Mamba-Enhanced Emotion Analysis TinyML Models for Embedded Devices Deployment

Xing Jin, Shakir Khan*, Mehdi Hosseinzadeh, Neeraj Kumar, Xiyin Wu

***Abstract*—The accuracy of emotion analysis has rapidly improved thanks to the breakthroughs of convolutional neural networks (CNNs) and Transformers. Moreover, the multi-head self-attention (MSA) mechanism of Transformer perfectly fits the modeling of the dependency relationship between expressions and different facial regions. However, CNNs struggle to capture global dependencies, and Transformers' quadratic complexity poses a big challenge to deploy on low-power devices. To resolve these issues, we design a robust and efficient hybrid Tiny machine learning (TinyML) model named HCMTMM for emotion recognition in ultra-low-power embedded devices. Specifically, we propose a hybrid deep model by combining a CNN and Mamba module, which relies on the state space models (SSM) framework, which can effectively exploit the local and global dependencies of different facial regions to enhance emotional recognition performance with linear computational complexity. Moreover, we leverage multi-loss distillation learning to enhance recognition performance. We conducted extensive comparative experiments on four publicly available datasets, and the experimental results showed that when running the family of CNNs, our proposed solution outperforms any other implementation in terms of accuracy and model size. Moreover, we port and test the proposed model on the embedded device ESP32 Cam platform. Our proposed model achieves remarkable results in inference speed.**

***Index Terms*—Tiny machine learning, facial expression recognition, Mamba, distillation learning, ESP32 Cam platform**

I. Introduction

EXPRESSION analysis technology is widely used in people's lives and production, including online education, healthcare, and smart furniture [1]. Psychologist Mehrabian has shown through extensive experimental research that approximately 55% of emotional information in people's daily communication is conveyed through their facial expressions [2] [3]. To systematically analyze and study facial emotions, Ekman [4] first defined seven basic human expressions through extensive scientific research. Subsequently, he proposed the facial action coding system (FACS) with Friesen [5] to explore the close relationship between facial muscle movements and different expressions, and defined action units (AUs) for these different facial muscles. With the rapid development of computers and information technology, human-computer interaction has become more diverse, and relying solely on touch-based mechanical operations is difficult to meet practical needs [6]. Therefore, by accurately recognizing facial expressions through computers, the personalization and intelligence of human-computer interaction can be effectively enhanced [7].

With the development of computer hardware technology, computer vision-based methods provide powerful support for facial emotion analysis, resulting in higher recognition accuracy and faster speed [8]. According to the different types of features used, existing facial expression recognition mainly includes hand-crafted features and deep learning features [9]. The former mainly utilizes existing prior knowledge to extract hand-crafted features from facial regions, and then uses different classifiers for emotion recognition. Common hand-crafted features include geometric features, statistical features, frequency domain features, and motion features [10]. Common classifiers include support vector machines (SVM), random forests, and Bayesian networks. Due to the fast training speed of hand-crafted features, it does not require a large amount of training data and computing resources [11]. However, hand-crafted features are easily affected by human factors and complex scenarios, leading to insufficient training of classifiers and a serious decline in the accuracy of sentiment analysis. To address this issue, deep learning features abandon the shortcomings of hand-crafted feature design and can automatically extract features closely related to tasks from a large number of training samples, including facial emotion analysis [12]. According to the different structures of networks, emotion analysis methods based on deep networks are mainly divided into convolutional neural networks (CNNs), recurrent neural networks, and hybrid networks [13]. CNNs can effectively extract facial texture feature information by utilizing their advantages of parameter sharing, translation invariance, and pooling operations. Jain et al. [14] proposed an effective convolutional network trained on CK+ and JAFFE datasets and achieved higher recognition accuracy than hand-crafted features. Additionally, by adding residual blocks, the performance of the convolutional network is further improved. CNNs mainly use static facial images as input; they are unable to mine temporal features. Among different facial expressions, temporal features can better reflect the movement changes of facial muscles. With the increasing scale of video emotion datasets, the performance of recurrent neural network-

Manuscript received February 2, 2025; revised September 7, 2025. This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) under Grant IMSIU-DDRSP-RP25.(Corresponding author: Shakir Khan.)

Xing Jin is with the College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing 210037, China (e-mail: xingjin@njfu.edu.cn).

Shakir Khan is with the Information Technology Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia (e-mail: sgkhan@imamu.edu.sa).

Mehdi Hosseinzadeh is with the Institute of Research and Development, and School of Engineering & Technology, Duy Tan University, Da Nang, Vietnam (e-mail:mehdihosseinzadeh@duytan.edu.vn).

Neeraj Kumar is with the Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala 147004, India (e-mail:neeraj.kumar@thapar.edu).

Xiyin Wu is with the College of Big Data and Informatics Engineering, Guizhou University, Guiyang 550025, China (e-mail: qqwu3@gzu.edu.cn).

based emotion recognition has been significantly improved. Jung et al. [15] extracted temporal feature information from image sequences and facial landmark sequences, and achieved the highest performance on CK+ and Oulu-CASIA datasets through ensemble learning mechanisms. Li et al. [16] integrated CNN and long short-term memory (LSTM) for spatiotemporal feature mining, further improving the accuracy of sentiment analysis. Compared to hand-crafted features, deep networks can automatically learn input features, achieving higher recognition accuracy and robustness in complex scenes [17].

With the popularity of consumer electronics, various microprocessors have gradually become integrated into daily life due to their advantages, such as low power consumption, low deployment cost, and low latency [18] [19]. However, existing deep networks require large storage and computing resources, making it difficult to effectively deploy on these microprocessors [20] [21]. In recent years, Tiny machine learning (TinyML) technology has been widely studied. This technology can deploy machine learning and deep learning models on microcontrollers and ensure the normal operation of the models with ultra-low power consumption [22]. One of the key directions of TinyML technology is lightweight network design and deployment [23]. To this end, MobileNet reduces computational complexity by 9 times by using depthwise separable convolutional operations. To further enhance the performance, MobileNetV2 introduced residual modules and ReLU activation functions to reduce information loss. MobileNetV3 introduced the squeeze-and-excitation (SE) attention mechanism, further improving the robustness of the model. To further reduce the parameters of the model, Han et al. [24] proposed the SqueezeNet model, which has a model size of less than 0.5MB and a parameter count 510 times smaller than AlexNet. In the facial expression recognition task, the MicroExp network [25] achieved good emotion recognition accuracy with a model parameter of 65K. To achieve a better balance between model accuracy and complexity, as well as model memory size, more and more research is designing more sophisticated network structures to improve performance [26]. Jaderberg et al. [27] proposed a spatial transformer module to extract deep features from regions of interest of the human face. Hu et al. [28] proposed the SE module that focuses on the influence of different channels. Due to the lack of spatial information in the SE module, Woo et al. [29] proposed the convolutional block attention module (CBAM) by utilizing both channel and spatial information. In recent years, Transformers based on multi-head attention mechanisms have shown superior performance in the fields of natural language and computer vision [30]. In the task of facial emotion analysis, the attention mechanism can fully explore the influence of different facial region AUs on emotions. The distribution of different AUs in the face is shown in Fig. 1. Although existing attention mechanisms improve the performance of the model, the overall computational complexity is high, such as the multi-head self-attention (MSA) mechanism requires quadratic complexity [31]. Therefore, from the perspective of mobile deployment, designing more efficient TinyML models has become one of the urgent directions for facial emotion recognition.

Fig. 1. Overview of the distribution of different AUs in the human face. We present some AUs, i.e., AU4, AU12,AU23 and AU24.

In recent years, benefiting from the advantage of state space models (SSM), a new attention mechanism called Mamba [32] has the advantage of linearly varying complexity calculations to achieve the effect of the MSA mechanism, which can improve model performance while also increasing inference speed. To design an efficient TinyML for sentiment analysis, this paper integrates convolutional networks and Mamba to create an efficient and robust deep model named HCMTMM. The entire model has only 66K parameters and can model the local and global dependencies of AUs. The experimental results on four publicly available datasets show that the proposed model achieves higher recognition accuracy with fewer model parameters. The testing results on the EPS32-CAM development board indicate that our proposed model has superiority in terms of model inference speed.

The core innovations of this paper are listed as follows:

(1) We design an efficient TinyML model for sentiment analysis, which uses convolutional operations to extract local features of AU and combines the Mamba module to exploit global dependencies of different AUs. The proposed TinyML model exploits the relationship between AUs and different facial expressions, effectively improving the performance.

(2) We leverage a multi-loss distillation learning mechanism, in which we introduce feature-based distillation loss and response-based loss to enhance the performance. The proposed model only has 66K parameters, which is more lightweight than most existing deep networks.

(3) We carry out extensive comparative experiments on four publicly available datasets. The experimental results indicate that the proposed model achieves optimal performance in terms of accuracy. In addition, by deploying the proposed model on the ESP32-CAM platform, the proposed TinyML model achieves higher inference speed.

The remainder of this paper is organized as follows: we present the related work in Section II; we provide a detailed introduction to the network structure and the Mamba mechanism in Section III; all experiments are conducted and summarized in Section IV; the last section is the conclusion of this paper.

II. Related Work

A. Machine leaning-based emotion recognition

Facial emotion analysis has become one of the key research directions in the fields of computer vision and artificial intelligence due to its significant academic and commercial potential [33] [34]. To extract features from static facial images and videos for sentiment analysis, researchers have proposed many

machine learning algorithms [35]. According to the types of features, the former requires designing hand-crafted features and using different classifiers. From the perspective of hand-crafted feature design, Active appearance models (AAM) use geometric feature information of facial features and facial contours for sentiment analysis. Due to the susceptibility of such algorithms to lighting, angle, and facial size, the accuracy of the model is relatively low [36]. To extract more image texture feature information, Shan et al. [37] used the local binary pattern (LBP) to construct image texture features for sentiment analysis. Although the LBP operator has simple operations and rotation invariance. However, it can lead to excessively high dimensionality of features and increase the complexity of the model. To improve the representation ability of features, the Gabor wavelet transform [38] is used to obtain the frequency domain features of facial images. This type of algorithm can effectively extract the detailed features of images. The above methods mainly extract features from static images, and the mining of temporal features can further improve the accuracy of facial expression recognition. Li et al. [39] extracted motion features based on optical flow from video sequences, which can effectively alleviate the impact of lighting on sentiment analysis. From the perspective of classifier selection, Wang et al. [40] applied the K-nearest neighbor (KNN) algorithm to sentiment classification tasks. However, the classifier has a slow training speed and the training results are unstable. Bayesian network-based classifiers can infer unknown expressions from known expression classification information [41]. In addition, the classifier of SVM [42] continuously optimizes the objective function to obtain the hyperplane with the largest difference between the features of different categories of samples. Traditional machine learning methods do not require a large number of samples for training and have a fast inference speed. However, manual features and existing classifiers are difficult to adapt to the complexity and diversity of facial sentiment analysis.

B. Deep learning-based facial expression recognition

To alleviate the inherent shortcomings of traditional hand-crafted feature design, deep learning based methods can automatically extract features from a large number of samples and perform feature classification simultaneously through supervised learning [43] [44]. In facial expression recognition tasks, two types of deep networks, convolutional neural networks and recurrent neural networks, are mainly used based on whether the extracted object is a static image or a video frame sequence [45]. To further improve the performance of traditional convolutional networks, Mollahosseini et al. [46] adopted two convolutional layers and four Inception layers to broaden the depth and width of the network. Hamester et al. [47] proposed a multi-channel convolutional neural network. This network uses autoencoders to assist standard convolutional neural network learning. To extract richer feature information, Cai et al. [48] proposed a sparse batch normalized convolutional neural network model and used larger convolutional kernels in the convolutional layer. On both the CK+and JAFFE datasets, the accuracy of facial expressions exceeds 95%. To extract discriminative spatiotemporal video features in video sequences, Zhang et al. [49] adopted a spatial CNN and a temporal CNN to extract features from static images and optical flow images. Then all spatial and temporal features are fed into a deep belief network (DBN) for expression recognition. To extract features from video sequences, Pan et al. [50] integrated CNN and LSTM networks for spatiotemporal feature mining. Although deep learning can effectively improve the accuracy and robustness of emotion recognition, it requires a large number of samples for network training [51]. In addition, the structure of the network is more complex, leading to an increase in the number of parameters and memory requirements of the model [52]. To deploy deep models on embedded devices, a lightweight model design is urgently needed. Lightweight network models such as ShuffleNet, MobileNet, GhostNet, etc. reduce the number of parameters and complexity of the model while maintaining accuracy as much as possible, and improve the training speed [53]. To build a real-time facial emotion recognition system, Li et al. [54] designed a lightweight CNN by exploiting pre-activation in the residual block. In this paper, we aim to design a very lightweight deep model which can be deployed on the ESP32-CAM platform for emotion analysis.

C. Attention mechanism in computer vision

Attention mechanism refers to the ability of a model to focus on important parts related to the task and ignore other irrelevant parts [55]. In computer vision tasks, attention mechanisms can help models better understand input images or videos by assigning different weights to features in different regions, thereby improving model performance [56]. At present, attention mechanisms mainly include spatial attention mechanisms, channel attention mechanisms, temporal attention mechanisms, and mixed attention mechanisms [57]. The spatial attention mechanism mines the importance of different positions in the input data and weights the input [58]. Jaderberg et al. [27] proposed the spatial Transformer networks, which complete preprocessing operations suitable for the task by learning the deformation of the input, and have made significant progress in image classification tasks. Zhao et al. [59] combined the bilateral U-Net network model with a spatial attention mechanism to increase the receptive field and enhance the context information fusion. The channel attention mechanism weights different channels of input data to improve the performance. Hu et al. [28] proposed the squeeze-and-excitation networks (SENet) in which each layer of a convolutional network has many convolutional kernels corresponding to a feature channel. The Squeeze operation takes the global spatial features of each channel as its representation, while the excitation operation learns the degree of dependence of each channel and adjusts different feature maps based on the degree of dependence. The temporal attention mechanism weights time series data to improve the model's understanding of time series data. Xia et al. [60] proposed a new temporal attention mechanism to learn the contribution of different historical appearance features at the same position to the current features. Due to the modeling capability of a single attention mechanism, an increasing number of studies involve hybrid attention

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. 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/TCE.2025.3614990

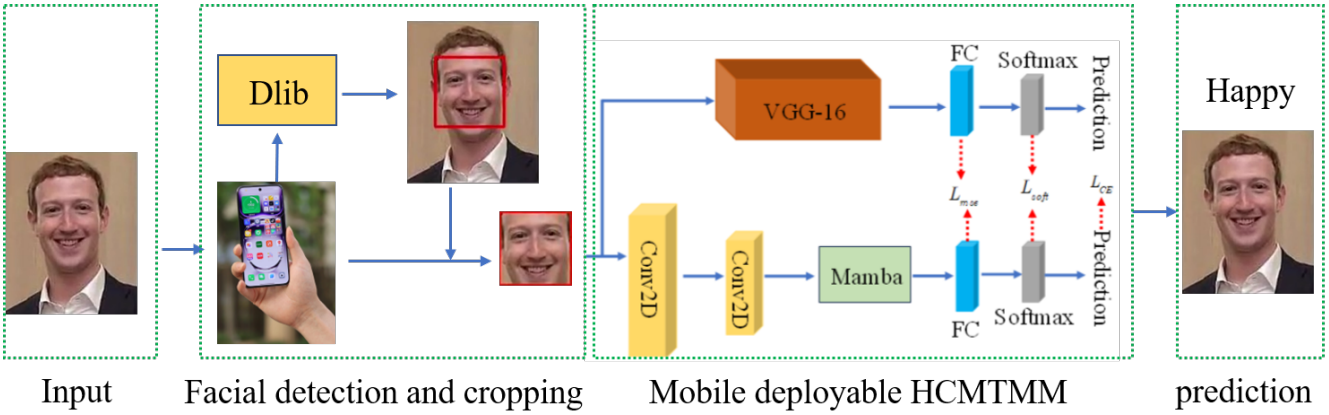

Fig. 2. Overview of emotion recognition system based on consumer electronics products. Electronic products are used for face detection and cropping through Dlib tools. We use the pre-trained VGG-16 as teacher network. In the student network, which only has 66K parameters, we use two convolution layers for deep feature extraction. To exploit the global dependencies of the AUs in different facial regions, we embed the Mamba module to construct an efficient global attention mechanism. Moreover, we introduce multi-loss distillation learning mechanism to enhance the performance.

mechanisms to further improve the performance of models. Huang et al. [61] designed a novel channel-spatial attention mechanism by fusing multi-stage features for object detection. Li et al. [62] used the spatial–temporal attention mechanism for remote sensing image change detection. In recent years, the MSA mechanism has been widely applied in the field of computer vision [63]. The MSA mechanism captures the global dependencies of a sequence through parallel operations , thereby obtaining richer semantic information and improving model performance. However, the MSA mechanism leads to quadratic computational complexity, making it difficult to deploy on mobile or embedded devices [64]. In emotion analysis, different muscle movements generate different expressions. This paper designs a TinyML model by introducing a novel attention mechanism to focus on different facial regions.

III. Our proposed HCMTMM framework

A. Motivation

Deep learning-based model plays an essential role in emotion analysis. However, due to high computation and memory footprints, enabling deep model deployment on embedded devices is becoming a major challenge. TinyML models have achieved remarkable performance to facilitate the operation of deep models on embedded devices. Toward this end, we design an effective TinyML model that benefits from the attention mechanism and the FACS system. More importantly, it integrates the Mamba module to explicitly leverage global dependencies of facial AUs, allowing the model to improve performance with linear time complexity compared to the MSA mechanism. The proposed model HCMTMM is shown in Fig. 2. In the proposed HCMTMM, we adopt two convolution layers to extract deep features of different facial regions. By introducing the Mamba module, we aim to use a more effective attention mechanism to enhance the performance of facial expression recognition. In addition, we introduce a multi-loss distillation learning mechanism to further improve the performance of the student network.

B. Facial emotion analysis network based on Mamba attention mechanism

In the FACS system, the changes in facial expressions are closely related to facial action units. To further describe the motion changes of facial muscles corresponding to different action units, we use residual connection convolution operation to extract local features of different regions of the face. Due to the advantages of parameter sharing, local receptive field, and low operational complexity in convolution operations, it has been widely used for extracting local features from images. Convolution operation is used to extract local features through sliding windows and achieve parameter sharing, effectively reducing the complexity of the model. In addition, to further improve the feature representation ability of the model, we introduce a residual connection mechanism. We use two 3×3 convolution operations and max pooling operations to obtain the deep features of local facial regions. The correspondence between different facial regions and deep feature is shown in Fig. 3.

Through convolution and pooling operations, we can obtain deep features of different facial regions. Inspired by FACS research, the introduction of the attention mechanism can further improve the recognition performance. In facial expression recognition, changes in different expressions may be closely related to multiple facial action units. In addition, modeling the dependency relationships of these action units is quite complex. Compared with commonly used attention mechanisms, i.e., channel attention mechanism, spatial attention mechanism, and hybrid attention mechanism, the MSA mechanism can achieve better performance in computer vision tasks. The MSA mechanism can achieve modeling of long-distance dependencies and parallelize the entire process, improving the efficiency of data processing. The calculation process of the entire MSA mechanism is as follows: Firstly, three projected matrices Query Q, Key K, and Value V are calculated as follows:

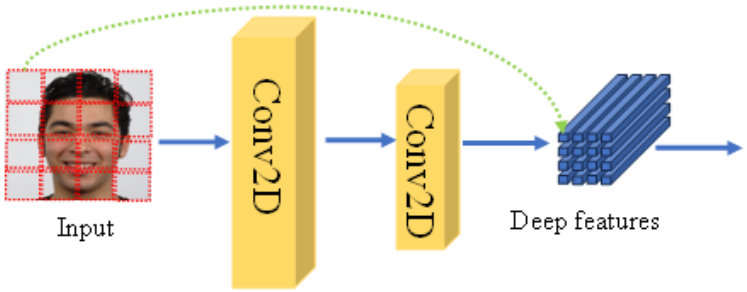

Fig. 3. Overview of correspondence between different facial regions and deep features. The facial image is processed through two layers of convolution and pooling operations to obtain the deep features of the face. These deep features have a clear correspondence with different regions of the face. According to the FACS system, these deep features are also closely related to the AUs of the human face.

$$\begin{aligned} Q &= XW^Q \\ K &= XW^K \\ V &= XW^V \end{aligned} \tag{1}$$

where X denotes the input features, W^Q, W^K, W^V denote three learnable matrices.

To construct the MSA mechanism, the one-head self-attention is defined as follows:

$$H(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{2}$$

where d is the scale factor, which is used to adjust the numerical value of the product.

The MSA mechanism can be defined as follows:

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head_1, head_2, ..., head_h)W_M \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{3}$$

where W_M, W_i^Q, W_i^K, W_i^V denote three learnable matrices.

The multi-head structure has improved the representation ability and flexibility of the model, but it also increases the computational complexity, especially when dealing with large amounts of data, which may lead to an increase in the demand for computing resources. In addition, the MSA mechanism has a large number of parameters and requires a large number of samples for model training. Moreover, the MSA mechanism needs quadratic computational complexity with the input size, making it difficult to handle very long sequence data. To alleviate these issues, a novel deep learning architecture called the Mamba model is designed based on the selective state space model and Mamba can achieve linear time complexity when processing long sequences [65]. Compared with the MSA mechanism, the Mamba is more efficient in processing long sequences. In addition, the Mamba model utilizes selective scanning techniques and a hardware-optimized design, effectively reducing memory overhead and further improving the efficiency of model training and inference.

Mamba's inspiration comes from traditional state space models, including discretization and convolution operations of the state space. Firstly, given the input $x(t)$, the output of the traditional definition of state space is as follows:

$$\begin{aligned} h^{'}(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t) + Dx(t), \end{aligned} \tag{4}$$

where $h(t)$ denotes the current state, t denotes time steps, A denotes the the state transition matrix, B denotes the impact of control variables on state variables, $y(t)$ denotes the output of system, C denotes the impact of the current state variable on the output, and D denotes the impact of the current control variable on the output.

As the traditional state space calculations involve continuous values, to further improve computational efficiency, it is necessary to discretize the state space model using the zero-order preservation method. Assuming that the solution of the state space ordinary differential equation is constant within the sampling period. The formula for calculating the solution of the differential equation is as follows:

$$h(t) = e^{A(t_{k+1}-t_k)}h(t) + \int_{t_k}^{t_{k+1}} e^{A(t-\tau)}Bx(\tau)d\tau, \tag{5}$$

where τ denotes integral variables, t_k and t_{k+1} denote starting and ending points of the sampling time interval.

The state space equation after discretization is as follows:

$$\begin{aligned} h_t &= \hat{A}h_{t-1} + \hat{B}x_t, \\ y_t &= Ch_t + Dx(t), \end{aligned} \tag{6}$$

where $\hat{A} = e^{\Delta A}$, $\hat{B} = (\Delta A)^{-1}(e^{\Delta A} - I) \cdot \Delta B$, Δ denotes the estimating learnable parameters for discrete intervals.

The convolution operations of the state space model are defined as follows:

$$y = x * \hat{K}, \tag{7}$$

where $\hat{K} = (C\hat{B}, C\hat{A}\hat{B}, ..., C\hat{A}^{M-1}\hat{B})$, M denotes the size of the convolution kernel.

The linear time-invariant structural state space models can not effectively exploit contextual information. Therefore, the selective state space model named Mamba which introduces input-related parameters to a time-varying system, can enhance the ability to model complex inputs. Two additional linear layers are added for matrices B and C to select the control and state variables of the input, enhancing the model's adaptability

to different input forms. The calculation process is defined as follows:

$$\begin{aligned} \hat{B} &= s_B(x), \\ \hat{C} &= s_C(x), \\ \Delta &= s_\Delta(x), \end{aligned} \tag{8}$$

where $s_B(\cdot)$ and $s_D(\cdot)$ denote two linear layers.

Due to the introduction of filtering mechanisms, the convolution window of the state space has changed, making it difficult to effectively perform convolution sequence operations. To speed up the computation, Mamba adopts multi-threading for parallel computing, using the combination law for out-of-order calculation for each sequence, and finally obtaining the result through accumulation. In order to simplify the model design, the Mamba block simplified the structure of Hungry Hungry Hippos (H3) and combined it with a gated multilayer perceptron (MLP), adding residual terms to prevent gradient vanishing. The entire model result is shown in Fig. 4.

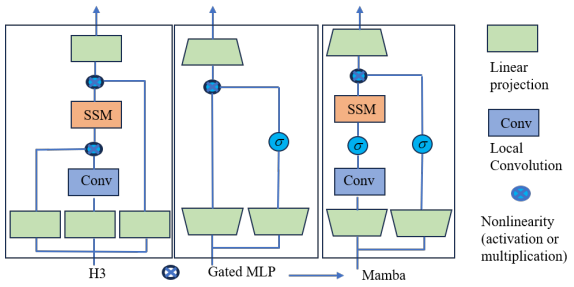

Fig. 4. Overview of the Mamba framework.

Compared with Transformer models based on the MSA mechanism, Mamba can achieve linear time complexity when processing long sequences. In addition, Mamba utilizes a selective state space model, which can be more efficient for modeling long sequences. Mamba uses parallel algorithms optimized for modern hardware, especially GPUs, to reduce memory requirements and improve computational efficiency. Finally, Mamba's structure is more concise, removing traditional attention and MLP blocks, providing better scalability and performance. To build an efficient TinyML model for sentiment analysis, we combined traditional convolution operations with the Mamba model. The Mamba model can achieve modeling of global dependencies in different regions of the human face with fewer model parameters and computational complexity.

C. Facial emotion analysis based on knowledge distillation mechanism

In the field of machine learning, complex models often come with training and inference costs, increasing the difficulty of deploying embedded devices for the models. To address this issue, the distillation learning mechanism effectively enhances the performance of student models by transferring knowledge from complex models (teacher models) to simple models (student models). The teacher model is usually a large and complex model with high accuracy and generalization performance. The student model, on the other hand, is a smaller and simpler model with lower computational and storage costs. To this end, we further improve the performance of the TinyML model proposed in this paper by introducing a multi-loss distillation learning mechanism. Given a training sample x_i and its corresponding label y_i. After obtaining the feature representations of the samples x through teacher and student models, the feature-based distillation loss is defined as follows:

$$L_{mse} = \frac{1}{N}\sum_{i=1}^{N} || \frac{f_i^s}{||f_i^s||} - \frac{f_i^t}{||f_i^t||} ||_2^2, \tag{9}$$

where N denotes the batch size, f_i^s and f_i^t denote the output of the last fully connected layers of the teacher and student models.

To further improve the distillation efficiency, we introduce response-based loss. Given sample x_i, the outputs of the softmax layer from the teacher and student models are u_i and v_i. The response-based loss is calculated as follows:

$$L_{soft} = KL(P_T, Q_T), \tag{10}$$

where $P_T = \frac{exp(u_i/T)}{\sum_{i=1}^{K} exp(u_i/T)}$, $Q_T = \frac{exp(q_i/T)}{\sum_{i=1}^{K} exp(q_i/T)}$, K denotes number of categories of facial expressions,T denotes the temperature parameter, $KL(\cdot)$ denotes the KL divergence.

The multi-loss of distillation learning is defined as follows:

$$L_{all} = L_{CE} + \alpha L_{soft} + \beta L_{mse}, \tag{11}$$

where L_{CE} denotes the cross entropy loss function, α and β denote hyper-parameters.

In summary, we introduce the Mamba structure to achieve global modeling of facial action units with less model complexity, effectively improving the performance of sentiment analysis. In addition, by introducing the multi-loss distillation learning mechanism, we constructed a more compressed and efficient TinyML model.

IV. Experiments and Analysis

In this section, we conducted extensive comparative experiments on four commonly used facial emotion datasets, RaFD [66], CK+ [67], FER2013 [68] and RAF-DB [69]. Moreover, we conducted comparative experiments to verify the effectiveness of distillation learning.

A. Datasets and experimental settings

The RaFD dataset is a high-quality facial emotion recognition dataset collected in a laboratory environment. This data collected facial images of 67 people in different poses, and three people's eye gaze in three directions. The categories of emotions include anger, disgust, fear, happiness, sadness, surprise and neutral. In our experiment, we selected 1407 frontal facial images. The CK+ dataset collects facial video sequences of 123 individuals in different emotional states. The total number of video sequences is 593, but only 327 video sequences have corresponding emotional labels added. CK+ has the same number of sentiment labels as the RaFD dataset. In our experiment, we extracted static facial images for

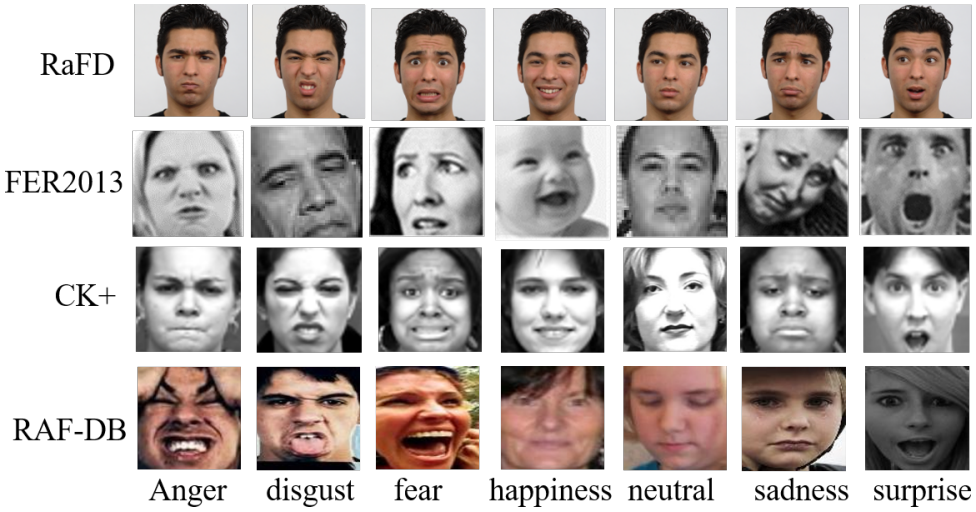

Fig. 5. Some samples from four facial expression datasets.

emotion analysis and used the last three frames from a labeled video sequence. In order to validate emotion recognition in complex scenarios, we chose the FER-2013 facial emotion dataset. This dataset, like the previous two datasets, has 7 basic facial emotion labels. The entire dataset consists of 35887 images, including 28708, 3589 and 3589 images in the training, validation, and testing sets. Due to the image resolution being only 48×48 and the presence of label noise and annotation errors, the human recognition rate is only about 70%. RAF-DB is a high-quality and challenging real-world emotion dataset that provides crucial resources for developing facial expression recognition systems capable of handling complex real-world scenarios. These facial emotion pictures all come from the Internet and contain various complex factors in the real world, such as lighting changes, head posture, and occlusion. This dataset contains approximately 30000 facial images in total. In our experiment, 12271 images were used for model training, and 3068 images were used for model testing. Some samples from four datasets are shown in Fig. 5.

We use the Dlib tool to crop facial regions and scale facial regions to 98×98. The other experimental settings on the four datasets are the same as the references [70] and [25]. For the distillation learning setting, we employed the VGG-16 as the teacher model. The hyperparameters α and β in distillation learning with multiple losses are set to 0.5. We made full use of existing research and combined relevant experiments to select the entire hyperparameter setting. In existing research [71] [72] [73], when the combination of cross entropy loss and KD is used for distillation learning, the hyperparameters of KD loss are usually set to 0 to 1. Therefore, we first obtain the hyperparameters of α on the RaFD dataset. Then fix the value of α and obtain the hyperparameters of β . The values of α and β are both set as 0.5 in our experiments. For a fair comparison, all comparative methods were run on an NVIDIA Titan RTX GPU using the PyTorch framework and Python 3.8 based on Anaconda virtual environment. During model training, we adopted the Adam optimizer with a learning rate of 5e-4. Moreover, we compare the inference speed on the ESP32-CAM embedded device. The ESP32-CAM module is equipped with an ESP32-S chip, an ultra-small OV2640 camera, 520 KB of SRAM, and a micro SD card slot. The ESP32-CAM module can be widely used in various Internet of Things (IoT) applications. ESP32-CAM is shown in Fig. 6.

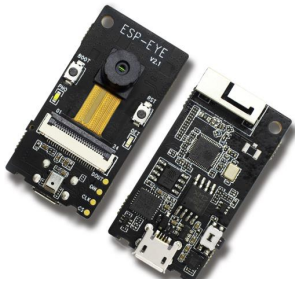

Fig. 6. Overview of ESP32-CAM embedded device.

B. Results on RaFD dataset

To verify the effectiveness of our proposed model, we selected different lightweight models and compared the accuracy and model parameters of different models on the RaFD dataset. We detail the results in Table I.

Table I displays the comparison results of different deep models in terms of accuracy and model parameters. Although some deeper models. i.e., compactCNN,ResMoNET and PeleeNet, achieve higher recognition accuracy, but these models have more model parameters. The compactCNN achieves a recognition rate of 96.83% with 20M model parameters. From the perspective of mobile deployment, the design of lightweight models has significant application value. Some lightweight models also achieve competitive recognition performance, i.e., DDRGCN achieves an overall accuracy of 94.48% with 110K parameters. In this paper, we combine the CNN and Mamba module for emotion analysis. The Baseline only has two convolution layers and two fully connected layers. The Baseline obtains 91.20% accuracy with 64K parame-

TABLE I
ACCURACY AND PARAMETERS OF DIFFERENT MODELS ON THE RAFD DATASET

Models	Parameters	Accuracy (%)
PeleeNet [74]	2123K	84.00
ResMoNET [75]	1721K	90.00
DDRGCN [70]	110K	94.48
MicroExpNet [25]	65K	90.80
compactCNN [66]	20M	96.83
TESGCN [76]	80K	94.64
MobileViT [77]	972K	92.30
SNNEs [78]	-	90.00
Baseline	64K	91.20
Baseline+MSA	69K	94.28
Baseline+Mamba	66K	95.25
HCMTMM	66K	96.08

ters. By introducing the Mamba module, the Baseline+Mamba exhibits improvements of 4.05% compared to Baseline. Moreover, HCMTMM has a significant 0.83% increase in accuracy compared to Baseline+Mamba by introducing the distillation learning mechanism.

C. Results of different models on CK+ dataset

We evaluate our proposed model and other deep models on the CK+ dataset and the results are reported in Table II.

TABLE II
ACCURACY OF DIFFERENT LIGHTWEIGHT MODELS ON CK+ DATASET

Models	Parameters	Accuracy (%)
EmotionNet Nano-A [79]	232K	97.60
RS-Xception [80]	1.92M	97.13
DDRGCN [70]	110K	94.32
MicroExpNet [25]	65K	90.71
ST-BLN wo/MCD [81]	132K	93.19
TESGCN [76]	80K	94.64
LITE-FER [82]	113K	89.69
LCNN [83]	1331K	98.00
SLW-LDT [84]	69K	98.90
Baseline	64K	90.84
Baseline+MSA	69K	93.87
Baseline+Mamba	66K	94.47
HCMTMM	66K	95.16

In Table II, RS-Xception and WRN obtain higher performance by exploiting complex models and spatiotemporal features. However, on resource-constrained devices, the deployment of these models presents significant challenges. Fortunately, various lightweight networks have also achieved good performance in facial sentiment analysis tasks. The EmotionNet Nano-A achieves 97.60% accuracy with 232K parameters. In this paper, we design a more lightweight model HCMTMM by integrating CNN and Mamba module with 66K parameters. The Baseline+Mamba obtains 95.06% accuracy, in which the Mamba can effectively leverage the relationship between facial movements and expressions from a global perspective. Moreover, the distillation learning mechanism in HCMTMM demonstrated a 0.69% improvement in accuracy. This demonstrates the effectiveness of the Mamba module and distillation learning in facial expression recognition tasks.

D. Results on FER2013 dataset

As shown in Table III, we further compare with HCMTMM and other deep models on the FER2013 dataset.

TABLE III
ACCURACY OF DIFFERENT LIGHTWEIGHT MODELS ON FER2013 DATASET

Models	Parameters	Accuracy (%)
DCN [46]	7300K	69.30
WRN [67]	18.35M	69.60
Xception [85]	1669K	67.01
DNNRL [86]	2.6M	71.33
RS-Xception [80]	1.92M	69.02
Li-CNN [87]	1387K	70.00
LANMSFF [88]	358K	70.44
MobileViT [77]	972K	62.20
Baseline	64K	64.23
Baseline+Mamba	66K	69.25
HCMTMM	66K	70.18

Although numerous experimental results have shown that deeper networks can achieve better performance in many tasks. However, in specific tasks, even with the use of a well-designed network structure, surprising results can be achieved. In Table III, the accuracy value of model LANMSFF exceeds that of model RS-Xception with fewer model parameters. This efficient lightweight model provides convenience for mobile deployment. But we propose an effective TinyML model by embedding the Mamba module in an existing CNN framework. The Baseline+Mamba shows improvement of 5.02% by introducing the attention mechanism. Moreover, our proposed HCMTMM model can utilize distillation learning to enhance facial expression recognition performance.

E. Results on RAF-DB dataset

To verify the emotion recognition performance of the proposed model in more complex scenarios, we compared the accuracy and model parameters of different models on the RAFDB dataset. All the results are listed in Table IV.

TABLE IV
ACCURACY AND PARAMETERS OF DIFFERENT MODELS ON THE RAF-DB DATASET

Models	Parameters	Accuracy (%)
CDLM-frame	2673K	65.52
MobileViT-XXS	1298K	73.89
L3Net	102K	55.03
DDRGCN [70]	110K	58.25
MicroExpNet [25]	65K	45.31
LWFER	770K	86.92
AFNet	26480K	80.30
Baseline	64K	52.26
Baseline+MSA	69K	59.14
Baseline+Mamba	66K	61.58
HCMTMM	66K	63.96

As shown in Table IV, we give the comparison results of different deep models from the perspective of accuracy and model parameters. Some deeper models. i.e., MobileViT-XXS and AFNet, still achieve high emotion recognition performance even in complex scenarios. Compared with the Baseline, the introduction of MSA and Mamba models has a positive

effect on performance improvement. Finally, by introducing a distillation learning mechanism, we further improved the performance of the model.

F. Effectiveness of distillation learning on four datasets

In our proposed model, we adopt distillation learning to further elevate model performance. In this section, we conduct comparative experiments to estimate the effectiveness of distillation learning. All results on four facial expression datasets are shown in Fig. 7.

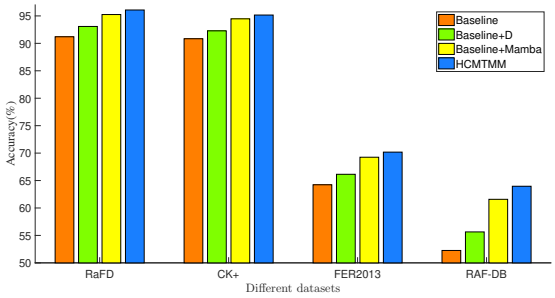

Fig. 7. Effectiveness of distillation learning on four datasets.

The models we compared include Baseline, Baseline+D, Baseline+Mamba, and HCMTMM. Baseline and Baseline+D use network structures within the system, while Baseline+D introduces a distillation learning mechanism. HCMTMM introduces the distillation learning mechanism based on Baseline+Mamba. From the results in Fig. 7, distillation learning can further improve the recognition performance.

G. Effectiveness of different attention mechanism

This paper constructs a global attention mechanism by embedding the Mamba module for modeling the relationship between action units in different facial regions, effectively improving the ability of sentiment analysis. Compared with the MSA mechanism, the Mamba model can model global dependencies with less complexity. Therefore, this experiment further validates the effectiveness of our model by comparing the effects of different attention mechanisms. All the results are shown in Fig. 8.

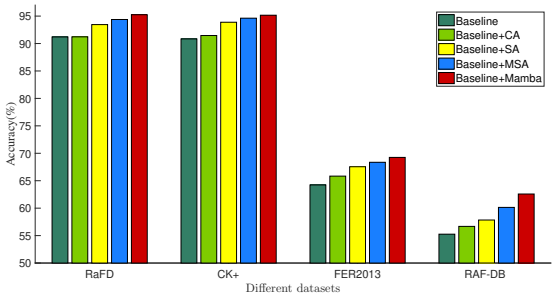

Fig. 8. Accuracy of different attention mechanisms on four datasets.

H. Model size analysis

To facilitate the deployment of deep models on resource-constrained mobile devices, the memory footprint of the model is an important metric. In this section, we calculate the memory footprint of different lightweight models. The memory footprint includes the memory requirements of the input, forward and backward processes, and model parameters. All statistical results are shown in Table V.

TABLE V
SIZE OF DIFFERENT MODELS

Model	Parameters	Size (MB)
MicroExpNet [25]	65K	0.93
DDRGCN [70]	110K	0.48
Baseline	64K	0.92
HCMTMM	66K	0.95

As the results shown in Table V, the MicroExpNet needs a 0.93MB memory footprint. Compared to the MicroExpNet, the model size of HCMTMM increases by 0.02M. However, in terms of the accuracy of emotion recognition, HCMTMM has significantly better accuracy compared to MicroExpNet.

I. Inference speed of different models on ESP32-CAM board

To verify the testing performance of our proposed HCMTMM on mobile devices, we used the commonly used embedded development board ESP32-CAM. We compared the derivation speed of several lightweight models. All experimental results are shown in Table VI.

TABLE VI
SPEED OF DIFFERENT MODELS ON ESP32-CAM

Model	Speed (ms)
MicroExpNet [25]	85
Baseline	88
Baseline+MSA	101
HCMTMM	94

Table VI reports the inference speed of different models. Although MicroExpNet achieves better performance compared to HCMTMM, the difference in speed between the two models is not significant. In terms of overall performance in terms of model size, accuracy, and speed, our proposed model is more suitable for mobile deployment.

V. CONCLUSION

The issues of limited computational resources and complex dependence relationships of AUs have long been focal points in the field of facial expression recognition. To address these issues, we design an effective multi-loss distillation-based TinyML model named HCMTMM by combining CNN and Mamba modules. Compared to MSA, the Mamba module can implement global dependency modeling with linear computational complexity. Moreover, we construct a multi-loss distillation learning to further enhance expression recognition performance. The proposed model only has 66K parameters. Comparative experiments on four publicly available datasets

demonstrate that our proposed model outperforms embedded deployment in terms of both accuracy and model size. In addition, the porting and testing results on ESP32-CAM show that the inference speed of our proposed model is also more competitive. Our work is the first exploration of the hybrid CNN-Mamba architecture in the real-time emotion analysis task, and we also hope to bring a new direction in this field. Due to the significant advantages in inference speed and memory requirements of binary networks, the impact of quantization errors can lead to a decrease in model performance, such as recognition accuracy. In our future work, we will investigate how to design high-performance binary networks for sentiment recognition in mobile devices and further improve our method's performance in more complex scenarios.

REFERENCES

[1] M. Khan, J. Ahmad, W. Gueaieb, G. De Masi, F. Karray, and A. El Saddik, "Joint multi-scale multimodal transformer for emotion using consumer devices," *IEEE Transactions on Consumer Electronics*, 2025.

[2] W.-Y. Hsu and T.-H. Chiang, "Triple-attribute perceptron facial expression recognition in real-world environments," *IEEE Transactions on Consumer Electronics*, vol. 7, no. 1, pp. 608–620, 2025.

[3] Y. Liu, K. H. Cheng, M. Savic, H. Chen, Z. Yu, and G. Zhao, "3d face de-identification with preserving multi-facial attributes: A benchmark," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025.

[4] M. A. Solis-Arrazola, R. E. Sanchez-Yañez, C. H. Garcia-Capulin, and H. Rostro-Gonzalez, "Enhancing image-based facial expression recognition through muscle activation-based facial feature extraction," *Computer Vision and Image Understanding*, vol. 240, p. 103927, 2024.

[5] H. Liu, Q. Zhou, C. Zhang, J. Zhu, T. Liu, Z. Zhang, and Y.-F. Li, "Mmatrans: Muscle movement aware representation learning for facial expression recognition via transformers," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 12, pp. 13 753–13 764, 2024.

[6] X. Wang, B. Yi, B. F. Felemban, A. A. Aly, W. Li, and J. Liu, "Sentiment analysis via trustworthy label enhancement for consumer electronics applications," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 1, pp. 1935–1944, 2025.

[7] X. Ji, Z. Dong, Y. Han, C. S. Lai, G. Zhou, and D. Qi, "Emsn: An energy-efficient memristive sequencer network for human emotion classification in mental health monitoring," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 4, pp. 1005–1016, 2023.

[8] A. A. Albraikan, J. S. Alzahrani, R. Alshahrani, A. Yafoz, R. Alsini, A. M. Hilal, A. Alkhayyat, and D. Gupta, "Intelligent facial expression recognition and classification using optimal deep transfer learning model," *Image and Vision Computing*, vol. 128, p. 104583, 2022.

[9] J. Noor, M. Daud, R. Rashid, H. Mir, S. Nazir, and S. A. Velastin, "Facial expression recognition using hand-crafted features and supervised feature encoding," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2020, pp. 1–5.

[10] G. V. Reddy, C. D. Savarni, and S. Mukherjee, "Facial expression recognition in the wild, by fusion of deep learnt and hand-crafted features," *Cognitive Systems Research*, vol. 62, pp. 23–34, 2020.

[11] W. Xie, L. Shen, and J. Duan, "Adaptive weighting of handcrafted feature losses for facial expression recognition," *IEEE transactions on cybernetics*, vol. 51, no. 5, pp. 2787–2800, 2019.

[12] G. I. Tutuianu, Y. Liu, A. Alamäki, and J. Kauttonen, "Benchmarking deep facial expression recognition: An extensive protocol with balanced dataset in the wild," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108983, 2024.

[13] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning," *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106621, 2022.

[14] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognition Letters*, vol. 120, pp. 69–74, 2019.

[15] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2983–2991.

[16] T.-H. S. Li, P.-H. Kuo, T.-N. Tsai, and P.-C. Luan, "Cnn and lstm based facial expression analysis model for a humanoid robot," *IEEE Access*, vol. 7, pp. 93 998–94 011, 2019.

[17] Y. Zhang, Y. Li, X. Liu, W. Deng *et al.*, "Leave no stone unturned: mine extra knowledge for imbalanced facial expression recognition," in *Advances in Neural Information Processing Systems*, 2024, pp. 1–13.

[18] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou, "Rethinking the learning paradigm for dynamic facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 958–17 968.

[19] Y. Liu, H. Du, D. Niyato, J. Kang, Z. Xiong, S. Mao, P. Zhang, and X. Shen, "Cross-modal generative semantic communications for mobile aigc: Joint semantic encoding and prompt engineering," *IEEE Transactions on Mobile Computing*, 2024.

[20] N. Bellarmino, R. Cantoro, M. Huch, T. Kilian, U. Schlichtmann, and G. Squillero, "Deep learning strategies for labeling and accuracy optimization in microcontroller performance screening," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024.

[21] W. Chen, X. Yu, F. Zhu, X. Chen, K. Zhang, C. Yu, H. Zhao, and A. V. Vasilakos, "Joint communication and control optimization of a multi-vehicle platooning system," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[22] J. Jiang, B. Wang, Q. Tang, G. Zhong, X. Tang, and J. J. Rodrigues, "Incremental semi-supervised learning for data streams classification in internet of things," *IEEE Transactions on Network and Service Management*, vol. 22, no. 3, pp. 2489–2501, 2025.

[23] F. Oliveira, D. G. Costa, F. Assis, and I. Silva, "Internet of intelligent things: A convergence of embedded systems, edge computing and machine learning," *Internet of Things*, vol. 26, pp. 101 153–101 164, 2024.

[24] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[25] I. Cugu, E. Sener, and E. Akbas, "Microexpnet: An extremely small and fast model for expression recognition from face images," in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2019, pp. 1–6.

[26] S. Khan, M. Fazil, V. K. Sejwal, M. A. Alshara, R. M. Alotaibi, A. Kamal, and A. R. Baig, "Bichat: Bilstm with deep cnn and hierarchical attention for hate speech detection," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 4335–4344, 2022.

[27] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 1–9.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[30] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[31] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[32] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[33] Y. Chen, J. Li, S. Shan, M. Wang, and R. Hong, "From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos," *IEEE Transactions on Affective Computing*, 2024.

[34] Y. Liu, J. Kauttonen, B. Zhao, X. Li, and W. Peng, "Towards emotion ai to next generation healthcare and education," p. 1533053, 2024.

[35] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Information Sciences*, vol. 582, pp. 593–617, 2022.

[36] C. Martin, U. Werner, and H.-M. Gross, "A real-time facial expression recognition system based on active appearance models using gray images and edge images," in *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2008, pp. 1–6.

[37] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[38] M. Abdulrahman, T. R. Gwadabe, F. J. Abdu, and A. Eleyan, "Gabor wavelet transform based facial expression recognition using pca and lbp," in *2014 22nd signal processing and communications applications conference (SIU)*. IEEE, 2014, pp. 2265–2268.

[39] Q. Li, S. Zhan, L. Xu, and C. Wu, "Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow," *Multimedia Tools and Applications*, vol. 78, pp. 29 307–29 322, 2019.

[40] X.-H. Wang, A. Liu, and S.-Q. Zhang, "New facial expression recognition based on fsvm and knn," *Optik*, vol. 126, no. 21, pp. 3132–3134, 2015.

[41] S. Liu, D. Zhao, Z. Sun, and Y. Chen, "Bpmb: Bayescnns with perturbed multi-branch structure for robust facial expression recognition," *Image and Vision Computing*, vol. 143, p. 104960, 2024.

[42] H. Aouani and Y. Ben Ayed, "Deep facial expression detection using viola-jones algorithm, cnn-mlp and cnn-svm," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 65, 2024.

[43] M. Fazil, S. Khan, B. M. Albahlal, R. M. Alotaibi, T. Siddiqui, and M. A. Shah, "Attentional multi-channel convolution with bidirectional lstm cell toward hate speech prediction," *IEEE Access*, vol. 11, pp. 16 801–16 811, 2023.

[44] L. Wang, C. Yan, H. Wu, F. Zhu, S. Kumari, and M. J. Alenazi, "A privacy protection scheme for consumer electronics data based on blockchain and ai," *IEEE Transactions on Consumer Electronics*, 2025.

[45] J. Wei, G. Hu, X. Yang, A. T. Luu, and Y. Dong, "Learning facial expression and body gesture visual information for video emotion recognition," *Expert Systems with Applications*, vol. 237, p. 121419, 2024.

[46] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[47] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[48] J. Cai, O. Chang, X.-L. Tang, C. Xue, and C. Wei, "Facial expression recognition method based on sparse batch normalization cnn," in *2018 37th Chinese control conference (CCC)*. IEEE, 2018, pp. 9608–9613.

[49] S. Zhang, X. Pan, Y. Cui, X. Zhao, and L. Liu, "Learning affective video features for facial expression recognition via hybrid deep learning," *IEEE Access*, vol. 7, pp. 32 297–32 304, 2019.

[50] X. Pan, G. Ying, G. Chen, H. Li, and W. Li, "A deep spatial and temporal aggregation framework for video-based facial expression recognition," *IEEE Access*, vol. 7, pp. 48 807–48 815, 2019.

[51] Y. Li, Y. Liu, A. Nguyen, H. Shi, E. Vuorenmaa, S. Järvelä, and G. Zhao, "Interactions for socially shared regulation in collaborative learning: an interdisciplinary multimodal dataset," *ACM Transactions on Interactive Intelligent Systems*, vol. 14, no. 3, pp. 1–34, 2024.

[52] A. Radoi and G. Cioroiu, "Uncertainty-based learning of a lightweight model for multimodal emotion recognition," *IEEE Access*, 2024.

[53] X. Liu, Z. Yu, J. Jiang, B. Wang, F. Zhu, X. Chen, and W. Pedrycz, "Improved cross-modal retrieval systems using self-reinforcement and quadruplet alignment hashing," *IEEE Transactions on Consumer Electronics*, 2025.

[54] Q. Li, Y. Q. Liu, Y. Q. Peng, C. Liu, J. Shi, F. Yan, and Q. Zhang, "Real-time facial emotion recognition using lightweight convolution neural network," in *Journal of Physics: Conference Series*, vol. 1827, no. 1. IOP Publishing, 2021, p. 012130.

[55] H.-W. Xu, W. Qin, Y.-N. Sun, Y.-L. Lv, and J. Zhang, "Attention mechanism-based deep learning for heat load prediction in blast furnace ironmaking process," *Journal of Intelligent Manufacturing*, vol. 35, no. 3, pp. 1207–1220, 2024.

[56] Y. Liu, G. Liu, R. Zhang, D. Niyato, Z. Xiong, D. I. Kim, K. Huang, and H. Du, "Hallucination-aware optimization for large language model-empowered communications," *arXiv preprint arXiv:2412.06007*, 2024.

[57] N. S. Chauhan and N. Kumar, "Confined attention mechanism enabled recurrent neural network framework to improve traffic flow prediction," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108791, 2024.

[58] Y. Chen, R. Xia, K. Yang, and K. Zou, "Dnnam: Image inpainting algorithm via deep neural networks and attention mechanism," *Applied Soft Computing*, vol. 154, p. 111392, 2024.

[59] G. Zhao, Y. Zhang, M. Ge, and M. Yu, "Bilateral u-net semantic segmentation with spatial attention mechanism," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 2, pp. 297–307, 2023.

[60] L. Xia and Z. Li, "A new method of abnormal behavior detection using lstm network with temporal attention mechanism," *The Journal of Supercomputing*, vol. 77, no. 4, pp. 3223–3241, 2021.

[61] Z. Huang, B. Xu, M. Xia, Q. Li, L. Zou, S. Li, and X. Li, "Mscs: Multi-stage feature learning with channel-spatial attention mechanism for infrared and visible image fusion," *Infrared Physics & Technology*, p. 105514, 2024.

[62] Z. Li, S. Cao, J. Deng, F. Wu, R. Wang, J. Luo, and Z. Peng, "Stade-cdnet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[63] J. Zhang, X. Li, J. Tian, H. Luo, and S. Yin, "An integrated multi-head dual sparse self-attention network for remaining useful life prediction," *Reliability Engineering & System Safety*, vol. 233, p. 109096, 2023.

[64] R. Fu, H. Liang, S. Wang, C. Jia, G. Sun, T. Gao, D. Chen, and Y. Wang, "Transformer-bls: An efficient learning algorithm based on multi-head attention mechanism and incremental learning algorithms," *Expert Systems with Applications*, vol. 238, p. 121734, 2024.

[65] R. Xu, S. Yang, Y. Wang, B. Du, and H. Chen, "A survey on vision mamba: Models, applications and challenges," *arXiv preprint arXiv:2404.18861*, 2024.

[66] D. P. Lopez, F. Z. Canal, G. G. Scotton, E. Pozzebon, and A. C. Sobieranski, "A real-time computational approach for human facial expression recognition based on landmark feature extraction," *Seven Editora*, pp. 73–93, 2024.

[67] J. Ni, X. Zhang, and J. Zhang, "Multiscale feature fusion attention lightweight facial expression recognition," *International Journal of Aerospace Engineering*, vol. 2022, no. 1, p. 6523234, 2022.

[68] M. Sun, W. Cui, Y. Zhang, S. Yu, X. Liao, B. Hu, and Y. Li, "Attention-rectified and texture-enhanced cross-attention transformer feature fusion network for facial expression recognition," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 12, pp. 11 823–11 832, 2023.

[69] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017.

[70] X. Jin, Z. Lai, and Z. Jin, "Learning dynamic relationships for facial expression recognition based on graph convolutional network," *IEEE Transactions on Image Processing*, vol. 30, pp. 7143–7155, 2021.

[71] M. Yan, M. Zhao, Z. Xu, Q. Zhang, G. Wang, and Z. Su, "Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[72] S. Zhou, Y. Wang, D. Chen, J. Chen, X. Wang, C. Wang, and J. Bu, "Distilling holistic knowledge with graph neural networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 387–10 396.

[73] S. Zhang, H. Liu, J. E. Hopcroft, and K. He, "Class-aware information for logit-based knowledge distillation," *arXiv preprint arXiv:2211.14773*, 2022.

[74] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," *Advances in neural information processing systems*, vol. 31, 2018.

[75] R. Ferro-Pérez and H. Mitre-Hernandez, "Resmonet: a residual mobile-based network for facial emotion recognition in resource-limited systems," *arXiv preprint arXiv:2020.07649*, 2020.

[76] X. Jin, X. Song, X. Wu, and W. Yan, "Transformer embedded spectral-based graph network for facial expression recognition," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 6, pp. 2063–2077, 2024.

[77] B. Jiang, N. Li, X. Cui, W. Liu, Z. Yu, and Y. Xie, "Research on facial expression recognition algorithm based on lightweight transformer," *Information*, vol. 15, no. 6, p. 321, 2024.

[78] M. F. Altaf, M. W. Iqbal, G. Ali, K. Shinan, H. E. Alhazmi, F. Alanazi, and M. U. Ashraf, "Neural network-based ensemble approach for multi-view facial expression recognition," *PloS One*, vol. 20, no. 3, p. e0316562, 2025.

[79] J. R. Lee, L. Wang, and A. Wong, "Emotionnet nano: An efficient deep convolutional neural network design for real-time facial expression recognition," *Frontiers in Artificial Intelligence*, vol. 3, p. 609673, 2021.

[80] L. Liao, S. Wu, C. Song, and J. Fu, "Rs-xception: A lightweight network for facial expression recognition," *Electronics*, vol. 13, no. 16, p. 3217, 2024.

[81] N. Heidari and A. Iosifidis, "Progressive spatio-temporal bilinear network with monte carlo dropout for landmark-based facial expression recognition with uncertainty estimation," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2021, pp. 1–6.

[82] E. Bicer and H. Kose, "Lite-fer: A lightweight facial expression recognition framework for children in resource-limited devices," in *2024*

IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, 2024, pp. 1–9.

[83] R. Grover and S. Bansal, "Enhancing facial expression recognition in uncontrolled environment: a lightweight cnn approach with preprocessing," *Neural Computing and Applications*, vol. 37, no. 10, pp. 7363–7378, 2025.

[84] Y. Chen, C. Peng, X. Wang, and Y. Zheng, "Self-learning weight network based on label distribution training for facial expression recognition," *IET Image Processing*, vol. 19, no. 1, p. e13326, 2025.

[85] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[86] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep neural networks with relativity learning for facial expression recognition," in *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2016, pp. 1–6.

[87] R. Grover and S. Bansal, "Efficient facial expression recognition through lightweight cnn technique on public datasets," *SN Computer Science*, vol. 6, no. 1, p. 15, 2024.

[88] A. Ezati, M. Dezyani, R. Rana, R. Rajabi, and A. Ayatollahi, "A lightweight attention-based deep network via multi-scale feature fusion for multi-view facial expression recognition," *arXiv preprint arXiv:2403.14318*, 2024.

Xing Jin received the Ph.D degree in computer science and technology from Nanjing University of Science and Technology of China in 2021. He is now a lecturer in College of Information Science and Technology & Artificial Intelligence at Nanjing Forestry University. His research interests include computer vision, facial expression recognition and deep learning.

Shakir Khan (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science, in 1999, 2005, and 2011, respectively. He is currently working as an Associate Professor with the College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia. He is teaching bachelor's and master's degree courses at the College of Computer, Imam University. He has around 15 years of teaching, research, and IT experience in India and Saudi Arabia. His research interests include big data, data science, data mining, machine learning, the Internet of Things (IoT), e-learning, artificial intelligence, emerging technology, open-source software, library automation, and mobile/web application. He published many research papers in international journals and conferences in his research domain. He is a member of the International Association of Online Engineering (IAOE). He is a reviewer for many international journals.

Mehdi Hosseinzadeh received his B.Sc. degree in computer hardware engineering from IAU, Dezful Branch, Iran in 2003. He also received his M.Sc. and the Ph.D. degree in computer system architecture from the SRBIAU, Tehran, Iran in 2005 and 2008, respectively. Mehdi is currently an associate professor in Institute of Research and Development, Duy Tan University, Da Nang, Vietnam. He is the author/co-author of more than 120 publications in technical journals and conferences, and his research interests include IoT, SDN, information technology, data mining, big data analytics, E-Commerce, E-Marketing, and social networks.

Neeraj Kumar (Senior Member, IEEE) received the Ph.D. degree in CSE from Shri Mata Vaishno Devi University, Katra (J&K), India. He was a Postdoctoral Research Fellow with Coventry University, Coventry, U.K. He is currently as a Full Professor with the Department of Computer Science and Engineering, Thapar University, Patiala, India. He is also a Visiting Professor with Coventry University, Coventry, U.K. He has published more than 300 technical research articles in leading journals and conferences from the IEEE, Elsevier, Springer, and John Wiley. Some of his research findings are published in top cited journals, such as the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS (TIE), the IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING (TDSC), the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS), the IEEE TRANSACTIONS ON CLOUD COMPUTING (TCC), the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (TVT), the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS (TCE), the IEEE NETWORK, the IEEE COMMUNICATIONS, the IEEE WIRELESS COMMUNICATIONS (WC), the IEEE INTERNET OF THINGS JOURNAL (IoTJ), the IEEE SYSTEMS JOURNAL (SJ), FGCS, JNCA, and ComCom. He has guided many Ph.D. and M.E./M.Tech. His research is supported by fundings from Tata Consultancy Service, Council of Scientific and Industrial Research (CSIR), and Department of Science and Technology. He has awarded best research paper awards from the IEEE ICC 2018 and the IEEE SYSTEMS JOURNAL 2018. He is also leading the research group Sustainable Practices for the Internet of Energy and Security (SPINES), where group members are working on the latest cutting edge technologies. He is a TPC Member and a Reviewer of many international conferences across the globe.

Xiyin Wu received the Ph.D degree in pattern recognition and intelligence systems from Nanjing University of Science and Technology of China in 2020 and the Master's degree in computer technology from Jiangnan University of China in 2015. She is now a lecturer in data science and big data technology at Guizhou University. Her research interests include computer vision and medical image processing.