Research on Face Expression Recognition Model Based on Improved CTAN

Yancong Zhou, *Member, IAENG,* Dongdong Wang

***Abstract*—Traditional expression recognition models often struggle with effectively allocating computational resources during feature extraction and fail to adequately focus on regions with significant expression changes. The instability during training results in existing models facing significant performance fluctuations. To address these issues, this study constructs a novel expression recognition model based on ConvNeXt-Tiny, called the ConvNeXt-Tiny-based Combined Attention Network (CTAN). This model incorporates a dual-attention mechanism and Exponential Moving Average (EMA) to enhance overall performance. The dual-attention mechanism significantly improves the model's ability to capture features in critical regions. During training, EMA is used to smooth and update model parameters, improving training stability and reducing fluctuations. Experimental results show that the improved CTAN network achieves classification accuracy values of 98.99% on the CK+ dataset and 74.00% on the FER2013 dataset, representing an improvement of 0.69%-5.97% and 0.42%-5.50%, respectively, compared to current expression recognition models. These findings underscore the proposed model's significant advancements in both accuracy and stability over existing expression recognition models.**

***Index Terms*—Affective Computing, Complex Expression Recognition, ConvNeXt, FER2013**

I. Introduction

FACIAL Expression Recognition (FER) is a task of significant importance in the field of computer vision, aiming to recognize and assess the emotional state of an individual by automatically analyzing facial images or videos. In recent years, with the rapid development of deep learning technology and the widespread use of image capture devices such as smartphones and surveillance cameras, facial expression recognition has attracted considerable attention and research. Particularly in the field of smart devices, Affective Computing has become one of the hot research areas[1]. Researchers aim to improve the accuracy and real-time performance of expression recognition to apply it in emotional interaction systems, intelligent customer service, mental health monitoring, and autonomous driving[2]. Studies have shown that Convolutional Neural Networks (CNNs) perform poorly when facing variations in lighting, posture, and facial occlusion. Additionally, due to the hierarchical design of CNNs, while high-level features are rich in semantic information, fine-grained expression details are often overlooked, leading to poor recognition of subtle features such as micro-expressions[3]. Although some enhancements, such as multi-scale feature fusion or attention mechanisms, have been introduced, these tend to increase the computational cost and complexity of the models[4].

In summary, deep learning-based facial expression recognition research has advanced significantly and laid the technological foundation for its development and practical applications. However, several limitations remain in the research process: First, existing facial expression recognition models often exhibit low recognition rates when facing real-world challenges. Second, current deep learning models rely on single-method feature extraction, which leads to incomplete feature extraction, difficulty in capturing key expression features, and the neglect of important information.

Main Contributions of this Paper:

1. An improved CTAN is proposed. This model integrates Efficient Attention and Shuffle Attention modules, enhancing the ability to capture subtle expression changes. Efficient Attention separates high- frequency and low-frequency features, while Shuffle Attention optimizes feature selection, leading to improved accuracy and robustness in expression recognition.
2. The Exponential Moving Average (EMA) technique is combined with a multi-level attention mechanism to smooth parameter updates, reduce training fluctuations, and prevent overfitting, leading to enhanced stability and performance during feature extraction.
3. The CTAN model was validated on FER2013, CK+, KDEF, and JAFFE datasets, including a combined FER_CK+_KDEF dataset. Results show significant accuracy improvements (0.69%-5.97%) on CK+ and (0.42%-5.50%) on FER2013, outperforming existing models in accuracy, robustness, and generalization.

II. Related work

In facial expression recognition, CNNs, despite wide application, have significant limitations. The primary issues are fixed convolutional kernels and limited receptive fields, making it hard to capture multiscale and subtle changes. Studies show CNNs' recognition drops drastically under

Manuscript received Nov 28, 2024; revised Mar 14, 2025.

This work was supported in part by the Tianjin Graduate Research Innovation Program under Project Number 2022SKYZ312.

Yancong Zhou is a professor of Tianjin University of Commerce, Tianjin 300134 China (corresponding author to provide phone: +86-156-9222-1287; fax: 022-26667577; e-mail: zycong78@126.com).

Dongdong Wang is a postgraduate student of Tianjin University of Commerce, Tianjin 300134 China (e-mail: wangdongdong0127@163.com).

lighting, pose, and occlusion changes[2]. Additionally, while CNNs' high-level features are strong, they often overlook fine-grained expression details, leading to poor micro-expression recognition [3]. Enhancement techniques like multi-scale feature fusion or attention mechanisms exist, but they increase computational cost and complexity [4]. The process of emotion recognition based on deep learning involves: After image preprocessing and face detection cropping, the facial expression dataset is input into a deep learning neural network model to extract expression features, and these features are then used for emotion recognition.

In face detection and preprocessing, Jiang et al.[5] used data augmentation to enhance sample diversity and network generalization. Cui et al.[6] combined an improved VGGNet with focal loss to mitigate the impact of mislabeled samples. To address complex model structures and large parameter sizes, Shen et al.[7] improved the inverted residual network to build a lightweight convolutional model, fusing features for expression recognition. Gao et al.[8] focused on facial, eye, and mouth regions for feature extraction and classification. Li et al.[9] designed base classifiers with depth separable convolutions for model lightweighting.

Regarding image feature extraction during model training, Huagang Liang et al.[10] introduced a compression excitation module that assigns weights to features from different channels, using varying compression rates in different convolution layers to enhance the network's ability to extract facial expression features. Peng Zhang et al.[11] enhanced the Inception structure by adding dilated convolutions to extract multi-scale features of facial expressions, and introduced a channel attention mechanism to improve the model's ability to emphasize important features, yielding good results. Kai Wang et al.[12] proposed a Regional Attention Network (RAN) that adaptively adjusts the importance of facial regions. They further designed a Regional Bias Loss (RB-Loss) function to encourage high attention weights for the most important regions, effectively addressing challenges posed by occlusions and pose variations in real-world expression recognition.

Recently, following Transformer's success in NLP, its application in CV tasks has gained traction. Self-attention-based models like ViT have been explored for expression recognition. However, Transformer models have drawbacks: they rely on global feature extraction, leading to overfitting with small datasets or low-res images[13]; they require high computational resources, making training and inference more expensive than CNNs for high-res images [14]. Although lightweight architectures like Swin Transformer [15], have been proposed, Transformers still lag behind CNNs in capturing local details for facial expression recognition [16].

The application of the Transformer architecture in computer vision (CV) began with the introduction of Vision Transformer (ViT). Unlike traditional CNNs, ViT divides an image into fixed-size patches and treats them as a sequence to be input into the Transformer model, enabling the learning of global image features[17]. Cho et al.[18] proposed the DA-ViT model, which adjusts ViT for robust feature learning across different visual domains. Xue et al.[19] addressed ViT's lack of adaptability and susceptibility to noise in facial expression recognition by introducing attention pooling modules (APP and ATP) to improve recognition accuracy and reduce the impact of irrelevant features. Indolia et al.[20] combined CNNs and ViT to propose a convolutional-patch-based visual transformer, solving the challenge of capturing both local information and global dependencies, thus advancing micro-expression recognition.

The Swin Transformer builds upon ViT by introducing a sliding window mechanism [15] and down sampling layers, enabling the model to efficiently handle high-resolution images and demonstrate superior performance in various vision tasks. Qin et al.[21] integrated a Multi-Level Channel Attention (MLCA) module into the Swin Transformer, proposing the Swin Face model, which significantly improved facial expression recognition accuracy. Feng et al.[22] proposed the FST-MWOS method, using fine-tuning and optimization strategies to further enhance the performance of facial expression recognition models in real-world environments. However, the advantages of these methods largely rely on the powerful capabilities of Transformers, while the inherent inductive biases of CNNs still present challenges[23]. Additionally, Transformer models face challenges in image processing, including high computational complexity, weaker performance in capturing local information, large numbers of parameters, and significant data requirements[24], [25], [26]. Given the limitations of Transformer models, the ConvNeXt network, as a convolutional architecture designed to match ViT, has shown advantages in handling local image information and reducing computational complexity.

In conclusion, current mainstream expression recognition methods, whether based on CNNs or Transformers, have certain limitations. While CNNs excel in hierarchical feature extraction, their performance in multi-scale feature capture and complex scenarios is limited. On the other hand, while Transformers can capture global information better, they have not shown significant advantages in expression recognition tasks due to their computational cost and dependency on large datasets[27]. Therefore, this study chooses to improve the ConvNeXt architecture for facial expression recognition, aiming to enhance the model's ability to capture local features while maintaining efficient computation. Practical applications in facial expression recognition still face challenges such as insufficient feature extraction[28], low recognition accuracy[29], and large model sizes[30]. To address these issues, this study proposes an improved ConvNeXt network, incorporating attention mechanisms and exponential moving averages (EMA) to enhance feature diversity and model stability, particularly in addressing the diverse variations of facial expressions, significantly improving the model's performance.

III. METHODOLOGY

A. Research Methodology

This study proposes an enhanced CTAN for facial expression recognition, aiming to improve the accuracy and robustness of the task by refining both feature extraction and network parameter update mechanisms. The model integrates multiple attention mechanisms and the EMA to optimize for diverse facial expression changes. The design emphasizes

feature diversity, model stability, and multi-level feature focus capability.

Initially, datasets FER2013, KDEF, and CK+ were integrated, covering eight expression categories: anger, contempt, disgust, fear, happiness, sadness, surprise, and neutrality (Fig 1a). Preprocessing and augmentation steps were applied to reduce variability between datasets. Despite offering a diverse training dataset, initial experiments using the baseline ConvNeXt-Tiny network did not yield the expected performance gains due to dataset heterogeneity.

Subsequently, as shown in Fig 1b, the ConvNeXt-Tiny architecture was improved by integrating Efficient Attention and Shuffle Attention modules to enhance feature extraction capabilities. Efficient Attention captures global and local feature dependencies, while Shuffle Attention refines the sensitivity to subtle expression changes through channel and spatial attention mechanisms. These enhancements address the challenges posed by expression diversity and subtle expression detection. Additionally, the EMA strategy was employed to stabilize training and improve generalization by smoothing parameter updates during the training process.

Fig 1c demonstrates the evaluation of the model on the integrated dataset and separately on FER2013, KDEF, and CK+ to validate generalization and robustness. Results are presented in Section 4.

B. Model Architecture

To improve facial expression recognition performance, this study used the ConvNeXt-Tiny model as the base network and introduced improvements. Proposed in 2023, ConvNeXt combines the efficiency of ConvNets with the potential to rival Vision Transformers (ViTs)[31]. ConvNeXt-Tiny, a lightweight variant, is inspired by ResNet and Swin Transformer, stacking deep convolutional modules and using larger kernels to maintain efficiency and robustness. Its simple structure and low computational cost make it well-suited for resource-constrained environments, achieving high accuracy with minimal overhead.

In this study, the improved CTAN model adopts a modular design, incorporating convolutional layers, attention modules, and an EMA module to enhance feature extraction capabilities and improve model stability. The input data first undergoes convolutional operations to extract primary features, which are then enhanced using the Efficient Attention and Shuffle Attention modules. By introducing attention mechanisms, the model can more precisely focus on the key information in the input images. Furthermore, the inclusion of the EMA module smooths parameter updates, enhancing the model's stability and generalization during both training and testing phases.

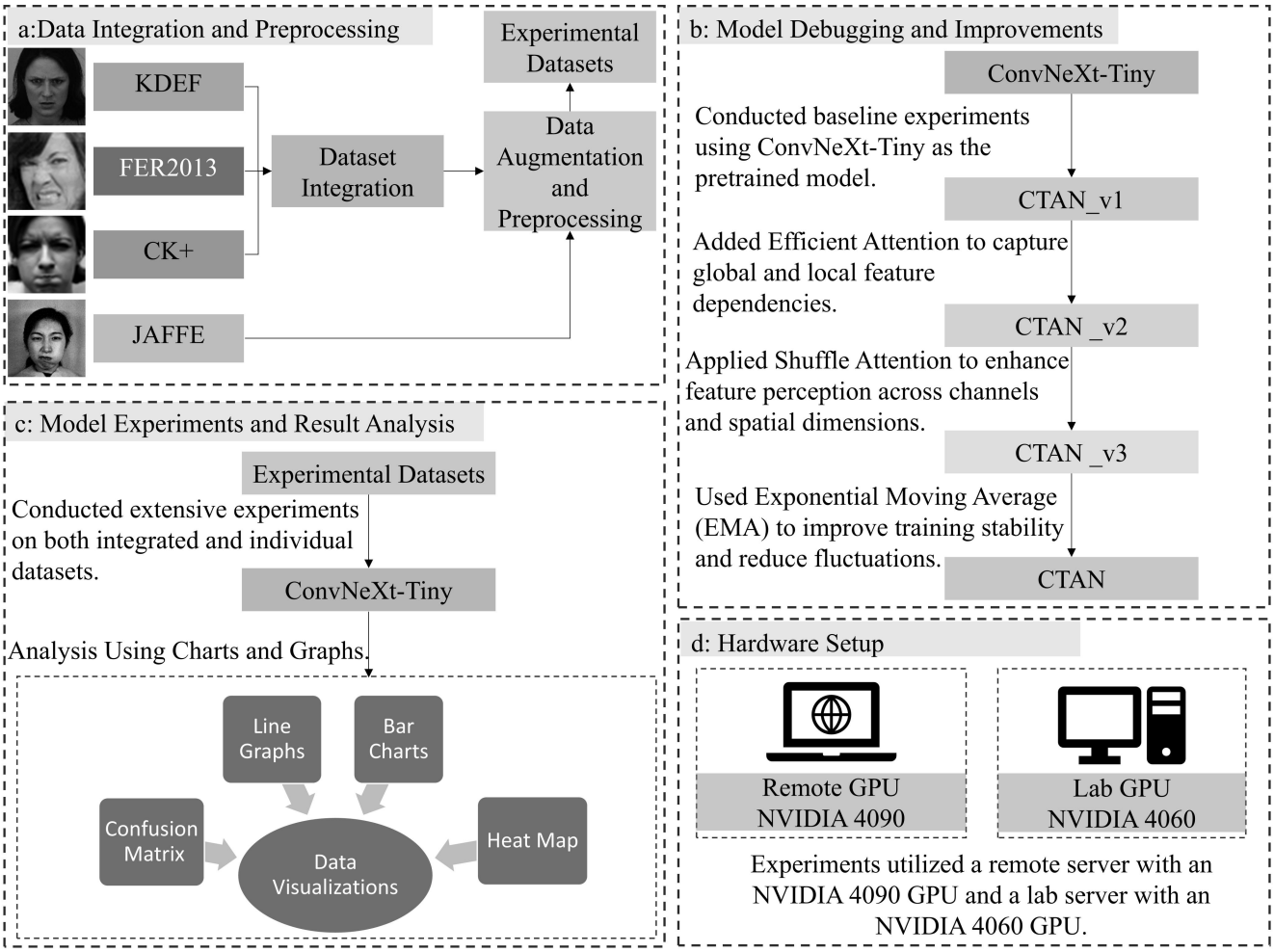

Fig. 1. Illustrates the research workflow:(a) Data integration and preparation; (b) Model optimization; (c) Experimental result analysis; (d) Hardware and software tools utilized in the study.

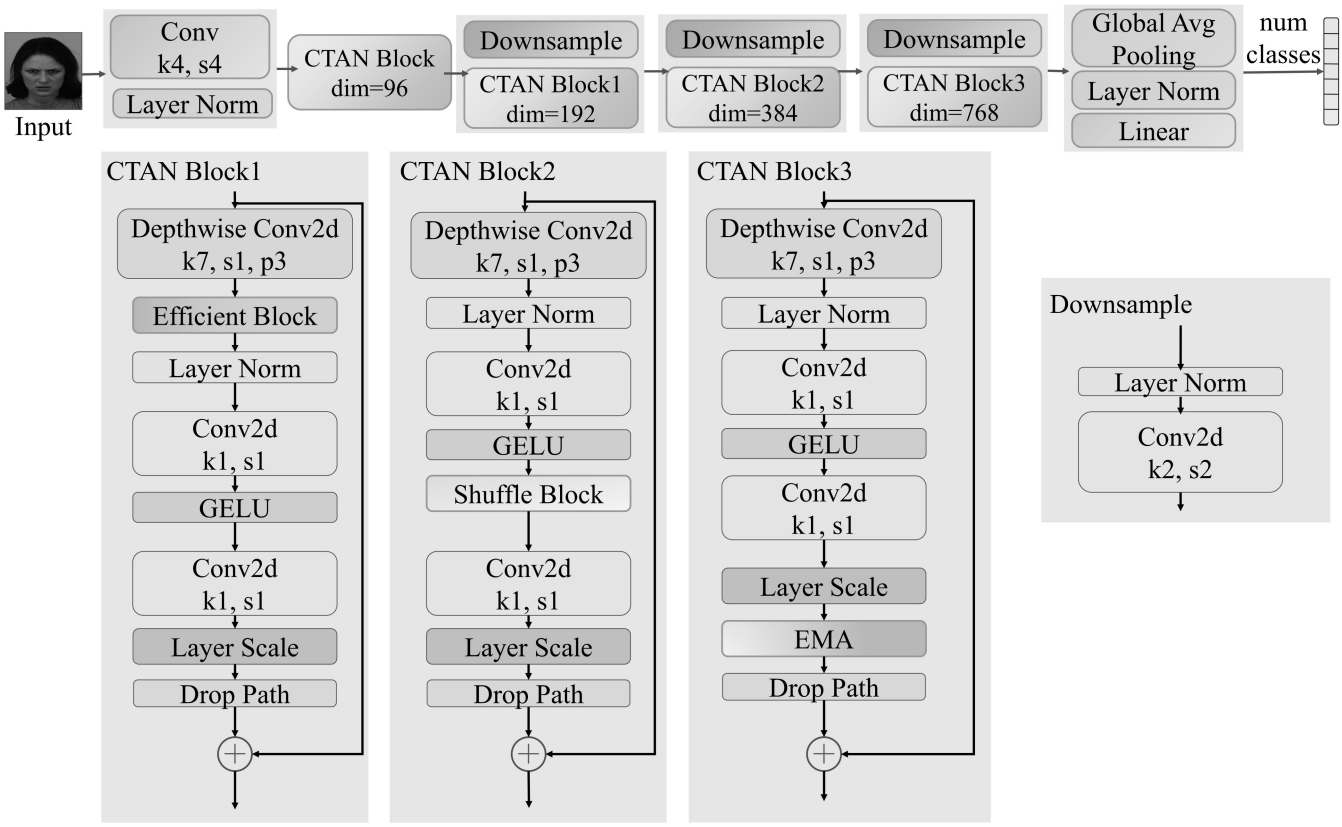

Fig. 2. Architecture of the Improved ConvNeXt-Tiny-based Combined Attention Network (CTAN), which uses Efficient Attention and Shuffle Attention to optimize the ConvNeXt-T network, and employs EMA to update the model parameters.

Fig 2 depicts the architecture of the improved CTAN model, including a simplified schematic design of the Efficient Attention and Shuffle Attention modules. The network is composed of multiple hierarchical layers, progressively extracting features of different levels from the input image to ensure comprehensive representation of expression features. The convolutional layers handle initial feature extraction, while the attention modules enhance feature expressiveness by integrating both local and global information. Afterward, the features undergo EMA optimization, which improves the weight update process and ensures robustness during training.

In the CTAN network, the EMA module and attention mechanisms collaborate effectively to extract critical features and ensure smooth parameter updates. This design is highly scalable, allowing flexible adjustments to the network's depth and width based on task requirements, thereby enhancing the model's generalization ability, stability, and performance, especially in expression recognition tasks. The internal parameters of the CTAN network are presented in Table 1.

C. *Efficient Attention Module*

The Efficient Attention module is designed to capture both global and local features, enhancing the model's ability to detect subtle expression changes. This is achieved by splitting the input features into two branches: one for global features and the other for local features. Each branch is processed using global channel attention and local channel attention, respectively, as shown in Fig. 3.

The global feature branch utilizes the global average pooling (GAP) method to compute the global feature map, as shown in Eq. (1):

$$F_G = \mathrm{GAP}(X) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{ij} \tag{1}$$

Where X is the input feature, and H and W represent the height and width of the feature map, respectively. The computed global features are used to generate attention weights, which are then multiplied element-wise with the original features to obtain the global attention features.

The local feature branch extracts local features through multiple depth convolutional layers and further processes them using an activation function (e.g., Swish). This process is represented by the following Eq. (2):

$$F_L = \mathrm{Activation}(\mathrm{Conv}(X)) \tag{2}$$

TABLE I
INTERNAL PARAMETERS OF THE CTAN

Module	In-put	Output	Kernel Size	Stride	Padding	Stacks
Stem	3	96	4	0	0	1
Basic RFB	96	96	-	-	-	1
ConvBlock1	96	192	7	1	3	3
ConvBlock2	192	384	7	1	3	3
ConvBlock3	384	768	7	1	3	9
ConvBlock4	768	768	7	1	3	3
Efficient Att	768	768	-	-	-	1
Shuffle Att	768	768	-	-	-	1
EMA	768	768	-	-	-	1
Layer Norm	768	768	-	-	-	1
FC Layer	768	num classes	-	-	-	1

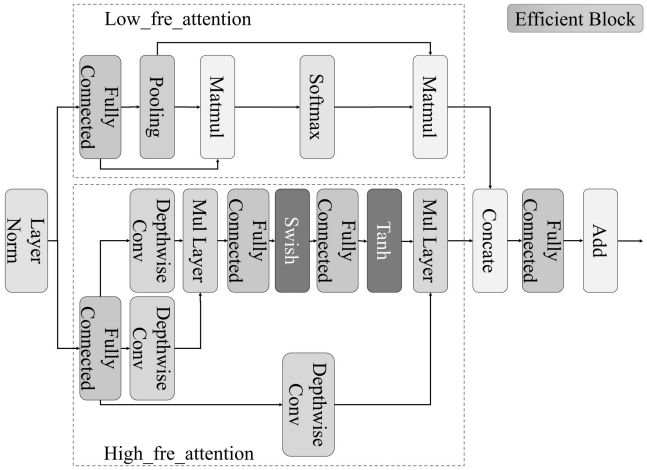

Fig. 3. Architecture of the Efficient Attention module, where the input features are divided into two branches—global and local. These branches are processed through high-frequency and low-frequency channels, respectively, and then, during the feature fusion stage, the two feature sets are concatenated to obtain the final feature representation.

In the feature fusion stage, the global and local features are concatenated to form the final feature representation. This design enables Efficient Attention to effectively balance long-range dependencies and local details, significantly enhancing the performance of face expression recognition.

D. Shuffle Attention module

The Shuffle Attention module is used to optimize the selection and representation of features, enhancing the model's ability to capture key information. This module improves feature enhancement through channel shuffling and the channel attention mechanism, as shown in Fig. 4. After the input features are rearranged through channel shuffling, the module employs adaptive pooling to generate channel attention weights, as shown in Eq. (3).

$$W_{att} = \mathrm{Softmax}(\mathrm{GAP}(X)) \tag{3}$$

Subsequently, the attention weights are multiplied element-wise with the original features to highlight the important features. This process enables the Shuffle Attention module to effectively enhance feature expressiveness while reducing redundant information, ultimately improving the accuracy of expression recognition.

E. The Role of Exponential Moving Average (EMA)

To further enhance model stability, the EMA mechanism is introduced into the network. EMA smooths parameter updates by maintaining a moving average of model parameters, which helps reduce model fluctuations and avoid overfitting. Compared to direct updates, EMA offers a more stable optimization method, leading to better performance on the validation and test sets. The EMA module smooths parameter updates, reducing volatility during training. The core idea behind EMA is to replace the current value with the average value from the previous time period, thereby enhancing model stability. Specifically, the EMA update formula is shown in Eq. (4):

$$\theta_{\mathrm{EMA}} = \alpha\theta_{\mathrm{current}} + (1-\alpha)\theta_{\mathrm{EMA}} \tag{4}$$

Where θ_{EMA} denotes the model parameters updated by EMA, $\theta_{current}$ denotes the model parameters updated by EMA, α is the decay coefficient controlling the weight of the historical parameters. At each training step, the EMA module helps reduce the risk of overfitting by maintaining the moving average of the historical parameters, improving the model's performance on the validation set. The incorporation of the EMA mechanism significantly enhances the model's generalization ability, ensuring stability and consistency across different datasets.

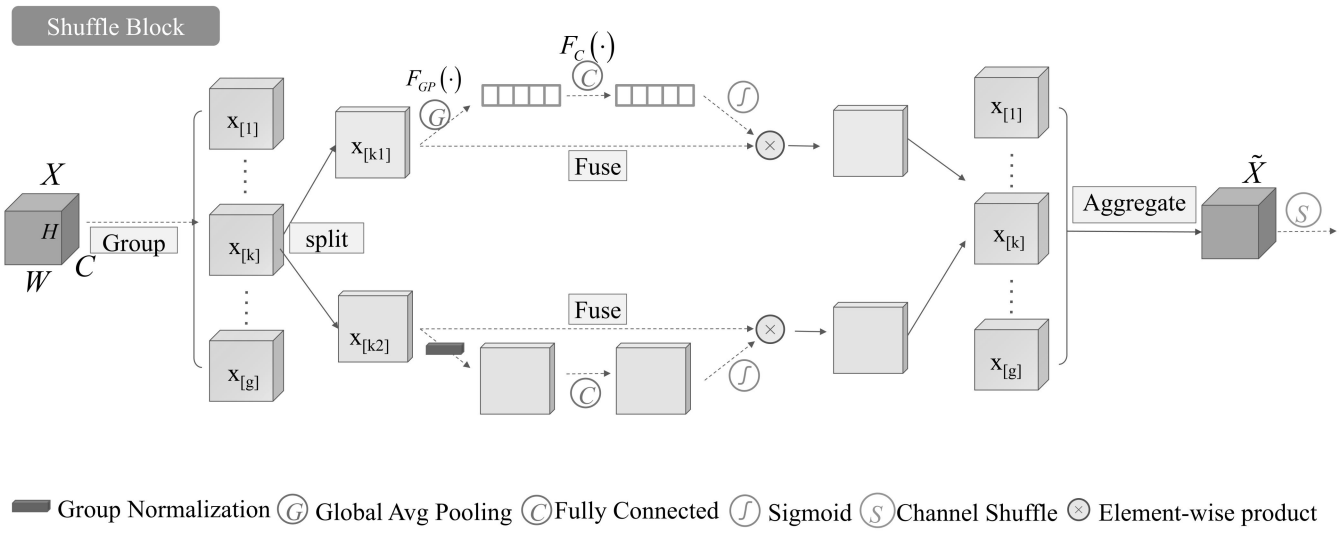

Fig. 4. Shuffle Attention module architecture, it groups channel features, extracts multiple sub-features, applies both spatial and channel attention mechanisms to each sub-feature, and finally pools all the sub-features, realizing the fusion of different feature groups through the Channel Shuffle operation.

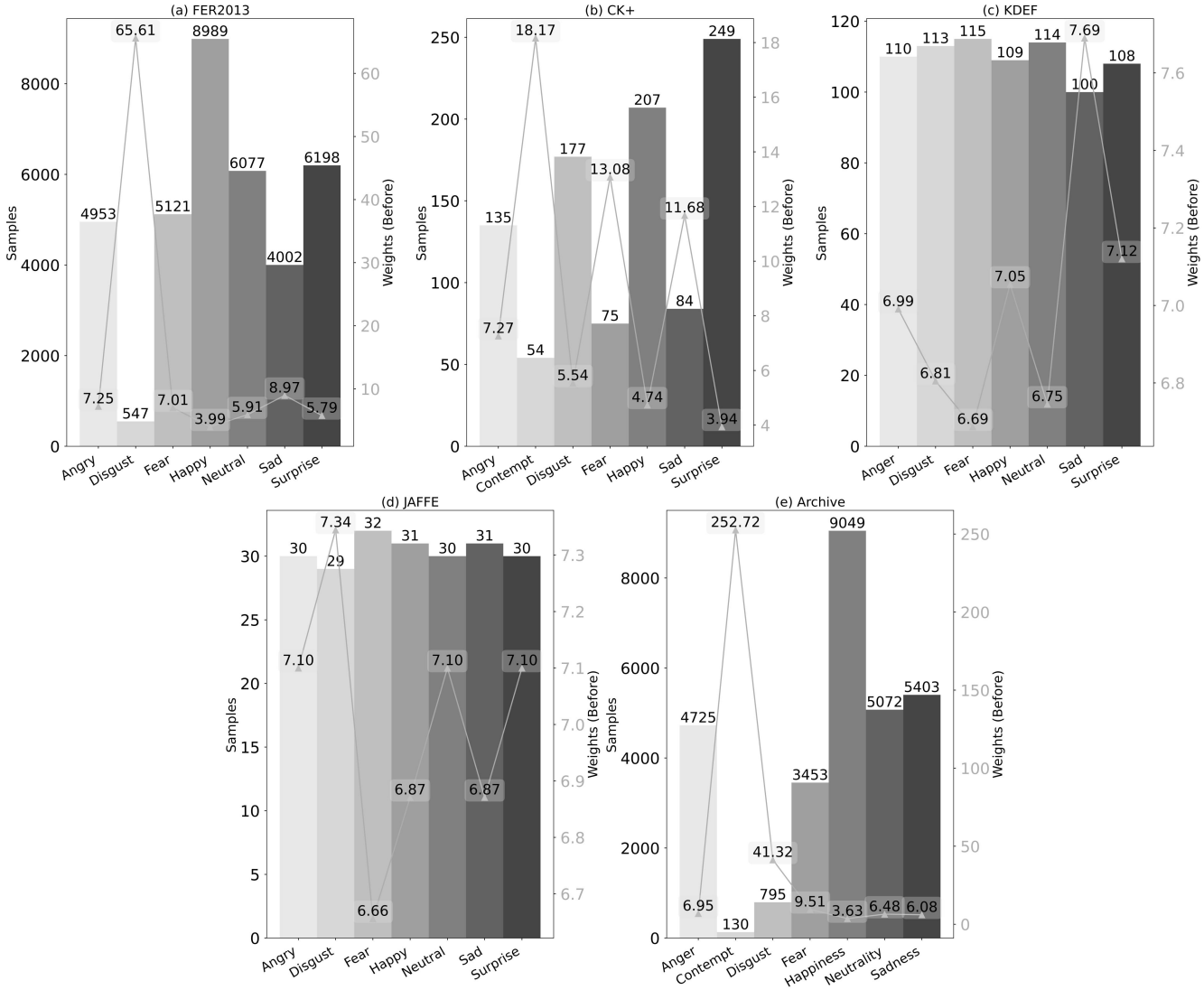

Fig. 5. Distribution of Sample Counts and Pre-Normalization Weight Values Across Different Datasets.

IV. Dataset and Experimental Setup

A. Experimental Datasets

In this study, we tested the improved ConvNeXt network on several publicly available facial expression recognition datasets, including FER2013 (Facial Expression Recognition 2013), KDEF (Karolinska Directed Emotional Faces), CK+ (Extended Cohn-Kanade Dataset), JAFFE (Japanese Female Facial Expression), and an integrated dataset.

FER2013 contains 35,887 grayscale images (28x28 resolution) covering 7 expressions, widely used for evaluating model robustness. CK+ includes image sequences covering 7 expressions, primarily for micro-expression and dynamic expression research. KDEF consists of 4,900 color images from 70 individuals across 7 expressions and 5 angles, ideal for multi-angle recognition. JAFFE contains 213 images from 10 Japanese females, suitable for small sample expression recognition.

The diversity of these datasets ensures comprehensive evaluation of our model's generalization and robustness, confirming the improved ConvNeXt network's ability to handle various challenges.

B. Data Preprocessing and Augmentation

The datasets were split into training, validation, and test sets (8:1:1 ratio). Fig. 5 shows the sample numbers and pre-normalized weights across datasets.

Prior to model training, we performed various preprocessing operations on the data. First, all images were resized to 224x224, and histogram equalization was applied to enhance the contrast of the images. Next, the images were transformed into tensors, and normalization was performed using mean values of [0.5, 0.5, 0.5] and a standard deviation of [0.5, 0.5, 0.5]. To address class imbalance, weighted loss function based on category sample sizes was adopted. Finally, the data were divided into training and validation sets using stratified sampling to maintain consistent category proportions.

With this method, smaller categories receive higher weights, ensuring more balanced contributions from different categories during training, effectively addressing the class imbalance problem. Fig 6 shows the normalized weight values of different categories in several datasets. Tables 2 to 6 show the weighted loss function weights for each expression category based on sample sizes in these datasets.

TABLE 7
ABLATION STUDY ANALYSIS

Dataset	Model	Epoch	Learning rate	Loss	Acc	Fps	Latency
CK+	v2	200	5E-5	0.384	93.00	140.57	0.0286s
	v2	300	5E-5	0.260	96.95	134.89	0.0297s
	v3	300	5E-5	0.267	98.99	136.34	0.0294s
FER2013	v1	200	2E-3	0.769	52.00	-	-
	v2	300	1E-4	0.351	85.77	105.35	0.0427s
	v3	100	5E-5	0.680	71.31	229.07	0.0175s
	v3	200	1E-4	0.658	80.27	115.52	0.0393s
	v3	300	5E-5	0.231	90.69	108.78	0.0300s

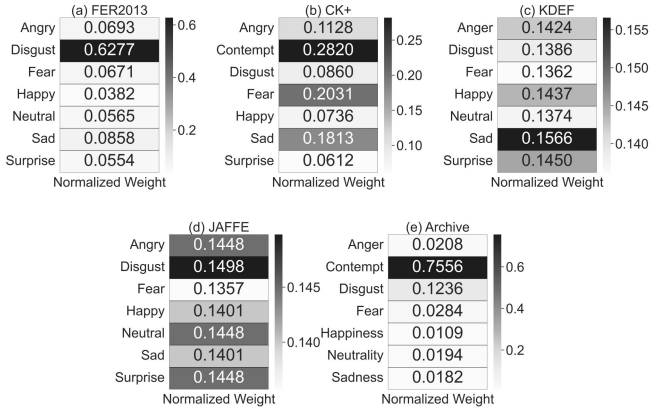

Fig. 6. Post-Normalization Weight Values for Each Expression Category Across Different Datasets.

With this method, smaller categories receive higher weights, balancing contributions across categories to address class imbalance. Fig 6 shows the normalized weight values of different categories of several data sets. Tables 2 to 6 show the weighted loss function weights for each expression category based on sample sizes in several datasets.

TABLE 2
WEIGHTS OF LOSS FUNCTION BASED ON CLASS SAMPLES IN FER2013

Expression Category	Sample Count	Computed Weight (Pre-Normalization)	Normalized Weight
Angry	4953	7.2455	0.0693
Disgust	547	65.6069	0.6277
Fear	5121	7.0078	0.0671
Happy	8989	3.9923	0.0382
Neutral	6077	5.9054	0.0565
Sad	4002	8.9673	0.0858
Surprise	6198	5.7901	0.0554

TABLE 3
WEIGHTS OF LOSS FUNCTION BASED ON CLASS SAMPLES IN CK+

Expression Category	Sample Count	Computed Weight (Pre-Normalization)	Normalized Weight
Angry	135	7.2667	0.1128
Contempt	54	18.1667	0.2820
Disgust	177	5.5424	0.0860
Fear	75	13.0800	0.2031
Happy	207	4.7391	0.0736
Sad	84	11.6786	0.1813
Surprise	249	3.9398	0.0612

TABLE 4
WEIGHTS OF LOSS FUNCTION BASED ON CLASS SAMPLES IN KDEF

Expression Category	Sample Count	Computed Weight (Pre-Normalization)	Normalized Weight
Anger	110	6.9909	0.1424
Disgust	113	6.8053	0.1386
Fear	115	6.6870	0.1362
Happy	109	7.0550	0.1437
Neutral	114	6.7456	0.1374
Sad	100	7.6900	0.1566
Surprise	108	7.1204	0.1450

TABLE 5
WEIGHTS OF LOSS FUNCTION BASED ON CLASS SAMPLES IN JAFFE

Expression Category	Sample Count	Computed Weight (Pre-Normalization)	Normalized Weight
Angry	30	7.1000	0.1448
Disgust	29	7.3448	0.1498
Fear	32	6.6562	0.1357
Happy	31	6.8710	0.1401
Neutral	30	7.1000	0.1448
Sad	31	6.8710	0.1401
Surprise	30	7.1000	0.1448

TABLE 6
WEIGHTS OF LOSS FUNCTION BASED ON CLASS SAMPLES IN ARCHIVE

Expression Category	Sample Count	Computed Weight (Pre-Normalization)	Normalized Weight
Anger	4725	6.9530	0.0208
Contempt	130	252.7154	0.7556
Disgust	795	41.3245	0.1236
Fear	3453	9.5143	0.0284
Happiness	9049	3.6306	0.0109
Neutrality	5072	6.4773	0.0194
Sadness	5403	6.0805	0.0182

C. *Test Setup*

The experimental environment is configured as follows: a 64-bit Windows 11 operating system, an Intel(R) Core(TM) i9-13900K CPU, 147GB of RAM, and an NVIDIA GeForce RTX 4090 GPU with 24GB of dedicated video memory. The experiments are conducted using the PyTorch 1.8 deep learning framework and Python 3.7 as the programming language, leveraging CUDA 11.8 and cuDNN 8.5 for GPU acceleration. This setup provides a powerful platform for efficiently running resource-intensive experiments.

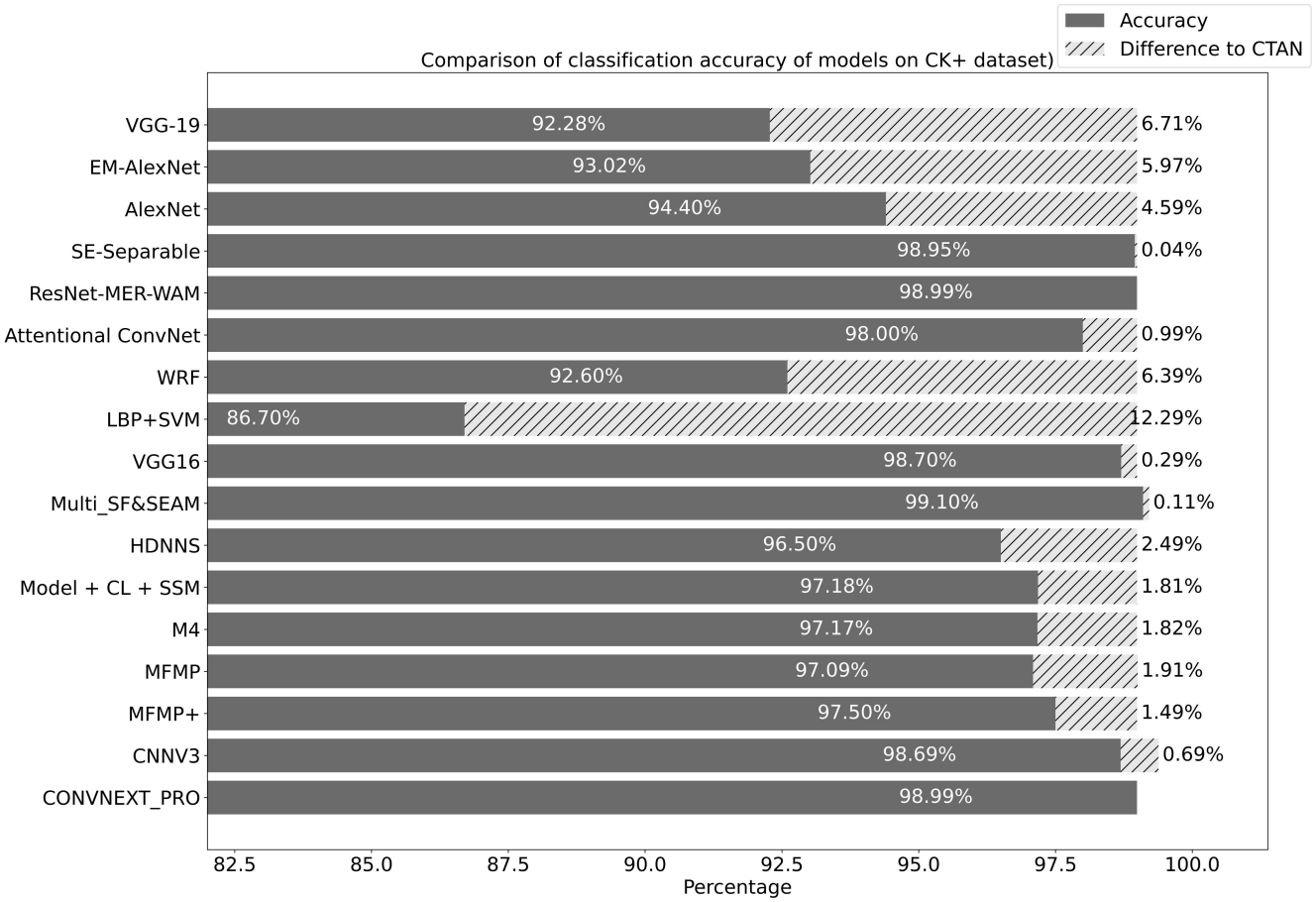

Fig. 7. Comparison of Classification Accuracy of Models on the CK+.

The optimizer used is Adam, with an initial learning rate of 5e-5 and a Cosine Annealing LR scheduler to gradually adjust the learning rate, with a minimum learning rate of 1e-9. The model is trained using the Cross Entropy Loss function with a batch size of 4 over 300 epochs. Drop Path is applied to enhance the model's generalization ability, with a drop path rate of 0.3. After each epoch, the model is evaluated on the validation set to select the optimal model.

D. Test Setup

The experimental environment is configured as follows: 64-bit Windows 11 operating system, Intel(R) Core(TM) i9-13900K CPU, 147GB RAM, and NVIDIA GeForce RTX 4090 GPU with 24GB of video memory. The experiments are based on the PyTorch 1.8 deep learning framework, with Python 3.7 as the programming language.

The optimizer used is Adam, with an initial learning rate of 5e-5 and a Cosine Annealing LR scheduler to gradually adjust the learning rate, with a minimum learning rate of 1e-9. The model is trained using the Cross Entropy Loss function with a batch size of 4 over 300 epochs. Drop Path is applied to enhance the model's generalization ability, with a drop path rate of 0.3. After each epoch, the model is evaluated on the validation set to select the optimal model.

E. Model evaluation setup

To improve the performance evaluation of each model, we used precision, recall, F1 score, and accuracy as the evaluation metrics to comprehensively assess the model's performance, as shown in Eq. (5) to (8).

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \tag{6}$$

$$\text{F1\_score} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \times 100\% \tag{7}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{8}$$

Where TP, TN, FP, and FN denote the number of true positive, true negative, false positive, and false negative samples, respectively. Precision is an estimate of how many of the predicted positive samples are actually positive, as shown in Eq. (5). Recall measures how many of the actual positive samples can be correctly predicted as positive, as shown in Eq. (6). The F1 score is the weighted average of precision and recall, as shown in Eq. (7). Accuracy is the most intuitive measure of model quality, as shown in Eq. (8).

V. Analysis of Experimental Results

A. Ablation Analysis

Table 7 presents a comparative analysis of models (v1, v2, v3 representing CTAN_v1, CTAN_v2, CTAN_v3, respectively, as shown in Fig. 1b) on CK+ and FER2013 datasets, based on training loss, accuracy, FPS, and latency,

after adjusting hyperparameters like epochs and learning rate.

As shown in Table 7, adjustments to the number of epochs, model structure, and hyperparameters lead to significant performance changes. In the CK+ dataset, CTAN_v3 outperforms CTAN_v2 in accuracy without a significant increase in training time, indicating better feature extraction and optimization. In the FER2013 dataset, despite higher FLOPs and MACs, CTAN_v1 struggles with accuracy and efficiency, especially at higher learning rates where overfitting occurs. However, after reducing the learning rate and using the Adam optimizer, the performance of CTAN_v2 and CTAN_v3 improves significantly. Especially, CTAN_v3 not only performs excellently in terms of accuracy but also maintains more stable FPS and lower latency, showing stronger generalization for large-scale datasets.

B. Model Performance on Datasets

Table 8 compares recognition accuracy for different models on the CK+ dataset. Classical models like VGG-19, AlexNet, and ResNet perform well, but recent improvements have further boosted accuracy. For example, SE-Separable and ResNet-MER-WAM models achieved accuracies of 98.95% and 98.99%, respectively, demonstrating a significant performance boost with the introduction of the attention mechanism. Notably, the CNNV3 model surpasses the other models with an accuracy of 99.68%, highlighting the strong capability of this network in feature extraction and classification. Additionally, the CONVNEXT_PRO model proposed in this paper achieved 98.99% accuracy on the CK+ dataset, demonstrating its competitiveness and robustness compared to state-of-the-art models.

TABLE 8
COMPARISON OF CLASSIFICATION ACCURACY OF MODELS ON THE CK+

Dataset	Models	Accuracy	Difference to CTAN
CK+	VGG-19	92.28%	-6.71%
	EM-AlexNet[32]	93.02%	-5.97%
	AlexNet	94.40%	-4.59%
	SE-Separable[10]	98.95%	-0.04%
	ResNet-MER-WAM[5]	98.99%	0.00%
	Att- Conv-Net[33]	98.00%	-0.99%
	Weighted Random Forest[34]	92.60%	-6.39%
	LBP+SVM[35]	86.70%	-12.29%
	VGG16[36]	98.70%	-0.29%
	Multi_SF&SEAM	99.10%	0.11%
	HDNNS[37]	96.50%	-2.49%
	Model + CL + SSM[38]	97.18%	-1.81%
	M4[39]	97.17%	-1.82%
	MFMP[39]	97.09%	-1.90%
	MFMP+[39]	97.50%	-1.49%
	CNNV3[40]	99.68%	0.69%
	CONVNEXT_PRO	98.99%	-

Meanwhile, the confusion matrix on the CK+ dataset shown in Fig 8 further illustrates the specific performance of the CONVNEXT_PRO model across different categories. It can be observed that the model achieves extremely high accuracy in most categories (e.g., 100% accuracy in Angry, Fear, Surprise, etc.). Although there is a slight decrease in accuracy for the "Sad" category (82%), overall, the model performs very stably and reliably in emotion classification tasks. These results validate the effectiveness and superiority of the CONVNEXT_PRO model in handling multi-class emotion recognition tasks.

TABLE 9
COMPARISON OF CLASSIFICATION ACCURACY OF MODELS ON THE FER2013

Dataset	Methods	Accuracy	Difference to CTAN
FER2013	MobileNetV2	38.85%	-35.15%
	Inception V3	39.06%	-34.94%
	ResNet-18	61.77%	-12.23%
	AlexNet	65.80%	-8.20%
	VGG-16	66.20%	-3.38%
	SE-Separable[10]	70.30%	-3.70%
	VGG-FL[6]	72.49%	-1.51%
	CS-AttNet-Trans[41]	73.37%	-0.63%
	RESNET50[42]	68.25%	-5.75%
	NGO-BILSTM[43]	51.29%	-22.71%
	Mobile ViT[44]	62.73%	-11.27%
	Region ViT[45]	56.03%	-17.97%
	Multi_SF&SEAM	68.50%	-5.50%
	Dual-branch CNN[46]	54.64%	-19.36%
	Light-CNN[46]	68.00%	-6.00%
	DAM-CNN[47]	66.20%	-7.80%
	CNN Model1[48]	65.77%	-8.23%
	CNN Model2[48]	65.23%	-8.77%
	SHCNN[49]	69.10%	-4.90%
	eXnetcutout[50]	71.92%	-2.08%
	eXnetcm[50]	73.54%	-0.46%
	eXnetmixup[50]	72.67%	-1.33%
	eXnet[50]	71.67%	-2.33%
	Deep-CNN[40]	66.40%	-7.60%
	CNNV3[40]	73.58%	-0.42%
	AsicNet[40]	72.42%	-1.58%
	MobiExpressNet[51]	67.96%	-6.04%
	BReG-NeXt[52]	68.50%	-5.50%
	CONVNEXT_PRO	74.00%	-

Table 9 presents a comparison of the classification accuracies of different methods on the FER2013 dataset. As shown, earlier models such as MobileNetV2 and Inception V3 perform moderately, with accuracy rates of 38.85% and 39.06%, respectively, which are significantly lower than those of more advanced models available today. With the increasing complexity of model structures and the introduction of new optimization strategies, accuracy has notably improved. For instance, the VGG-FL model achieves an accuracy of 72.49%, and the CS-AttNet-Trans further boosts this to 73.37%. Notably, the CONVNEXT_PRO

model proposed in this paper achieves an accuracy of 74.00% on the FER2013 dataset, outperforming all other methods and validating its superior performance in expression recognition tasks. This significant improvement underscores the model's ability to handle the complexities of facial expression recognition, making it a promising candidate for further integration into real-time applications.

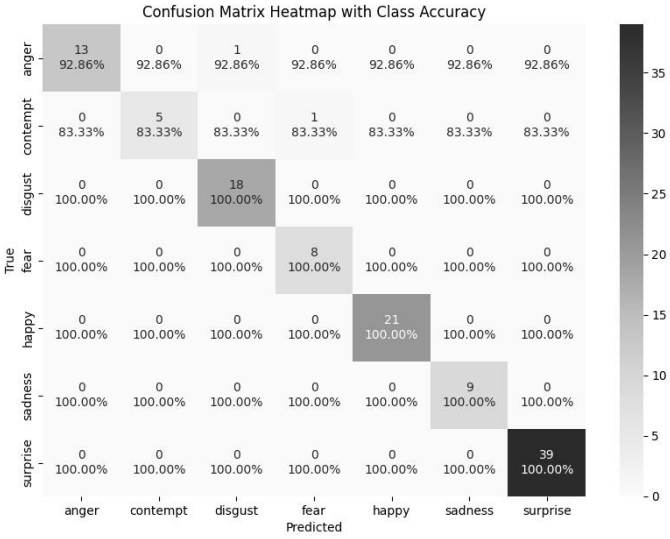

Fig. 8. Confusion Matrix of the Model on the CK+.

In addition, the confusion matrix on the FER2013 dataset further illustrates the specific performance of the CONVNEXT_PRO model in recognizing various emotions. It can be observed that the model excels in the "Happy" and "Surprise" categories, with F1 scores of 0.85 and 0.81, respectively, showcasing its strong classification capability for these emotions. However, the F1 scores for the "Fear" and "Disgust" categories are relatively lower, at 0.60 and 0.65, respectively, indicating that the model still has room for improvement when dealing with these more complex or less common emotion categories. These results suggest that further refinement of the model may be necessary to achieve better performance on these specific emotions. Nonetheless, the model's performance across most categories highlights its overall robustness and versatility. Overall, the CONVNEXT_PRO model performs excellently on the FER2013 dataset, achieving a good balance across most emotion categories and demonstrating performance enhancement through the introduced optimization mechanisms.

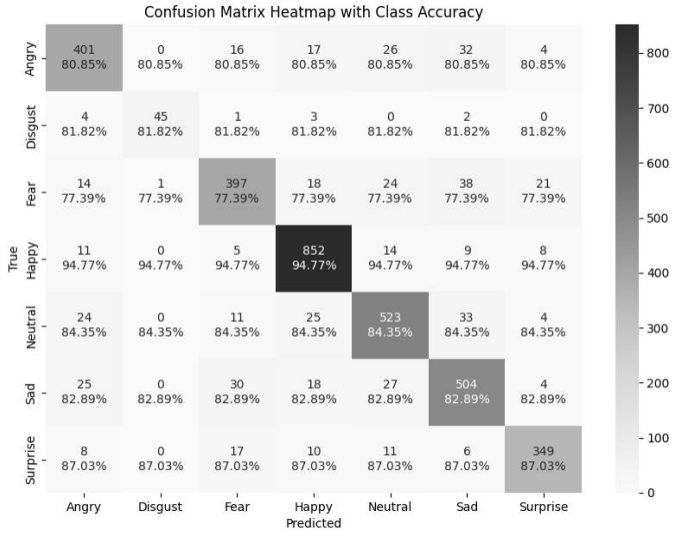

Fig. 9. Confusion Matrix of the Model on the FER2013.

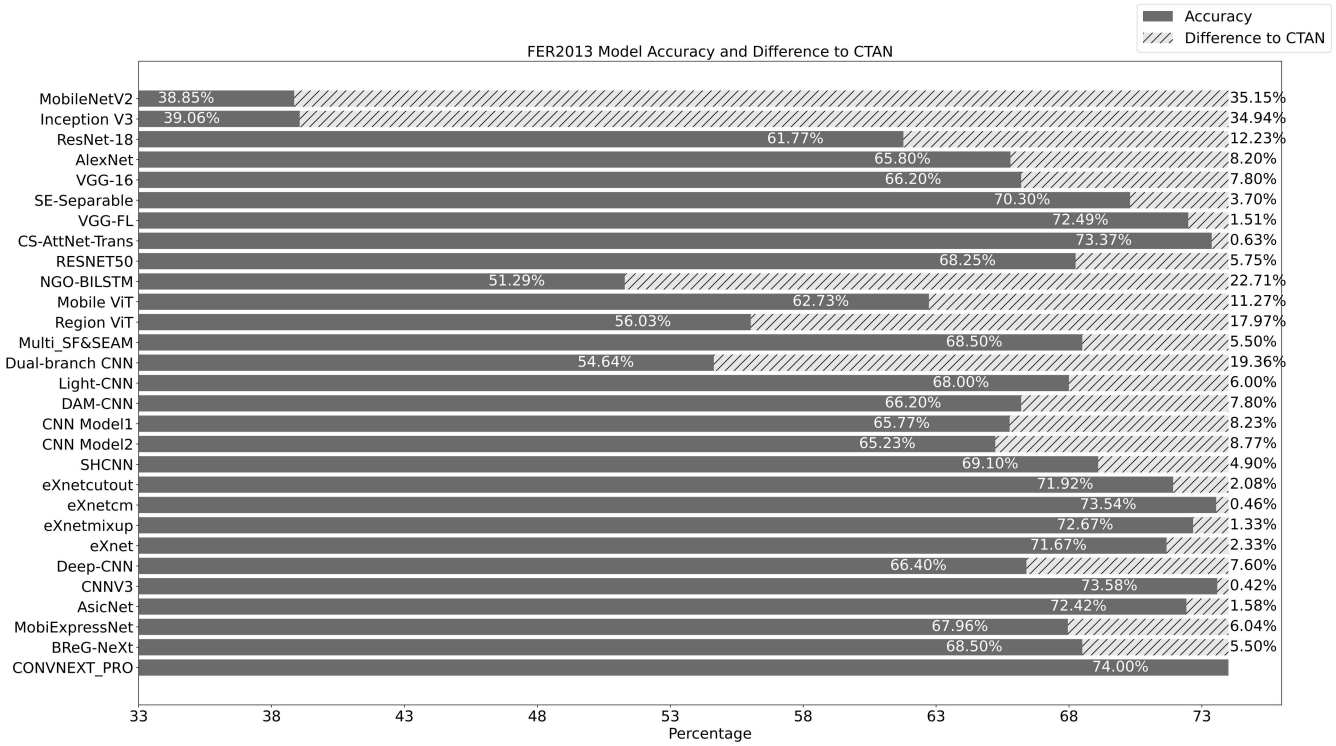

Fig. 10. Comparison of Classification Accuracy of Models on the FER2013.

VI. Conclusions

In this study, we propose the improved CTAN model for facial expression recognition, aimed at improving both classification accuracy and real-time performance. The model integrates Efficient Attention and Shuffle Attention modules to enhance feature extraction capability, while the EMA enhances training stability. Validated on multiple public and integrated datasets, the improved CTAN model achieved classification accuracies of 98.99% on CK+ and 74.00% on FER2013, outperforming state-of-the-art models by 0.69%-5.97% and 0.42%-5.50%, respectively, demonstrating superior accuracy and stability. We also analyzed the impact of hyperparameters like learning rate and batch size, providing insights for future optimization. In summary, the improved CTAN network, combining advanced architecture with effective optimization, offers a new approach to facial expression recognition. This research lays a foundation for applications in affective computing and human-computer interaction, with potential for further improvement in complex real-world environments. Future work will explore its adaptability and real-time performance.

References

[1] T. Deschamps-Berger, 'Emotion Recognition In Emergency Call Centers: The challenge of real-life emotions', in *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Nara, Japan: IEEE, Sep. 2021, pp. 1–5. doi: 10.1109/ACIIW52867.2021.9666308.

[2] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, 'Recurrent Neural Networks for Emotion Recognition in Video', in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, in ICMI '15. New York, NY, USA: Association for Computing Machinery, Nov. 2015, pp. 467–474. doi: 10.1145/2818346.2830596.

[3] J. Zeng, S. Shan, and X. Chen, 'Facial Expression Recognition with Inconsistently Annotated Datasets', in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 227–243. doi: 10.1007/978-3-030-01261-8_14.

[4] Y. Li, J. Zeng, S. Shan, and X. Chen, 'Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism', *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019, doi: 10.1109/TIP.2018.2886767.

[5] JIANG Yuewu, ZHANG Yujin, and SHI Jianxin, 'Expression Recognition Method Combining Key Points and Residual Network of Weight Allocation', *Computer Engineering and Applications*, vol. 58, no. 17, pp. 181–188, 2022.

[6] Z. CUI *et al.*, 'Facial Expression Recognition Combined with Improved VGGNet and Focal Loss', *Computer Engineering and Applications*, vol. 57, no. 19, pp. 171–178, 2021.

[7] Shen Hao, Meng Qinghao, and Liu Yinbo, 'Facial Expression Recognition by Merging Multilayer Features of Lightweight Convolutional Networks', *Laser & Optoelectronics Progress*, vol. 58, no. 6, pp. 148–155, 2021.

[8] Gao Jingwen, Cai Yongxiang, and He Zongyi, 'TP-FER: facial expression recognition method of tri-path networks based on optimal convolutional neural network', *Application Research of Computers*, vol. 38, no. 7, pp. 2213–2219, 2021, doi: 10.19734/j.issn.1001-3695.2020.07.0263.

[9] LI Chunhong and LU Yu, 'Facial expression recognition based on depthwise separable convolution', *Computer Engineering and Design*, vol. 42, no. 5, pp. 1448–1454, 2021, doi: 10.16208/j.issn1000-7024.2021.05.034.

[10] LIANG Huagang and LEI Yixiong, 'Expression Recognition with Separable Convolution Channel Enhancement Features', *Computer Engineering and Applications*, vol. 58, no. 2, pp. 184–192, 2022.

[11] ZHANG Peng, KONG Weiwei, and TENG Jinbao, 'Facial Expression Recognition Based on Multi-scale Feature Attention Mechanism', *Computer Engineering and Applications*, vol. 58, no. 1, pp. 182–189, 2022.

[12] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, 'Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition', *IEEE Trans. on Image Process.*, vol. 29, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.

[13] A. Dosovitskiy *et al.*, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', presented at the International Conference on Learning Representations, Oct. 2020. Accessed: Dec. 03, 2023. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, 'End-to-End Object Detection with Transformers', in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 213–229. doi: 10.1007/978-3-030-58452-8_13.

[15] Z. Liu *et al.*, 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[16] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, 'Training data-efficient image transformers & distillation through attention', in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 10347–10357. Accessed: Oct. 20, 2024. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html

[17] W. Sun *et al.*, 'Vicinity Vision Transformer', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12635–12649, Oct. 2023, doi: 10.1109/TPAMI.2023.3285569.

[18] Y. Cho, J. Yun, J. Kwon, and Y. Kim, 'Domain-Adaptive Vision Transformers for Generalizing Across Visual Domains', *IEEE Access*, vol. 11, pp. 115644–115653, 2023, doi: 10.1109/ACCESS.2023.3324545.

[19] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, 'Vision Transformer With Attentive Pooling for Robust Facial Expression Recognition', *IEEE Trans. Affective Comput.*, vol. 14, no. 4, pp. 3244–3256, Oct. 2023, doi: 10.1109/TAFFC.2022.3226473.

[20] S. Indolia, S. Nigam, R. Singh, V. K. Singh, and M. K. Singh, 'Micro Expression Recognition Using Convolution Patch in Vision Transformer', *IEEE Access*, vol. 11, pp. 100495–100507, 2023, doi: 10.1109/ACCESS.2023.3314797.

[21] L. Qin *et al.*, 'SwinFace: A Multi-Task Transformer for Face Recognition, Expression Recognition, Age Estimation and Attribute Estimation', *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2223–2234, Apr. 2024, doi: 10.1109/TCSVT.2023.3304724.

[22] H. Feng, W. Huang, D. Zhang, and B. Zhang, 'Fine-Tuning Swin Transformer and Multiple Weights Optimality-Seeking for Facial Expression Recognition', *IEEE Access*, vol. 11, pp. 9995–10003, 2023, doi: 10.1109/ACCESS.2023.3237817.

[23] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, 'A ConvNet for the 2020s', in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 11966–11976. doi: 10.1109/CVPR52688.2022.01167.

[24] K. Han *et al.*, 'A Survey on Vision Transformer', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[25] P. Xu, X. Zhu, and D. A. Clifton, 'Multimodal Learning With Transformers: A Survey', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113–12132, Oct. 2023, doi: 10.1109/TPAMI.2023.3275156.

[26] K. T. Chitty-Venkata, M. Emani, V. Vishwanath, and A. K. Somani, 'Neural Architecture Search for Transformers: A Survey', *IEEE Access*, vol. 10, pp. 108374–108412, 2022, doi: 10.1109/ACCESS.2022.3212767.

[27] 'MSSTNet: A Multi-Scale Spatio-Temporal CNN-Transformer Network for Dynamic Facial Expression Recognition'. Accessed: Oct. 20, 2024. [Online]. Available: https://arxiv.org/html/2404.08433v1

[28] YE Jihua, ZHU Jintai, JIANG Aiwen, LI Hanxi, and ZUO Jiali, 'Facial Expression Recognition: A Survey', *Journal of Data Acquisition and Processing*, vol. 35, no. 1, pp. 21–34, 2020, doi: 10.16337/j.1004-9037.2020.01.002.

[29] M. Jampour and M. Javidi, 'Multiview Facial Expression Recognition, A Survey', *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2086–2105, Oct. 2022, doi: 10.1109/TAFFC.2022.3184995.

[30] G. Zhao, H. Yang, and M. Yu, 'Expression Recognition Method Based on a Lightweight Convolutional Neural Network', *IEEE Access*, vol. 8, pp. 38528–38537, 2020, doi: 10.1109/ACCESS.2020.2964752.

[31] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, 'A ConvNet for the 2020s', Mar. 02, 2022, *arXiv*: arXiv:2201.03545. Accessed: Mar. 25, 2024. [Online]. Available: http://arxiv.org/abs/2201.03545

[32] Yang Xu and Shang Zhenhong, 'Facial Expression Recognition Based on Improved AlexNet', *Laser & Optoelectronics Progress*, vol. 57, no. 14, pp. 243–250, 2020.
[33] S. Minaee, M. Minaei, and A. Abdolrashidi, 'Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network', *Sens.*, vol. 21, no. 9, Art. no. 9, Jan. 2021, doi: 10.3390/s21093046.
[34] M. Jeong and B. C. Ko, 'Driver's Facial Expression Recognition in Real-Time for Safe Driving', *Sens.*, vol. 18, no. 12, Art. no. 12, Dec. 2018, doi: 10.3390/s18124270.
[35] M. Patil and S. Veni, 'Driver Emotion Recognition for Enhancement of Human Machine Interface in Vehicles', in *2019 International Conference on Communication and Signal Processing (ICCSP)*, Apr. 2019, pp. 0420–0424. doi: 10.1109/ICCSP.2019.8698045.
[36] B. Verma and A. Choudhary, 'A Framework for Driver Emotion Recognition using Deep Learning and Grassmann Manifolds', in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Nov. 2018, pp. 1421–1426. doi: 10.1109/ITSC.2018.8569461.
[37] J.-H. Kim, B.-G. Kim, P. P. Roy, and D.-M. Jeong, 'Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure', *IEEE Access*, vol. 7, pp. 41273–41285, 2019, doi: 10.1109/ACCESS.2019.2907327.
[38] X. Liu and F. Zhou, 'Improved curriculum learning using SSM for facial expression recognition', *Vis Comput*, vol. 36, no. 8, pp. 1635–1649, Aug. 2020, doi: 10.1007/s00371-019-01759-7.
[39] S. L. Happy, A. Dantcheva, and F. Bremond, 'Expression recognition with deep features extracted from holistic and part-based models', *Image and Vision Computing*, vol. 105, p. 104038, Jan. 2021, doi: 10.1016/j.imavis.2020.104038.
[40] S. Saurav, A. K. Saini, R. Saini, and S. Singh, 'Deep learning inspired intelligent embedded system for haptic rendering of facial emotions to the blind', *Neural Computing and Applications*, vol. 34, no. 6, pp. 4595–4623, Mar. 2022, doi: 10.1007/s00521-021-06613-3.
[41] GAO Hongxia and GAO Wei, 'Facial Expression Recognition Integrating Key Point Attributes and Attention Representation', *Computer Engineering and Applications*, vol. 59, no. 3, pp. 118–126, 2023.
[42] A. P. Fard and M. H. Mahoor, 'Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild', *IEEE Access*, vol. 10, pp. 26756–26768, 2022, doi: 10.1109/ACCESS.2022.3156598.
[43] J. Zhong, T. Chen, and L. Yi, 'Face expression recognition based on NGO-BILSTM model', *Front. Neurorobot.*, vol. 17, p. 1155038, Mar. 2023, doi: 10.3389/fnbot.2023.1155038.
[44] S. Mehta and M. Rastegari, 'MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer', *ArXiv*, Oct. 2021, Accessed: Apr. 09, 2024. [Online]. Available: https://www.semanticscholar.org/paper/MobileViT%3A-Light-weight%2C-General-purpose%2C-and-Mehta-Rastegari/da74a10824193be9d3889ce0d6ed4c6f8ee48b9e
[45] R. Chen, R. Panda, and Q. Fan, 'RegionViT: Regional-to-Local Attention for Vision Transformers', presented at the International Conference on Learning Representations, Apr. 2022. Accessed: Apr. 09, 2024. [Online]. Available: https://research.ibm.com/publications/regionvit-regional-to-local-attention-for-vision-transformers
[46] J. Shao and Y. Qian, 'Three convolutional neural network models for facial expression recognition in the wild', *Neurocomputing*, vol. 355, pp. 82–92, Aug. 2019, doi: 10.1016/j.neucom.2019.05.005.
[47] S. Xie, H. Hu, and Y. Wu, 'Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition', *Pattern Recognition*, vol. 92, pp. 177–191, Aug. 2019, doi: 10.1016/j.patcog.2019.03.019.
[48] A. Agrawal and N. Mittal, 'Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy', *Vis Comput*, vol. 36, no. 2, pp. 405–412, Feb. 2020, doi: 10.1007/s00371-019-01630-9.
[49] S. Miao, H. Xu, Z. Han, and Y. Zhu, 'Recognizing Facial Expressions Using a Shallow Convolutional Neural Network', *IEEE Access*, vol. 7, pp. 78000–78011, 2019, doi: 10.1109/ACCESS.2019.2921220.
[50] M. N. Riaz, Y. Shen, M. Sohail, and M. Guo, 'eXnet: An Efficient Approach for Emotion Recognition in the Wild', *Sensors*, vol. 20, no. 4, Art. no. 4, Jan. 2020, doi: 10.3390/s20041087.
[51] S. F. Cotter, 'MobiExpressNet: A Deep Learning Network for Face Expression Recognition on Smart Phones', in *2020 IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2020, pp. 1–4. doi: 10.1109/ICCE46568.2020.9042973.
[52] B. Hasani, P. S. Negi, and M. H. Mahoor, 'BReG-NeXt: Facial Affect Computing Using Adaptive Residual Networks With Bounded Gradient', *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 1023–1036, Apr. 2022, doi: 10.1109/TAFFC.2020.2986440.