represents the true-positive rate. The larger the area under the curve (AUC) value surrounded by the ROC curve and the horizontal and vertical coordinates, the better the model's performance. In Fig. 11(a), In the anger ER task, the ROC curve of the model proposed in this study is significantly higher than those of the other three comparative methods. It has a larger overall coverage area, and its final AUC value is 0.91. This is better than the values of EfficientNet (0.84), DenseNet (0.85), and Xception-DeepLab (0.82). As shown in Fig. 11(b), the model presented in this study still performs best for sadness detection, with an AUC value of 0.87. This demonstrates its ability to accurately perceive and model emotional changes. Notably, the ROC curves of other methods in both emotion categories show different degrees of jitter or a tendency to be close to the diagonal. In contrast, the proposed model can obtain a high true positive rate when the false positive rate is low. This indicates that the model possesses strong early differentiation ability while ensuring accuracy. The models are evaluated based on their performance in classification consistency for multimodal ER, as shown in Figure 12. This metric measures the consistency of a model's predicted results for each emotion type across multiple independent runs on the same sample set. Its computational formula is shown in Equation (12).

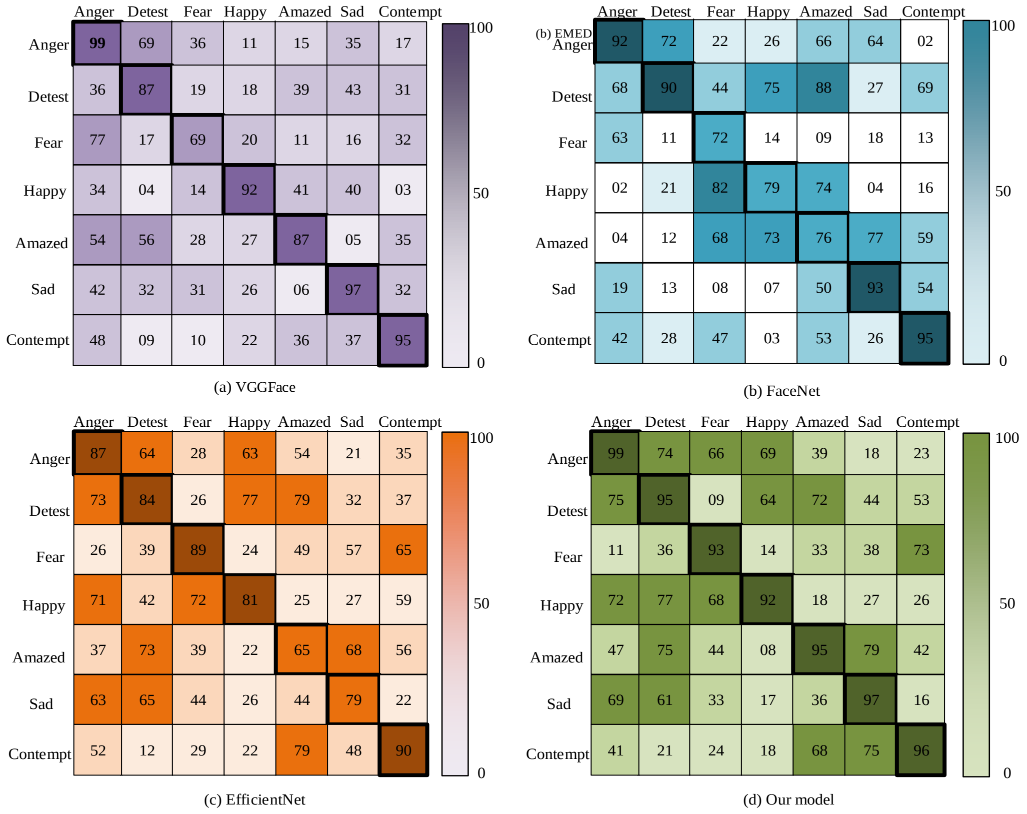
$$ECC = \frac{1}{C}\sum_{i=1}^{C}\frac{n_i^{consistent}}{n_i} \tag{12}$$

In Equation (12), C represents the total number of emotion categories. n_i represents the total number of samples in the test set for category i emotions. $n_i^{consistent}$ represents the number of samples with consistent prediction results over multiple times in this category. This indicator takes the value range of [0,1], which can be converted to percentage expression (unit: %). The higher value represents the more stable classification results of the model on the category of emotions. The study continues with emotion classification consistency as a metric. The results are shown in Figure 12.

Figure 12(a) shows the different model emotion classification consistency test results under IEMOCAP dataset. Figure 12(b) shows the results of emotion classification consistency test for different models under EMED dataset. In Figure 12(a), the emotion classification consistency of the proposed model is always maintained at a high level in both types of data. It also rapidly approaches the ideal variation curve with the increase of the number of samples, and its emotion classification consistency reaches up to 96.24%. In contrast, the VGGFace model has the lowest consistency of 85.74%. When the number of samples is large, its classification stability significantly decreases. The FaceNet model performs more consistently in the medium sample size range, but its agreement rate is only about 90.23% at high sample sizes, failing to improve further. The EfficientNet model has a slightly higher agreement rate than FaceNet, but when dealing with complex emotion categories, its agreement rate increases significantly less than that of the proposed model. In summary, studying the proposed model shows better recognition accuracy and stability in a variety of ER tests and is suitable for complex ER tasks. To further evaluate the adaptability of the proposed models under different data distribution and acquisition scenarios, this study conducts cross-dataset validation experiments. Specifically, the models undergo cross-training on the IEMOCAP dataset and cross-testing on the EMED dataset. This examination assesses the models' ability to recognize emotions in unseen samples from the target domain. Table 4 summarizes the main evaluation metrics of each model in this generalization test, including emotion intensity, detection latency, and mean accuracy of recognition (mAP). The goal is to validate the robustness and stability of the models in a cross-domain migration environment. Among them, the "detection latency" reported in Table 4 refers to the actual wall-clock time, in seconds, that the model experiences during the inference phase, from when the input modal features are loaded to when the emotion classification results are output. The measurement process includes the complete inference process of feature preprocessing (e.g., MFCC extraction and image normalization), multimodal coding, Transformer interaction modeling, and final classification prediction.

Table 3: Multiple indicator test results

Data set	Model	P/%	R/%	F1/%	Specificity/%
IEMOCAP	VGGFace	84.32	83.47	83.89	85.14
	FaceNet	86.45	85.62	86.03	87.28
	EfficientNet	89.58	88.72	89.15	90.34
	Our model	92.84	91.96	92.44	93.12
EMED	VGGFace	85.21	84.36	84.78	86.19
	FaceNet	87.48	86.53	86.99	88.42
	EfficientNet	90.15	89.42	89.78	91.03
	Our model	94.37	93.64	94.21	94.85



	Anger	Detest	Fear	Happy	Amazed	Sad	Contempt
Anger	99	69	36	11	15	35	17
Detest	36	87	19	18	39	43	31
Fear	77	17	69	20	11	16	32
Happy	34	04	14	92	41	40	03
Amazed	54	56	28	27	87	05	35
Sad	42	32	31	26	06	97	32
Contempt	48	09	10	22	36	37	95

(a) VGGFace

	Anger	Detest	Fear	Happy	Amazed	Sad	Contempt
Anger	92	72	22	26	66	64	02
Detest	68	90	44	75	88	27	69
Fear	63	11	72	14	09	18	13
Happy	02	21	82	79	74	04	16
Amazed	04	12	68	73	76	77	59
Sad	19	13	08	07	50	93	54
Contempt	42	28	47	03	53	26	95

(b) FaceNet

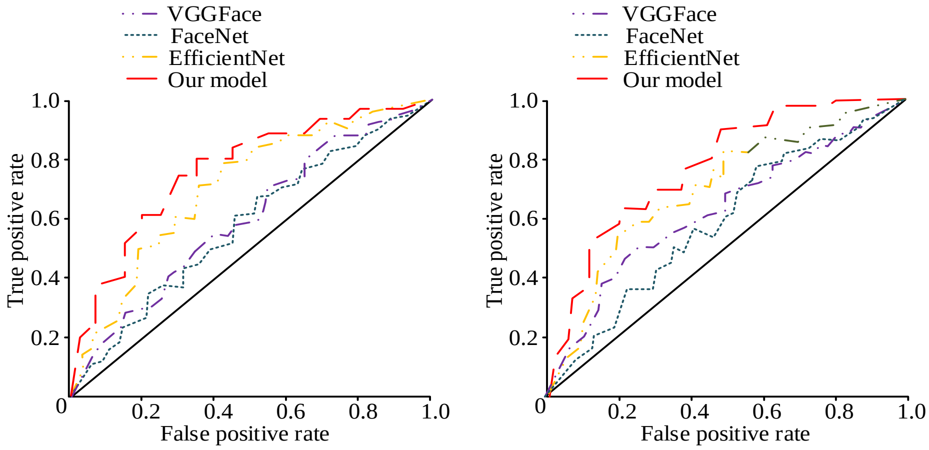
	Anger	Detest	Fear	Happy	Amazed	Sad	Contempt
Anger	87	64	28	63	54	21	35
Detest	73	84	26	77	79	32	37
Fear	26	39	89	24	49	57	65
Happy	71	42	72	81	25	27	59
Amazed	37	73	39	22	65	68	56
Sad	63	65	44	26	44	79	22
Contempt	52	12	29	22	79	48	90

(c) EfficientNet

	Anger	Detest	Fear	Happy	Amazed	Sad	Contempt
Anger	99	74	66	69	39	18	23
Detest	75	95	09	64	72	44	53
Fear	11	36	93	14	33	38	73
Happy	72	77	68	92	18	27	26
Amazed	47	75	44	08	95	79	42
Sad	69	61	33	17	36	97	16
Contempt	41	21	24	18	68	75	96

(d) Our model

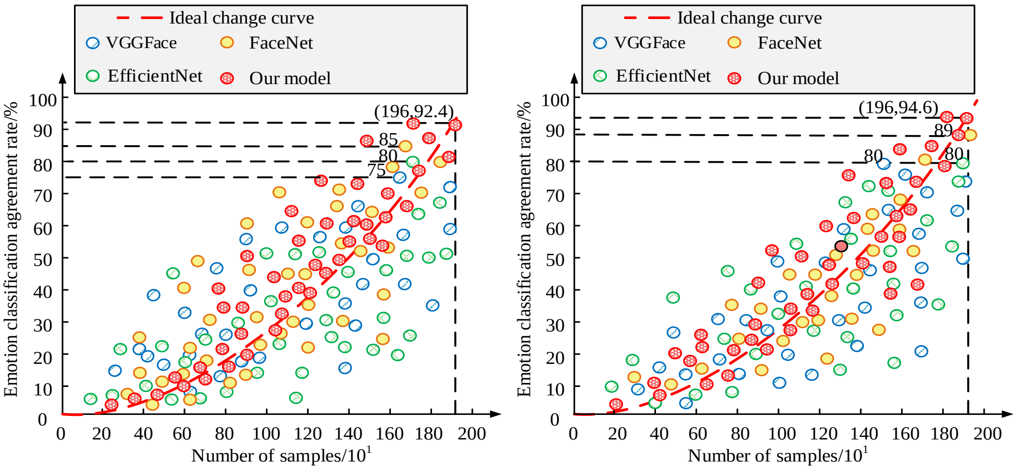
Figure 10: Confusion matrix results for seven types of facial emotion recognition based on video modality in IEMOCAP dataset by different models



(a) Anger

(b) Sadness

Figure 11: Statistical results of AUC indicators



(a) IEMOCAP

(b) EMED

Figure 12: Results of emotion classification consistency test for different models

Table 4: Test results of emotion recognition index

Data set	Model	Emotional intensity/%	Detection delay/s	mAP/%	Inference cost/MB
IEMOCAP	VGGFace	82.45	0.72	85.32	73
	FaceNet	85.62	0.68	88.45	88
	EfficientNet	89.14	0.62	91.78	91
	Our model	93.67	0.55	95.24	94
EMED	VGGFace	80.38	0.75	84.19	73
	FaceNet	84.21	0.69	87.56	88
	EfficientNet	88.92	0.63	91.03	91
	Our model	94.13	0.56	96.12	94

In Table 4, on the IEMOCAP dataset, the study's proposed model performs optimally in terms of sentiment strength, detection latency, and mAP. Its computational inference cost is 94 MB per sample, which is an acceptable overhead for guaranteed accuracy. In terms of detection latency, the model outperforms VGGFace's 0.72 seconds and EfficientNet's 0.62 seconds with a minimum response time of 0.55 s. This reflects a clear real-time advantage. For the mAP, the model reaches 95.24%, which is an improvement of 6.79% and 3.46% compared to FaceNet and EfficientNet, respectively. In the EMED dataset, the model achieves a sentiment strength of 94.13% and an mAP of 96.12%. The detection latency is 0.56 seconds, maintaining the model's leading position and demonstrating its ability to adapt to varying inference loads. Overall, the proposed model improves multimodal fusion accuracy by introducing the dynamic AM while considering inference efficiency and computational cost. This demonstrates good practicality and balance. On the other hand, VGGFace has the weakest overall performance due to its ineffective modeling of complex features between modalities.

4 Discussion

Aiming at the problems of rough modal fusion and insufficient dynamic dependency modeling in multimodal ER for the elderly, the study constructed a fusion model combining the EfficientFace-Transformer structure with an improved AM.

The results of the ablation experiments demonstrated that improved ER performance stemed not only from the efficient feature extraction backbone network but also from the dynamic importance factor and the modal interaction compensation term introduced in the AM layer. The former could dynamically adjust the importance of different modal features, significantly enhancing the model's ability to adapt to heterogeneous cross-modal information. The latter mitigated the problem of modal misalignment by introducing a compensation mapping that made intermodal interaction more adequate. In comparison experiments in which each module was introduced one by one, recognition accuracy (mAP) improved from 91.78% to 95.24%. This verified the improved strategy's contribution to the performance of the multimodal fusion layer.

In terms of computational resources, although the proposed model added an attention enhancement layer, the overall model size was only 94.00 MB and the inference latency was controlled within 0.55 seconds. The model considered both accuracy and inference efficiency. It was suitable for smart home edge device environments where both response time and resource overhead were limited. Meanwhile, the fusion mechanism proposed by the study has better interpretability than the direct splicing structure of the original EfficientFace and Transformer. By analyzing the change of the attention weight matrix, one can track the attention area and interaction mode of the model in different modes, which aids in understanding the process of ER. In terms of validation, the study tested the model on the publicly available multimodal IEMOCAP dataset and introduced the real-life EMED collection scenario dataset for cross-validation. This further enhances the generalizability and stability of the experimental results. A professional psychological annotator labels the EMED dataset based on the simultaneous evaluation of multichannel signals. This approach provides a higher degree of labeling consistency and clinical reliability. In the cross-dataset test, the model's strong robustness verifies its generalization ability. Meanwhile, the combined structure of the original EfficientFace and Transformer was established as a unified baseline. It was then reproduced using the same hardware, parameter configuration, and training strategy to ensure the fairness and reproducibility of the experimental findings. Despite the progress made in the study, it should be pointed out that the model is currently sensitive to high-noise modes, and the modal credibility estimation mechanism and adaptive gating fusion module can be further introduced in the future to improve the dynamic stability under multi-source imbalance conditions. At the same time, it can be extended to a high-dimensional multimodal ER framework that includes more channels such as text and physiological signals, in order to enhance its usefulness and breadth in intelligent interaction systems for the elderly.

5 Conclusion

For the multimodal elderly ER task, an end-to-end recognition model integrating EfficientFace-Transformer structure and improved AM was proposed. It integrated dynamic importance factor, modal interaction compensation term, and CMF enhancement module. Moreover, it strengthened the semantic alignment and feature synergy between audio and video modalities. The experiments were carried out on IEMOCAP and EMED datasets with multiple rounds of testing. The results showed that the proposed model achieved 95.24% in mAP, 93.67% in affective strength, and 0.55 s in detection latency on IEMOCAP. Meanwhile, this model achieved 96.12% in mAP, 94.13% in affective strength, and 0.56 s in latency on EMED, which were all significantly better than VGGFace, FaceNet with EfficientNet, and other comparison methods. To verify the statistical significance of the performance improvement, paired t-tests were used to analyze the detection latency and recognition accuracy. In this case, the p-value of the mAP improvement was less than 0.01, as well as the *p*-value of the difference in detection latency. This indicated that the proposed method had significant advantages in terms of both accuracy and efficiency. Additionally, cross-dataset validation demonstrates the model's robustness and generalization ability. This makes it applicable to the elderly's ER needs in complex environments, such as smart homes.

Statements and declarations

Competing Interests: Author declares no conflict of interest.

Data availability: Data can be obtained from the authors upon reasonable request.

References

[1] Mangano G, Ferrari A, Rafele C, Vezzetti E, Marcolin F. Willingness of sharing facial data for emotion recognition: a case study in the insurance market. AI & SOCIETY, 2024, 39(5): 2373-2384. https://doi.org/10.1007/s00146-023-01690-5

[2] Fei Guan, Zelin Su. Income, income gap and health of the elderly: evidence from China. Advanced Management Science, 2023, 12(1)

[3] Zhu X, Huang Y, Wang X, Wang R. Emotion recognition based on brain-like multimodal hierarchical perception. Multimedia Tools and Applications, 2024, 83(18): 56039-56057. https://doi.org/10.1007/s11042-023-17347-w

[4] Feng G, Wang H, Wang M, Zheng X, Zhang R. A Research on Emotion Recognition of the Elderly Based on Transformer and Physiological Signals. Electronics, 2024, 13(15): 3019-3028. https://doi.org/10.3390/electronics13153019

[5] Sreevidya P, Veni S, Ramana Murthy O V. Elder emotion classification through multimodal fusion of intermediate layers and cross-modal transfer learning. Signal, image and video processing, 2022, 16(5): 1281-1288. https://doi.org/10.1007/s11760-021-02079-x

[6] Park H, Shin Y, Song K, Yun C, Jang D. Facial emotion recognition analysis based on age-biased data. Applied Sciences, 2022, 12(16): 7992-7998. https://doi.org/10.3390/app12167992

[7] Lu J, Liu Y, Lv T, Meng L. An emotional-aware mobile terminal accessibility-assisted recommendation system for the elderly based on haptic recognition. International Journal of Human–Computer Interaction, 2024, 40(22): 7593-7609. https://doi.org/10.1080/10447318.2023.2266793

[8] Du J, Yin J, Chen X, Hassan A, Fu E, Li X. Electroencephalography (EEG)-based neural emotional response to flower arrangements (FAs) on normal elderly (NE) and cognitively impaired elderly (CIE). International Journal of Environmental Research and Public Health, 2022, 19(7): 3971. https://doi.org/10.3390/ijerph19073971

[9] Kelvin L, Anna S. (2023). An Exploratory Study of How Emotion Tone Presented in A Message Influences
Artificial Intelligence (AI) Powered Recommendation System. Journal of Technology & Innovation, 3(2): 80-84.

[10] Zhao Y, Guo M, Sun X, Chen X, Zhao F. Attention-based sensor fusion for emotion recognition from human motion by combining convolutional neural network and weighted kernel support vector machine and using inertial measurement unit signals[J]. IET Signal Processing, 2023, 17(4): e12201. https://doi.org/10.1049/sil2.12201

[11] Saganowski S. Bringing emotion recognition out of the lab into real life: Recent advances in sensors and machine learning. Electronics, 2022, 11(3): 496-503. https://doi.org/10.3390/electronics11030496

[12] Zhou S, Wu X, Jiang F, Huang Q, Huang C. Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks. International Journal of Environmental Research and Public Health, 2023, 20(2): 1400-1431. https://doi.org/10.3390/ijerph20021400

[13] Lin W, Li C. Review of studies on emotion recognition and judgment based on physiological signals. Applied Sciences, 2023, 13(4): 2573-2577. https://doi.org/10.3390/app13042573

[14] Cai Y, Li X, Li J. Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review. Sensors, 2023, 23(5): 2455-2461. https://doi.org/10.3390/s23052455

[15] Tamer Ghareeb B, Tarek F A, Said H A. FER_ML: Facial Emotion Recognition using Machine Learning. Journal of Computing and Communication, 2023, 2(1): 40-49. https://doi.org/10.21608/jocc.2023.282094

[16] Lin W, Li C. Review of studies on emotion recognition and judgment based on physiological signals. Applied Sciences, 2023, 13(4): 2573-2579. https://doi.org/10.3390/app13042573

[17] Zhang T, El Ali A, Wang C. Weakly-supervised learning for fine-grained emotion recognition using physiological signals. IEEE Transactions on Affective Computing, 2022, 14(3): 2304-2322. https://doi.org/10.1109/taffc.2022.3158234

[18] Liu X, Huang C, Zhu H. State-of-the-Art Elderly Service Robot: Environmental Perception, Compliance Control, Intention Recognition, and Research Challenges. IEEE Systems, Man, and Cybernetics Magazine, 2024, 10(1): 2-16. https://doi.org/10.1109/msmc.2023.3238855

[19] Almukadi W. Smart Scarf: An IOT-based Solution for Emotion Recognition. Engineering, Technology & Applied Science Research, 2023, 13(3): 10870-10874. https://doi.org/10.48084/etasr.5952

[20] Kumar, Himanshu, Aruldoss, Martin.Advanced Optimal Cross-Modal Fusion Mechanism for Audio-Video Based Artificial Emotion Recognition.Informatica (Slovenia), 2025, 49(12):61-77. https://doi.org/10.31449/inf.v49i12.7392

[21] Younis E M G, Zaki S M, Kanjo E. Evaluating ensemble learning methods for multi-modal emotion recognition using sensor data fusion. Sensors, 2022, 22(15): 5611-5617. https://doi.org/10.3390/s22155611

[22] Sun, Sulian, Wu, Libo.Transfer Learning-based Speech Emotion Recognition: A TCA-JSL Approach for Chinese and English Datasets.Informatica (Slovenia), 2025, 49(13):221-226. https://doi.org/10.31449/inf.v49i13.7640