Research and Implementation of Facial Expression Recognition Algorithm Based on Machine Learning

Xinjiu Xie*, Jinxue Huang
School of Modern Information Industry, Guangzhou College of Commerce, Guangzhou 511363, China

***Abstract*—Traditional information security management methods can provide a degree of personal information protection but remain vulnerable to issues such as data breaches and password theft. To bolster information security, facial expression recognition offers a promising alternative. To achieve efficient and accurate facial expression recognition, we propose a lightweight neural network algorithm called T-SNet (Teacher-Student Net). In our approach, the teacher model is an enhanced version of ResNet18, incorporating fine-grained feature extraction modules and pre-trained on the MS-Celeb-1M facial dataset. The student model uses the lightweight convolutional neural network ShuffleNetV2, with the model's accuracy further improved by optimizing the distillation loss function. This design carefully considers the key features of facial expressions, determines the most effective extraction techniques, and classifies and recognizes these features. To evaluate the performance of our algorithm, we conducted comparative experiments against state-of-the-art facial expression recognition methods. The results show that our approach outperforms existing methods in both recognition accuracy and efficiency.**

***Keywords*—Facial expression; expression recognition; convolutional neural network; deep learning**

I. Introduction

Facial Expression Recognition (FER) technology has become increasingly prevalent across various fields, significantly enhancing human-computer interaction and automation systems. In healthcare, FER is utilized to diagnose and monitor mental health conditions by analyzing patients' emotional states. For example, individuals with depression or anxiety often display specific emotional characteristics that FER can help clinicians identify, leading to more personalized treatment plans. In marketing and retail, FER is employed to assess customer satisfaction and engagement by observing their reactions to products and advertisements. Businesses can use this technology to analyze emotional changes during shopping, allowing them to optimize product displays and advertising strategies, ultimately increasing conversion rates. Security systems leverage FER to detect suspicious behaviors and potential threats based on facial cues. In public spaces and critical facilities, FER can monitor the emotional state of crowds in real-time, quickly identifying abnormal behaviors to prevent potential security risks. In education, FER assists in evaluating students' comprehension and engagement during remote learning. By analyzing students' facial expressions during classes, teachers can better understand their attention levels and emotional states, allowing them to adjust teaching methods for improved educational outcomes.

The significance of FER lies in its ability to provide deeper insights into human emotions and intentions, which is crucial for enhancing communication and interaction between humans and machines. By accurately interpreting facial expressions, systems can respond more appropriately to users' needs, thereby improving user experience and efficiency. For instance, in customer service, FER-equipped automated systems can detect signs of customer frustration and promptly offer assistance, leading to increased satisfaction. These systems can analyze facial expressions in real-time during interactions with service representatives, immediately alerting the representative to take appropriate action when signs of confusion or dissatisfaction are detected, thus improving service quality.

Neural networks, especially deep learning models, have been instrumental in advancing facial expression recognition (FER). These models can automatically learn and extract complex features from facial images, outperforming traditional methods that rely on handcrafted features. For example, Convolutional Neural Networks (CNNs) are particularly effective at recognizing subtle facial expressions by capturing the spatial hierarchy of facial features. Through multiple convolutional and pooling layers, CNNs create hierarchical feature representations from raw images, which are essential for distinguishing various facial expressions. Moreover, Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) enhance FER by incorporating temporal dynamics, enabling the analysis of expression sequences over time. These networks are well-suited for handling time-series data, as they capture dependencies between different time points—critical for understanding the dynamic nature of expression changes. The integration of neural networks into FER systems significantly enhances their accuracy and robustness, making these systems more reliable and practical for real-world applications. This technological progress expands the potential applications of facial expression recognition across various domains, driving further innovation in human-computer interaction technologies.

Section II introduces the current development status, and Section III presents the methods we propose. The superiority of our proposed method has been demonstrated through experiments in Section IV. Finally, a summary and outlook were made in Section V.

II. Literature Review

In this section, we introduce the development of recognition algorithms, and facial expression recognition algorithms, and summarize the research gaps.

*Corresponding Author.

A. Recognition Algorithm

With the advancement of technology, recognition algorithms have become increasingly sophisticated. These algorithms are not only prevalent in the field of image processing but are also applied across various other domains. Li et al. [1] proposed a method employing a differential evolution algorithm to optimize convolutional neural network (CNN) parameters for music emotion recognition tasks. Chen et al. [2] addressed the issues of slow convergence and weak generalization capability of CNNs in-vehicle feature recognition by proposing an improved bee colony algorithm to optimize CNN-based vehicle recognition strategies (ibsa-cnn). Li et al. [3] introduced a super-automatic algorithm combining non-local convolution and three-dimensional convolutional neural networks to address the shortcomings of missing critical feature information when processing long-time series video behavior features. Shi et al. [4] utilized data collected from multiple sensors measuring the Earth's magnetic field and employed a one-dimensional convolutional neural network algorithm for gesture recognition. Mao et al. [5] selected the Mel-frequency cepstral coefficient (MFCC) and filter bank (Fbank) as feature parameters to recognize English speech. Zhou et al. [6] proposed a method for lip print recognition based on convolutional neural networks.

In summary, recognition methods based on convolutional neural networks have been applied in a wide range of fields. Therefore, with the continuous development of technology, we can achieve more accurate facial expression recognition based on neural networks.

B. Facial Expression Recognition Algorithm

Currently, many scholars are dedicated to researching facial expression recognition algorithms. Zheng et al. [7] proposed a facial expression recognition method called TransformerKNN (TKNN), which integrates information about the state of the eyebrows and eyes in scenarios where the face is partially covered by a mask. Dong et al. [8] introduced a basic center regularization term based on the variance between basic centers to ensure that the learned expression features possess adequate discriminative capability. Wang et al. [9] proposed an Expression Complementary Disentanglement Network (ECDNet), which accomplishes the Facial Expression Disentanglement (FED) task during the face reconstruction process to handle all facial attributes in the disentanglement process. Yan et al. [10] introduced a new neonatal facial expression database for pain analysis. Win et al. [11] proposed a method for synthesizing complex facial expression images from learned expression representations without specifying emotion labels as input. Naveen et al. [12] addressed the problem of facial expression recognition under occlusion and proposed a robust facial expression recognition framework.

In summary, most current practice systems design a common practice path that all students must follow, lacking personalization. Due to the absence of necessary feedback, students' enthusiasm for practice is low, leading to low system utilization and poor results.

C. Research Gaps

The intelligent tutoring system proposed in this paper must address the following key challenges:

1) Effectiveness in capturing subtle differences: Current models struggle to effectively capture the nuanced differences between similar facial expressions, which is crucial for accurate facial expression recognition.

2) Complexity of recognition algorithms: Many existing facial expression recognition algorithms are overly complex due to their attempt to capture a wide range of features. This complexity often hinders performance, highlighting the need for a more lightweight model that can enhance recognition accuracy without unnecessary computational overhead.

These challenges are essential to consider in the development of facial expression recognition algorithms. The following sections of this paper will explore how deep learning methods can be utilized to achieve precise facial expression recognition while addressing these challenges.

III. Proposed Method

Our facial expression recognition model, designed using knowledge distillation, consists of a teacher model and a student model. The teacher model is an enhanced version of ResNet18, incorporating fine-grained feature extraction modules to capture deeper, multi-scale, and more detailed facial expression features. The student model utilizes the lightweight ShuffleNetV2 network, which maintains high recognition accuracy while being more computationally efficient.

The student model is trained and its parameters are updated by optimizing a distillation loss function. This function combines the probability distributions output by both the teacher and student models, enabling effective knowledge transfer. Through this process, the student model captures the essential knowledge from the teacher model, achieving high performance while remaining lightweight. Distillation training, therefore, facilitates the transfer of knowledge from a complex, high-performance model to a more efficient, lightweight model. The structure of this lightweight facial expression recognition network based on knowledge distillation is illustrated in Fig. 1.

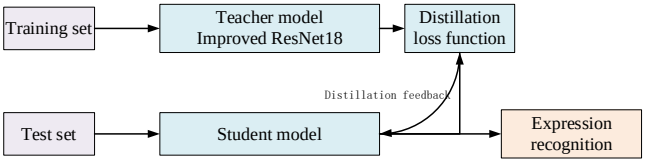

Fig. 1. Lightweight facial expression recognition network based on knowledge distillation.

In conclusion, our approach effectively tackles the challenge of distinguishing subtle differences between similar facial expressions while addressing the need for a more efficient and compact model. By utilizing knowledge

distillation, we transfer the strengths of the teacher model to the student model, ensuring that the latter maintains high accuracy while being computationally efficient and suitable for real-time applications. This method offers a balanced solution, combining the depth and robustness of complex models with the speed and efficiency of lightweight models, making it an ideal choice for facial expression recognition tasks.

A. Fine-grained Feature Extraction Module (FFE)

To tackle the challenge of small intra-class variations and the difficulty in extracting features from different facial expression categories, we designed a fine-grained feature extraction module inspired by Res2Net [13]. This module enables the extraction of multi-scale features at a fine-grained level from critical facial regions within a single basic block, as illustrated in Fig. 2.

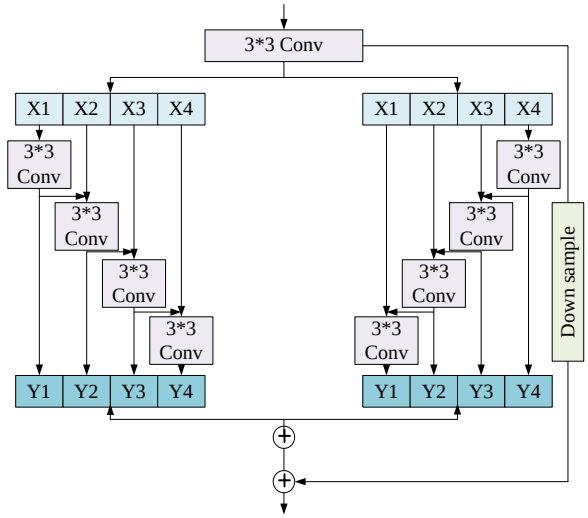

Fig. 2. Fine-grained Feature Extraction module (FFE).

Our module focuses on capturing subtle differences within the same category of facial expressions by emphasizing key facial areas. By leveraging the Res2Net-inspired architecture, the module enhances the model's ability to discern intricate details crucial for accurate facial expression recognition. This approach improves the model's capacity to distinguish between expressions with minimal variation, thereby increasing the overall accuracy and robustness of the recognition system.

The fine-grained feature extraction module is seamlessly integrated into our knowledge distillation framework, enhancing the student model's ability to learn from the teacher model. By including this module, we ensure that the lightweight student model retains the high-resolution, multi-scale features essential for precise facial expression analysis. This integration leads to a more effective and efficient facial expression recognition system, well-suited for real-world applications with limited computational resources.

In summary, the fine-grained feature extraction module effectively addresses the challenge of feature extraction in facial expression recognition by improving the system's ability to capture and analyze subtle facial cues. This innovation, coupled with our knowledge distillation approach, advances the development of more accurate and efficient facial expression recognition models.

The fine-grained feature extraction module designed in this paper utilizes a symmetrical structure to learn multi-scale features within a single basic block. This design ensures that feature subsets from both preceding and succeeding stages contain richer scale information. Specifically, after applying a 3×3 convolution to the input feature map X, the map is evenly divided along the channel axis into n subsets, denoted as X_i, where $i \in \{1, 2, \ldots, n\}$. Each subset X_i retains the same spatial dimensions as the original feature map X but with $1/n$ of the channels.

These subsets are then processed using a 3×3 convolution, denoted as $P_i^p(\cdot)$, where $p \in \{left, right\}$ indicates the position within the symmetrical structure. The output of each convolution operation, Y_i^p, can be expressed using Eq. (3). This method enables the extraction of fine-grained, multi-scale features, enhancing the model's capability to capture detailed facial expression characteristics.

$$Y_i^{left} = \begin{cases} P_i^{left}(X_i), & i = 1 \\ P_i^{left}(X_i + Y_{i-1}^{left}), & 1 < i \le n \end{cases} \quad (1)$$

$$Y_i^{right} = \begin{cases} P_i^{right}(X_i), & i = 1 \\ P_i^{right}(X_i + Y_{i-1}^{right}), & 1 < i \le n \end{cases} \quad (2)$$

$$Y_i = Y_i^{left} + Y_i^{right} \quad (3)$$

According to Eq. (1), each operation of $P_i^{left}(\cdot)$ captures feature from all subsets $\{X_j, \leqslant i\}$. Conversely, Eq. (2) shows that each operation of $P_i^{right}(\cdot)$ captures feature from subsets $\{X_j, \geqslant j \geqslant i\}$. Each operation involves applying a 3×3 convolution to a split feature X_i. As a result, the output Y_i^{left} has a larger receptive field compared to $\{Y_k, <i\}$, while Y_i^{right} has a larger receptive field compared to $\{Y_k, k>i\}$. This design enables the extraction of comprehensive multi-scale features from both preceding and succeeding stages, enhancing the model's ability to analyze fine-grained details.

Each output Y_i^p contains facial features at varying scales and quantities. To achieve a richer diversity of multi-scale features, all Y_i^p are aggregated along the channel axis, integrating the fine-grained details from each subset. While increasing n enriches the feature representation, it also raises computational costs. To strike an optimal balance between performance and efficiency, we set n = 4.

This design ensures that the fine-grained feature extraction module effectively captures multi-scale facial features crucial for accurate and robust facial expression recognition. By aggregating features from different scales, the model enhances its ability to detect subtle variations in facial expressions, thereby improving accuracy and reliability for practical applications.

Setting n = 4 allows us to balance feature richness with computational efficiency, making our method practical for real-world applications where both accuracy and speed are

critical. This design choice demonstrates our commitment to optimizing performance while maintaining computational feasibility.

By utilizing this approach, we ensure that both initial and subsequent feature subsets capture detailed multi-scale information. This enhances the model's ability to detect fine-grained variations in facial expressions, resulting in more accurate and robust facial expression recognition.

Fig. 2 provides a detailed illustration of the symmetrical structure of our fine-grained feature extraction module. The figure demonstrates how the module employs symmetrical convolutions to process feature maps, enhancing the extraction of multi-scale features.

This innovative module integrates seamlessly into our facial expression recognition framework, allowing the lightweight student model to effectively learn complex, multi-scale features from the teacher model. As a result, the system achieves high performance while maintaining computational efficiency, making it well-suited for deployment in real-world applications with limited resources.

B. The Construction of Teacher Model and Student Model

We selected ResNet18 for its robust feature extraction and high recognition accuracy as the backbone of our teacher model. To further enhance its performance, we integrated a fine-grained feature extraction module designed to capture intricate facial details, especially from critical regions such as the eyes and mouth. This improvement enables the teacher model to provide more precise supervision and guidance to the student model. The enhanced teacher model, as illustrated in Fig. 3, processes facial images of size 224×224×3. The image first passes through standard convolutional layers, followed by batch normalization, ReLU activation, and max pooling, producing feature maps of size 112×112×64. Two residual blocks then generate feature maps of size 28×28×128. These are subsequently processed through two sequential fine-grained feature extraction modules, resulting in multi-scale, high-resolution features of size 224×224×3. Finally, these features are fed into fully connected layers for facial expression classification. This configuration allows the teacher model to effectively capture and leverage detailed facial features, ensuring that the student model benefits from enhanced guidance during training.

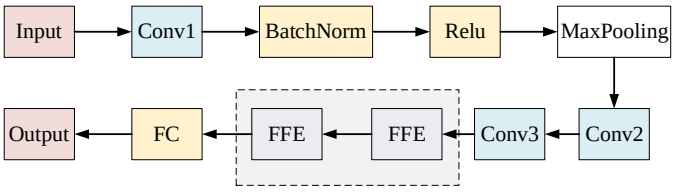

Fig. 3. Teacher model.

We selected ShuffleNetV2 as the backbone for the student model due to its computational efficiency and lightweight design. This network excels in real-time applications thanks to its use of depth wise separable convolutions, which significantly reduce model parameters and speed up computations. This makes ShuffleNetV2 particularly well-suited for deployment on edge devices for real-time facial expression recognition.

The ShuffleNetV2 architecture, depicted in Fig. 4, processes 224×224×3 facial images. It starts with a 3×3 standard convolutional layer followed by max pooling. The network then progresses through several stages, including down sampling blocks and basic blocks that utilize depth-wise separable convolutions. After these stages, the model passes through fully connected layers to produce the final facial expression classification. This design balances efficiency with performance, making it ideal for practical applications requiring real-time processing.

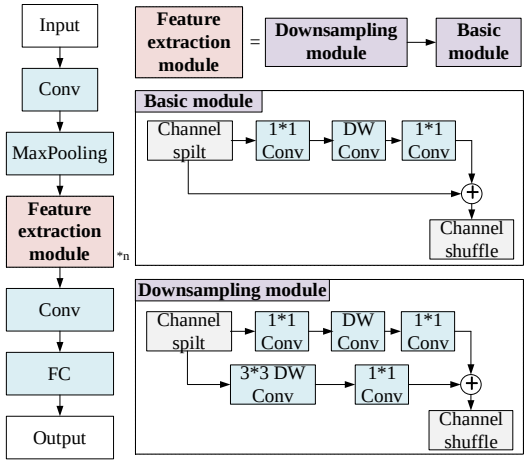

Fig. 4. Student model.

C. Facial Expression Recognition Network Based on Knowledge Distillation

Knowledge distillation is a widely utilized model compression technique that enhances the performance and accuracy of a lightweight student model by leveraging the expertise of a larger, more complex teacher model. This technique enables the student model to benefit from both the "hard labels" of the dataset and the "soft labels" provided by the teacher model's probabilistic outputs. The essence of knowledge distillation is to transfer the nuanced insights and detailed feature representations learned by the teacher model to the student model. In the T-SNet framework, this is achieved through a dual-branch network architecture, which includes both the teacher and student models. As illustrated in Fig. 5, the teacher model provides comprehensive guidance by generating rich, informative soft labels that the student model uses for training. This setup ensures that the student model captures not just the direct class labels but also the underlying distributions and patterns recognized by the teacher, leading to improved performance and accuracy in a more compact and efficient model.

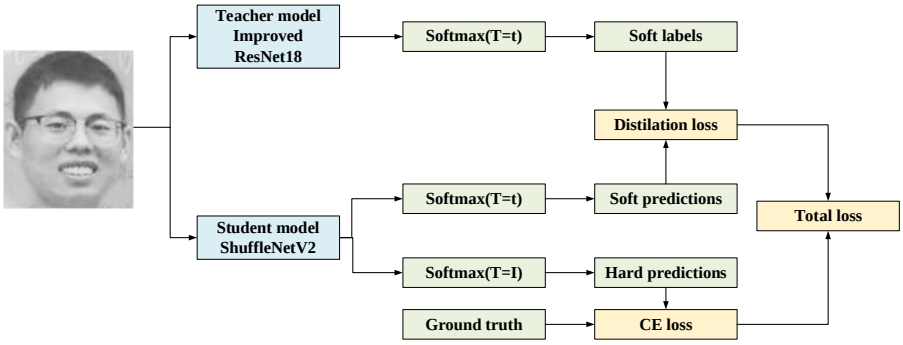

Fig. 5. Knowledge distillation model.

In the T-SNet framework, the dual-branch architecture operates as follows:

1) Teacher model branch: This branch performs global fine-grained feature extraction and classification. It leverages its complex architecture to capture detailed facial features and nuances.

2) Student model branch: This branch trains the student model using distilled features from the teacher model. It adapts these features to its more compact and efficient design for facial expression classification.

The framework employs a joint loss function to guide the training of both models. This ensures the student model retains essential knowledge from the teacher model, enhancing efficiency and performance while reducing computational requirements. This approach is especially suitable for real-time facial expression recognition where resources are constrained.

During the knowledge distillation process, facial expression images are simultaneously inputted into both the teacher and student models for feature extraction. The teacher model generates a Softmax distribution at a high temperature T, which serves as soft labels. Ultimately, the classifier outputs from both the teacher and student networks determine the expression categories. Introducing a Softmax temperature function helps smooth the probability distribution of predicted expressions, thereby providing additional class-specific feature information inherent to the teacher network.

D. Loss Function

During the knowledge distillation training process, the loss function L_{kd} for the student model comprises two key components: L_{soft} and L_{hard}, corresponding to learning from "soft labels" and "hard labels", respectively. L_{soft} uses Kullback-Leibler (KL) divergence to measure the difference between the predictions of the teacher model and the student model. Specifically, it calculates the cross-entropy loss between the Softmax output of the student model under the same temperature T and the soft labels provided by the teacher model. The L_{hard} component uses the cross-entropy loss function to compute the difference between the predictions of the student model and the true facial expression labels. The final loss for optimization is the sum of these two parts.

The L_{soft} loss function is formulated as follows:

$$\mathrm{Y_i} = \mathrm{Y_i^{left}} + \mathrm{Y_i^{right}} \tag{4}$$

where p_i^T denotes the Softmax output of the teacher model for class i at temperature T, and q_i^T denotes the Softmax output of the student model for class i at temperature T. The expressions for p_i^T and q_i^T are given by:

$$\mathrm{p_i^T} = \frac{\mathrm{e^{v_i/T}}}{\sum_k^N \mathrm{e^{v_k/T}}} \tag{5}$$

$$\mathrm{q_i^T} = \frac{\mathrm{e^{z_i/T}}}{\sum_k^N \mathrm{e^{z_k/T}}} \tag{6}$$

where v_i is the raw logits from the output layer of the teacher model, z_i is the raw logits from the output layer of the student model, and N is the total number of labels.

The L_{hard} loss function is defined as Eq. (7):

$$\mathrm{L_{hard}} = -\sum_i^N \mathrm{c_i log(q_i^1)} \tag{7}$$

$$\mathrm{q_i^1} = \frac{\mathrm{e^{z_i}}}{\sum_k^N \mathrm{e^{z_k}}} \tag{8}$$

$$\mathrm{Totleloss} = \alpha \mathrm{T^2 L_{soft}} + (1-\alpha)\mathrm{L_{hard}} \tag{9}$$

where q_i^1 is the Softmax output of the student model for class i without temperature scaling, given by Eq. (8). The total loss $Totleloss$ is a combination of L_{soft} and L_{hard}.

where α is a balancing factor between the distillation loss and the cross-entropy loss.

Because the gradient magnitude of soft labels is scaled down by $1/T^2$, L_{soft} is multiplied by T^2 to ensure consistency between the true label values and the probability distribution of the teacher model. The temperature T and the coefficient α are hyperparameters that need to be empirically determined, which we explore further in subsequent sections through ablation experiments.

IV. Experiment and Verification

In this chapter, we verify the reliability and validity of the proposed method through experiments.

A. Experimental Environment

This study validated the algorithm's effectiveness using an environment comprising an 11th Gen Intel(R) Core (TM) i7-11700K @ 3.60GHz CPU with 32.0 GB of RAM, running

Python 3.6. For facial expression recognition experiments, a dataset was created from randomly captured images manually annotated. The dataset consists of 8500 images, divided into training, testing, and validation sets in a 6:2:2 ratio. The testing set comprises 1700 three-dimensional images with faces, randomly grouped into five sets of 340 images each. Each group of images was processed using the system for facial expression recognition.

B. Evaluation Parameter

First, the data set is aligned to the face [14], then the pre-processed face image is input to the teacher model training, the teacher model parameters are saved, and then the teacher model is trained by knowledge distillation to assist the student model [15]. Finally, the accuracy and the number of model parameters are calculated on the test set to verify the performance of the model.

We choose number of Accuracy, Precision, Accuracy, and the cords FLOPs as evaluation index, accurate calculation method is as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{10}$$

$$\text{ACC} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{11}$$

Where TP represents true positives, meaning instances classified as positive that are actually positive; FN represents false negatives, where instances classified as negative but are actually positive; FP represents false positives, where instances classified as positive but are actually negative; and TN represents true negatives, where instances classified as negative that are actually negative.

AUC (Area Under the Curve) is a metric used to evaluate the performance of binary classification models, indicating the probability that the model ranks a randomly chosen positive example higher than a randomly chosen negative example. Firstly, True Positive Rate (TPR) refers to the proportion of actual positive cases correctly predicted as positive, calculated as $\frac{TP}{TP+FN}$. False Positive Rate (FPR) is the proportion of actual negative cases incorrectly predicted as positive, calculated as $\frac{FP}{TN+FP}$. The ROC (Receiver Operating Characteristic) curve plots TPR against FPR, with AUC representing the area under this curve. A higher AUC indicates better model performance in distinguishing between positive and negative examples.

Parameter count is a metric measuring model complexity and storage requirements, calculated based on the model's architecture and layer types [16]. Parameters are computed using the model.parameters function in PyTorch. Computational complexity (FLOPs) measures the total number of floating-point operations performed by the model during inference or training. Additionally, this study evaluates inference time to assess model processing speed on embedded devices.

During training, stochastic gradient descent (SGD) optimizes the loss function with the following hyperparameters: weight decay of 1e-4, momentum of 0.9, dropout rate of 0.5, and initial learning rate of 0.01. Learning rate adjustments are dynamically managed using ReduceLROnPlateau, decreasing the learning rate by a factor of 10 if the loss does not decrease after 3 epochs. The total training epochs are set to 150 with a batch size of 128. The best model is saved after 150 epochs based on the highest achieved accuracy.

C. Test and Evaluation

We set up ablation experiments to demonstrate the superiority of our parameter selection. Table I shows the comparison of results of different parameters. In the knowledge distillation process, the distillation temperature T and α are set to 8 and 0.4 respectively. The ablation experiment verified the best effect of this parameter setting.

TABLE I. ABLATION EXPERIMENT

T	α	**precision**
4	0.4	0.899
6	0.2	0.893
6	0.4	0.988
6	0.6	0.906
8	0.4	0.891

We tested the results of teacher-only model training, student-only model training and the results of the combined training of teacher model and student model after adding knowledge distillation on FER2013Plus and RAF-DB datassets. To ensure the fairness of experimental results, the same hyperparameters were used for both teacher model and student model. In order to evaluate the performance of knowledge distillation algorithm, experiments were conducted on FER2013Plus and RAF-DB datasets respectively to test the accuracy and number of parameters of the improved teacher model, student model and distilled student model. Table II shows the test results of teacher model and student model on FER2013Plus and self-made datasets. The faculty model is named EFF-ResNet18(Enhanced Fine- ResNet18), the student model is named ShuffleNetV2, and the distilled student model is named T-SNet.

TABLE II. TEACHER MODEL AND STUDENT MODEL TEST RESULTS

Datasets	Model	parameters	Accuracy
FER2013Plus	EFF-ResNet18	11.8	86.23
	ShuffleNetV2	1.3	88.15
	T-SNet	1.2	91.62
Our datasets	EFF-ResNet18	12.3	88.57
	ShuffleNetV2	1.3	90.45
	T-SNet	1.2	94.69

According to Table II analysis, on the FER2013Plus dataset, knowledge distillation improved the accuracy of the student model from 86.23% to 91.62%. Similarly, on our proprietary dataset, performance enhancement was observed, increasing from 88.57% to 94.69%. Across both datasets, the distilled student model surpassed the teacher model in accuracy, while reducing parameter count by 10.60 million, indicating significant improvement. Experimental results

demonstrate that knowledge distillation enhances the generalization performance of the student model by transferring knowledge from the teacher model, achieving model lightweighting in the process. This highlights the effectiveness of training lightweight models with performance comparable to large-scale network models under supervision of feature extraction from large network models.

To assess the performance of T-SNet in facial expression recognition tasks, this study compared T-SNet with other mainstream algorithms including RCL-Net [17], ResNet18, TFEN [18], ECDNet, etc. Comparative results on our dataset are presented in Table III.

TABLE III. COMPARISON OF T-SNET AND SOTA METHODS

Models	Precision	Accuracy	FLOPs
TFEN	0.898	0.906	4.5
ResNet18	0.903	0.893	11.18
RCL-Net	0.901	0.891	6.4
ECDNet	0.897	0.903	3.5
Ours	0.931	0.988	1.3

Through comparison, it can be observed that the recognition accuracy of our proposed method on the data set is as high as 98.8%, which is much higher than the second method. This shows that the proposed method is superior to other SOTA methods in terms of accuracy. In addition, in terms of the number of parameters and the amount of computation, our method has the lowest parameter number of 1.3M. Our method has very good advantages in recognition rate and parameter number, which can be said to be a qualitative leap for the system with high real-time requirements. The experimental results show that the lightweight expression recognition model based on knowledge distillation proposed in this paper has strong competitiveness and can be used as one of the choices for a real-time facial expression recognition system.

TABLE IV. COMPARISON OF RECOGNITION EFFECT OF DIFFERENT EXPRESSIONS

Expression type	FER2013Plus	Our datasets
Angry	0.95	0.96
disgust	0.90	0.91
happy	0.98	0.99
sad	0.92	0.93
surprised	0.96	0.97
neutral	0.97	0.98
contempt	0.91	0.93

We respectively tested the recognition effects of different expressions on FER2013Plus and self-made data sets with our proposed method, as shown in Table IV. From the confusion matrix of the two data sets, it can be concluded that happy expression is the most easily recognized expression, mainly because happy data in the two data sets has the highest amount and more abundant features. Happiness is also highly recognizable in the real world. Neutral is easier to identify because some uncertain subtle expressions are generally marked as neutral during data set annotation. The recognition rates of surprised, angry and sad expressions declined successively, which is consistent with human visual characteristics.

In order to more intuitively verify the effectiveness of T-SNet, this paper uses GradCAM to generate visual activation diagram for visual analysis of T-SNet [19]. Among them, the data samples are from the FER2013Plus test set. The visualized result is shown in Fig. 6. T The first column is the original data sample, the second column is the ECDNet model activation map, and the third column is the T-SNet activation map. The visual activation graph can verify the importance of the network to the key areas of the image, and the brighter the color, the more important the content of the area is for the recognition of the network. Red indicates high activation and blue indicates low activation. The comparison between the second and third column shows that T-SNet after knowledge distillation can absorb the fine-grained feature extraction experience of the teacher model, and the extracted feature semantic information is richer, and the receptive field is larger than that of the baseline model. By focusing attention on important facial areas, T-SNet can more accurately identify expression categories, which verifies T-SNet's excellent expression recognition performance.

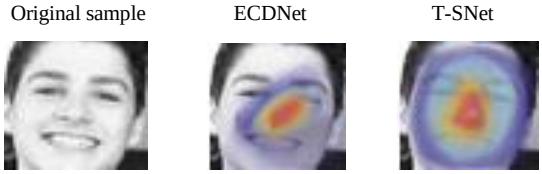

Fig. 6. Visual analysis.

V. CONCLUSION

We propose a lightweight algorithm for facial expression recognition based on neural networks, termed T-SNet. Addressing the challenge of deploying complex neural networks on smart terminals and the increasing demand for lightweight facial expression recognition models, we introduce T-SNet, a knowledge-distillation-based lightweight facial expression recognition network. We selected a refined ResNet18 with enhanced fine-grained feature extraction module as the teacher model backbone and ShuffleNetV2 as the student model backbone. By constructing distillation loss, we transfer rich classification information from the teacher model to the student model, leveraging the teacher model's experience to train the student model and improve its accuracy in facial expression recognition. The distilled student model, trained as a lightweight facial expression recognition model, is the outcome of our approach. Experimental results demonstrate that our proposed T-SNet model outperforms other mainstream facial recognition models in terms of accuracy and parameter efficiency. Despite the achievements in lightweight facial recognition, real-world scenarios still pose challenges due to factors such as environment, lighting, and occlusion.

While our method shows promising performance on current datasets, demonstrating excellent effectiveness and efficiency,

further validation in practical applications is essential. Future advancements in deep learning technology and dataset enhancements will likely expand the application of facial recognition into more domains. To ensure effective deployment in diverse applications, rigorous empirical research on feasibility and efficacy is needed. This will further validate the practical potential of our approach.

ACKNOWLEDGMENT

This study received funding from three sources: the Campus Level Quality Engineering Project , titled " Research on the Construction and Application of Smart Classrooms Supporting the Development of Advanced Thinking in the Digital Age", sponsored by Guangzhou College of Commerce (No.: 2024JXGG47); Higher Education Scientific Research Topic Project, titled "Research on the Construction of Digital Teaching Resources in Applied Universities under the Background of School-Enterprise Collaborative Education", sponsored by Office of Guangdong Provincial Leading Group for Education Science Planning (No.: 2023GXTK581); Guangdong Province Ordinary University Characteristic Innovation Project, titled "Research on Key Technologies for Campus Network Security Situation Awareness", sponsored by scientific research division of Guangdong Provincial Department of Education(No.: 2020KTSCX169).

REFERENCES

[1] J. Li, S. Soradi-Zeid, A. Yousefpour, D. Pan, “Improved differential evolution algorithm based convolutional neural network for emotional analysis of music data,” Applied Soft Computing, 153. 2024. https://doi.org/10.1016/j.asoc.2024.111262

[2] X. Chen, “Vehicle Feature Recognition via A Convolutional Neural Network with An Improved Bird Swarm Algorithm,” *Journal of Internet Technology*, *24*(2), 421–432. 2023. https://doi.org/10.53106/160792642023032402020

[3] J. Li, J. Liu, C. Li, F. Jiang, J. Huang, S. Ji, Y. Liu, “A hyperautomative human behaviour recognition algorithm based on improved residual network,” *Enterprise Information Systems*, *17*(10). 2023. https://doi.org/10.1080/17517575.2023.2180777

[4] B. Shi, X. Chen, Z. He, H. Sun, R. Han, “Research on Gesture Recognition System Using Multiple Sensors Based on Earth’s Magnetic Field and 1D Convolution Neural Network,” *Applied Sciences (Switzerland)*, *13*(9), 2023. https://doi.org/10.3390/app13095544

[5] C. Mao, S. Liu, “A Study on Speech Recognition by a Neural Network Based on English Speech Feature Parameters,” *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *28*(3), 679–684, 2024. https://doi.org/10.20965/jaciii.2024.p0679

[6] H. Zhou, “Lip Print Recognition Algorithm Based on Convolutional Network,” *Journal of Applied Mathematics*, *2023*. https://doi.org/10.1155/2023/4448861

[7] K. Zheng, L. Tian, Z. Li, H. Li, J. Zhang, “Incorporating eyebrow and eye state information for facial expression recognition in mask-obscured scenes,” *Electronic Research Archive*, *32*(4), 2745–2771. 2024. https://doi.org/10.3934/ERA.2024124

[8] R. Dong, K. M. Lam, “Bi-Center Loss for Compound Facial Expression Recognition,” *IEEE Signal Processing Letters*, *31*, 641–645, 2024. https://doi.org/10.1109/LSP.2024.3364055

[9] S. Wang, H. Shuai, L. Zhu, Q. Liu, “Expression Complementary Disentanglement Network for Facial Expression Recognition,” *Chinese Journal of Electronics*, *33*(3), 742–752, 2024. https://doi.org/10.23919/cje.2022.00.351

[10] J. Yan et al., “FENP: A Database of Neonatal Facial Expression for Pain Analysis. *IEEE Transactions on Affective Computing*, *14*(1), 245–254. 2023. https://doi.org/10.1109/TAFFC.2020.3030296

[11] S. S. K. Win, P. Siritanawan, K. Kotani, “Compound facial expressions image generation for complex emotions,” *Multimedia Tools and Applications*, *82*(8), 2023. 11549–11588. https://doi.org/10.1007/s11042-022-14289-7

[12] P. Naveen, “Occlusion-aware facial expression recognition: A deep learning approach,” *Multimedia Tools and Applications*, *83*(11), 32895–32921, 2024. https://doi.org/10.1007/s11042-023-17013-1

[13] S H Gao, M M Cheng, K Zhao, et al., “Res2net: A new multi-scale backbone architecture,” IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 652-662.

[14] R R Adyapady, B. Annappa, “A comprehensive review of facial expression recognition techniques,” Multimedia Systems, 29(1): 73-103, 2023.

[15] Xinchen Pan, Ling Qin, Xiaojian Yang, “Facial expression recognition in complex scenes based on multi-region detection network,” Data Acquisition and Processing, 38(06): 1422-1433, 2023.

[16] Yahui Nan, Qingyi Hua, “Occluded face expression recognition deep learning Progress in law research,” Application Research of Computers, 39 (2): 321-330, 2022.

[17] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, G. He, “Facial Expression Recognition Methods in the Wild Based on Fusion Feature of Attention Mechanism and LBP,” *Sensors*, *23*(9), 2023. https://doi.org/10.3390/s23094204

[18] J. Teng, D. Zhang, W. Zou, M. Li, D. J. Lee, “Typical Facial Expression Network Using a Facial Feature Decoupler and Spatial-Temporal Learning,” *IEEE Transactions on Affective Computing*, *14*(2), 1125–1137, 2023. https://doi.org/10.1109/TAFFC.2021.3102245

[19] R R Selvaraju, M Cogswell, A Das, et al., “Grad-cam: Visual explanations from deep networks via gradient-based localization,” IEEE international conference on computer vision. 618-626, 2017.