Children's Expression Recognition Based on Multi-Scale Asymmetric Convolutional Neural Network

Pengfei Wang, Xiugang Gong*, Qun Guo, Guangjie Chang, Fuxiang Du
School of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong, 255000, China

***Abstract*—This paper proposes a multi-scale asymmetric convolutional neural network (MACNN), specifically designed to tackle the challenges encountered by traditional convolutional neural networks in the realm of children's facial expression recognition. MACNN addresses problems like low accuracy from facial expression changes, poor generalization across datasets, and inefficiency in traditional convolution operations. The model introduces a multi-scale convolution layer for capturing diverse features, enhancing feature extraction and recognition accuracy. Additionally, an asymmetric convolutional layer is integrated to learn directional features, improving robustness and generalization in facial expression analysis. Post-training, this layer can revert to a standard square convolutional layer, optimizing efficiency for child expression recognition. Experimental results indicate that the proposed algorithm achieves a recognition accuracy of 63.35% on a self-constructed children's expression dataset, under the configuration of a GPU Tesla P100 with 16GB video memory. This performance exceeds all comparative algorithms and maintains efficient recognition. Furthermore, the algorithm attains a recognition accuracy of 78.26% on the extensive natural environment expression dataset RAF-DB, highlighting its robustness, generalization capability, and potential for practical application.**

***Keywords*—Children's expression recognition; convolutional neural network; multi-scale asymmetric convolutional neural network; asymmetric convolutional layers**

I. INTRODUCTION

Facial expressions are crucial in human communication. According to Mehrabian, facial expressions provide about 55% of emotional expression, with only about 7% conveyed through spoken words [1]. However, their complexity and variety pose a challenge for automatic recognition. With the rise of deep learning, Convolutional Neural Networks (CNNs) have become widely utilized in automatic expression recognition due to their strong feature extraction ability and adaptability, resulting in significant advancements in this field. Wen XY et al. used convolutional neural networks to extract deep semantic features and shallow geometric features, while introducing a channel self-attention mechanism to reduce the impact of occlusion and pose changes on expression recognition [2]. Ying He et al. recently proposed a novel multi-layer feature recognition algorithm based on a three-channel convolutional neural network (CNN), which significantly improved the accuracy of convolutional neural network expression recognition [3]. Cuiping Shi et al. proposed a facial expression recognition algorithm based on multi-branch cross-connected convolutional neural network (MBCC-CNN). Compared with traditional machine learning algorithms, the proposed algorithm can extract expression features more effectively [4]. Jung Hwan Kim et al. proposed customized visual geometry group-19 (CVGG-19), which combines the designs of visual geometry group (VGG), inception-v1, residual neural network (ResNet), and xception to improve expression recognition performance while reducing computational cost [5]. Yinggang He et al. developed a multi-branch attention convolutional neural network based on a multi-branch structure to recognize facial expressions, which is more efficient in extracting image features [6]. Qian Dong et al. proposed a VIT-based multi-scale Attention Learning network (MALN) that learns facial expression features in a multi-scale manner [7]. Aly Walaa et al. designed a deep convolutional neural network combining residual spatial channel attention (RSCA) and spatial Pyramid pooling (ASPP) to improve the expression recognition effect of the model for low-resolution images [8]. Chen Bin et al. proposed a residual rectified dense convolutional neural network, which linearly rectified the residual block through the activation function embedded in the convolutional layer to improve the model's ability to extract complex expression features [9]. Qi H et al. designed a Pyramid convolutional attention residual network (PCARNet) based on ResNet-18, which combines the pyramid convolution module and the improved convolutional attention mechanism to effectively extract expression features and achieve high-precision facial expression recognition [10]. Tataji KNK et al. proposed a cross-connected convolutional neural network (CC-CNN), which has been shown to be effective in extracting local and global facial features [11]. Kalsum T et al. proposed a new lightweight deep convolutional neural network (DCNN) model, which improved the recognition effect of facial expression while reducing the complexity of the model [12]. Liu Y et al. used three parallel multi-channel convolutional networks to learn and fuse local and global features from different facial regions, which enhanced the expression feature extraction ability of convolutional neural networks [13]. Mukhopadhyay et al. proposed a algorithm combining local binary pattern (LBP) and convolutional neural network. The LBP processed images were trained by CNN, which improved the efficiency of convolutional neural network for expression recognition [14]. Jing Li et al. combined LBP features and attention mechanism and achieved good results [15]. Saad Saeed et al. proposed an automated framework for face detection using CNN, which includes four convolutional layers and two hidden layers to improve expression recognition accuracy [16].

The aforementioned researches on facial expression recognition algorithms predominantly concentrate on adult facial expressions, whereas, in comparison, recognizing children's expressions holds greater practical significance. This is because children are in a stage where their language system is still developing, rendering it primarily reliant on facial expressions and behaviors to ascertain their emotional state [17].

Corresponding Author.

By analyzing children's emotions and intentions through facial expression recognition, researchers, parents, and educators can gain a deeper understanding of children's inner world. This approach can facilitate a deeper understanding of children's needs and emotional states among adults, while also enabling timely detection and effective handling of any issues or troubles faced by children. Although children's expression recognition faces many challenges, including difficulties in data acquisition, and ethical and privacy issues [18], some researchers have still achieved important research results in this field. Nagpal Shruti et al. developed a mean-supervised deep Boltzmann machine (msDBM) for the classification of children's expressions, marking the first application of a deep learning-based algorithm in this domain [19]. Manish Rathod et al. utilized seven distinct convolutional neural network (CNN) architectures for the task of recognizing children's facial expressions. Their results indicate that the 152-layer residual network (ResNet-152) configuration demonstrates superior performance, achieving significant accuracy improvements [20]. Alejandro Lopez-Rincon et al. developed a lightweight CNN for NAO robot-based children's expression recognition. Despite its design, the accuracy is insufficient [21]. Adish Rao et al. designed an algorithm combining facial geometric features and a deep neural network (DNN) to study the effectiveness of neural networks in recognizing adult and children's expression features. Findings show that children's features are more challenging to extract, necessitating advanced feature extraction for accurate recognition [22]. Wenming Wang et al. believe that the feature extraction and generalization ability of traditional CNN is insufficient, and designed the multi-scale mixed attention mechanism network (MMANet). The network combines the multi-scale convolutional layer, mixed attention module and VGG16. It improves the accuracy and generalization ability of children's expression recognition, but reduces the computational efficiency of children's expression recognition [23]. Ulya Mahsa Anandiwa and her colleagues introduced the couple local binary patterns (LBP) and local ternary patterns (LTP) methods of children-learning readiness recognizing facial expression (Co-ChiLeRFE) algorithm, tailored for recognizing expressions specific to children. This algorithm leverages both LBP and LTP to extract meaningful features effectively. For classification, a support vector machine (SVM) is utilized to achieve accurate recognition of children's expressions. Nevertheless, the algorithm's recognition performance is constrained when confronted with complex expressions [24].

The advantages and disadvantages of the aforementioned child expression recognition algorithms are as follows. msDBM performs exceptionally well in unsupervised learning, but it lags behind specifically designed models such as CNN when dealing with large-scale image datasets. CNN is suitable for child expression recognition due to its robust processing capabilities for image data. However, it can suffer from insufficient robustness and generalization capabilities when trained on limited data, which may lead to overfitting. To reduce computational costs and adapt to low-power devices, lightweight CNN has been proposed, but it sacrifices some recognition accuracy in order to achieve this. MMANet improves child expression recognition by enhancing accuracy and generalization capabilities, but this increase in performance comes with a corresponding increase in model computational cost. Finally, Co-ChiLeRFE possesses a good descriptive ability for textures and local details, but its performance is limited when faced with complex expression features.

Aiming at the problems of low recognition accuracy, poor generalization ability and low efficiency of ordinary convolutional neural network in children's expression recognition, this paper proposes a multi-scale asymmetric convolutional neural network (MACNN). The network incorporates multi-scale convolutional blocks to enhance its capacity to extract diverse-sized features, subsequently boosting the model's recognition accuracy. Additionally, an asymmetric convolution layer is utilized to fortify the convolutional kernel's skeletal aspect, thereby enhancing the model's rotational robustness and generalizability. After training, the model in deep learning will utilize the learned parameters to perform inference, where it receives input data and generates corresponding output, without further gradient calculation or parameter updates. Therefore, the asymmetric convolution kernel can be replaced by the original square convolution kernel after training to reduce the amount of calculation, which improves the efficiency of the model to recognize expression.

Data serves as a crucial prerequisite for conducting research on expression recognition. With the development of expression recognition algorithms, high-quality expression data with rich samples and accurate labels is particularly important for designing robust expression recognition models [25]. Currently, most high-quality expression datasets consist primarily of images of adult subjects, including the JAFFE dataset [26], the KDEF dataset [27], the Multi-PIE dataset [28], the AFEW 7.0 dataset [29], the ExpW dataset [30], the fer2013 dataset [31], the Oulu-CASIA dataset [32], the EmotionNet dataset [33], the CHEVAD dataset [34], the RAF-DB dataset [35], the BU-3DFE dataset [36], and the AffectNet dataset [37], among others. However, due to the rapid change of children's expressions and the difficulty of capturing them, the low obedience of children to the experimenter, and the government's privacy protection of minors, it is difficult to establish a children's expression database.

In recent years, researchers have increasingly acknowledged the significance of establishing expression datasets specifically tailored for children, given their unique characteristics [38]. Consequently, despite encountering numerous challenges, several such datasets have been successfully developed. The Radboud Faces Database contains expression images of children aged 8-12 [39], the NIMH-CHEFS child emotional faces picture set contains images of children aged 10-17 [40], and the CAFE dataset contains facial photos of children aged 2-8 [41]. The EmoReact dataset contains 1,102 videos of children aged 4-14 [42]. The Liris-cse dataset contains 208 facial expression videos of 12 children aged 6-12 [43]. The ChildEFES dataset contains images and videos of the expressions of children aged 4-6 [44]. Although these datasets are related to children's expressions, their small size and limited public accessibility hinder comprehensive research efforts. Given the absence of publicly available large-scale children's expression datasets, this paper establishes a large-scale children's expression dataset independently. This dataset comprises public children's expression images voted on by 10 volunteers, offering a more comprehensive resource for research into children's expression recognition.

The contributions of this paper are summarised as follows:

- In this paper, we propose a children's expression recognition algorithm based on MACNN. The multi-scale convolutional layer is designed to extract richer children's expression features, while the asymmetric convolutional layer enhances the kernel skeleton of the convolutional neural network, thereby improving the model's performance. In addition, the asymmetric convolutional layer can be converted into a square convolutional layer after training, thereby enhancing the efficiency of children's expression recognition.
- In light of the limited availability of extensive and publicly accessible children's expression datasets, this paper introduces a new dataset comprising 15,157 images of children's expressions. The dataset has been meticulously curated from a variety of sources, predominantly from publicly available images on the Internet. Notably, this compilation includes images from well-established datasets such as the fer2013 dataset and the RAF-DB dataset, which have been instrumental in the field of facial expression recognition. Each image was further scrutinized by a panel of 10 volunteers to ensure the accuracy of the expression labels. This comprehensive dataset serves as an invaluable resource for the study of children's expression recognition, bridging the gap in the current landscape of available datasets.
- Experimental analysis is carried out on the children's expression dataset, which verifies the effectiveness and superiority of the proposed MACNN. In addition, to verify the robustness and generalization of the algorithm in natural scene expression recognition, detailed experimental analysis was carried out on the RAF-DB dataset, and the designed model was verified.

The structure of the paper is organized as follows. Section II provides a review of the related work, initially discussing the VGG16 network that forms the core of the algorithm in this study, followed by an examination of asymmetric convolution. Section III introduces the datasets used in this research, starting with a description of the process for constructing the in-house children's facial expression dataset, and then detailing the RAF-DB dataset. Section IV delves into the MACNN network proposed in this paper, primarily elucidating the multi-scale asymmetric convolutional units that are central to our approach. Section V describes the experimental setup and the environment in which the experiments were conducted. Section VI presents the results of applying the proposed algorithm to the children's facial expression dataset and the RAF-DB dataset, beginning with a presentation of the outcomes and then proceeding to an analysis of these results. Section VII concludes the paper by summarizing the findings and proposing potential avenues for future work.

II. Related Work

A. VGG16 Network Model

VGG16 is a convolutional neural network model proposed by Simonyan [45]. In the 2014 ImageNet image recognition challenge, it secured the second position and has subsequently gained widespread application in various computer vision tasks, including image classification and object detection. The VGG16 network employs consecutive small convolutional kernels (3×3) and pooling layers to construct a deep neural network with a depth of up to 16 layers. The network structure of VGG16 is illustrated in Fig. 1.

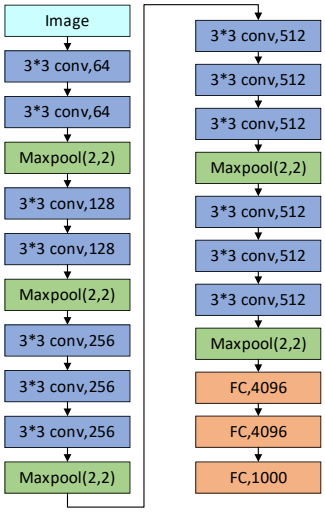

Fig. 1. VGG16 network.

The main characteristics of VGG16 are that the network structure is relatively deep, and the number of convolutional layers and pooling layers is large, so that the network can learn more high-level abstract features. In addition, the convolution layer of VGG16 uses 3×3 convolution kernels, and multiple 3×3 convolution kernels in series can form a convolution kernel with larger receptive field. After two 3×3 convolution kernels in series, the same receptive field as a 5×5 convolution kernel with step size 2 is obtained, and the amount of calculation is smaller. A 7×7 receptive field can be obtained by concatenating three 3×3 kernels. Therefore, the VGG16 network uses multiple 3×3 convolution kernels, which can increase the receptive field and improve the efficiency and accuracy of feature extraction.

B. Asymmetric Convolution

Asymmetric convolution is a convolution operation where the size of the convolution kernel is not square, but rectangular with different length and width.

One role of asymmetric convolution is to approximately replace square convolution. A d×d kernel can be replaced by 1×d and d×1 kernels to reduce the number of parameters [46], [47]. For example, in InceptionV3, a sequence of 1×7 and 7×1 convolutions replaces the 7×7 convolution kernel [48]. Efficient neural network (ENet) decomposes the 5×5 convolution kernel into two convolution kernels of 1×5 and 5×1, which reduces the amount of calculation of the network and makes the network run on embedded devices [49]. Efficient dense modules with asymmetric convolution (EDANet) divides the 3×3 convolution kernel into two kernels: 3×1 and 1×3. This division reduces the number of parameters and required computation by one-third [50]. If the rank of a 2D kernel is 1, it means that the kernel has only one eigenvector in the 2D space. In this case, this 2D kernel

can be represented equivalently by a sequence of 1D convolution operations. However, in deep learning networks, convolutional kernels typically encompass multiple feature values. Directly converting the 2D kernel into a sequence of 1D kernels can result in significant loss of information, as it disregards other directional feature information encapsulated within the convolutional kernel [51].

Another consequence of asymmetric convolution is its enhancement of the model's rotational robustness. This enhancement arises from the fact that, in comparison to square convolutions of dimension d×d, horizontal convolutions of size d×1 exhibit horizontal flip invariance, while vertical convolutions of size 1×d demonstrate vertical flip invariance. From a mathematical standpoint, the elements within each row remain invariant under horizontal flipping, and when flipped vertically, each column maintains the integrity of its elements. The Asymmetric Convolutional Network (ACNet) employs parallel convolution kernels of dimensions d×d, 1×d, and d×1, in lieu of the original d×d kernels. It aggregates the outputs of these three convolutional operations, preserving the entirety of the convolutional kernel information. This approach enhances both the rotational robustness and generalization capability of the model. Furthermore, post-training, the network maintains the same number of parameters as the original d×d convolutional kernel, ensuring computational efficiency [52].

III. Dataset Description

A. Children's Expression Dataset

Advancement in children's expression recognition systems is often hindered by data scarcity. This paper introduces a novel dataset to address this limitation, specifically curated for the study of children's facial expressions. Acknowledging the stringent security and ethical requirements for collecting children's facial imagery, the dataset was constructed from publicly accessible images on the Internet, thereby upholding children's privacy. The majority of these images were sourced from established, publicly available datasets, including the fer2013 and the RAF-DB, recognized for their contributions to the field of facial expression analysis. In alignment with the World Health Organization's classification, the term 'children' refers to individuals under the age of 14. To ensure the dataset's accuracy, a majority voting system was employed, with each image evaluated by ten annotators. An image was classified as a child's expression if it received positive confirmation from at least six annotators. The resulting dataset comprises 15,157 instances of children's facial expressions, categorized into seven primary emotional expressions. The distribution of these instances is delineated in TABLE I.

Due to the diverse origins of the extracted children's expression images sourced from publicly available repositories on the Internet, variations in both size and format are observed. Consequently, as a means of standardization, all images within the dataset have been resized and converted to 48×48 pixels in PNG format for consistency and comparability across the dataset. To reduce computational complexity and mitigate the influence of image color on expression recognition, all images in this study were converted to grayscale. An illustrative example image from the children's expression dataset established in this study is presented in Fig. 2.

TABLE I. Number of Images of each Expression in the Children's Expression Dataset

Expression category	*Number of pictures*
Angry	1213
Disgust	496
Fear	875
Happy	5208
Netural	2872
Sad	2943
Surprise	1550

Fig. 2. Example images of children's expression dataset.

B. RAF-DB Dataset

The RAF-DB dataset is a highly regarded real-world dataset that includes 15,339 images of facial expressions, each with a resolution of 100×100 pixels. It features the seven universal emotional expressions, which have been meticulously labeled by 40 independent annotators. These images are subject to variations in occlusion, pose, and lighting conditions, making them representative of the diversity and complexity of expressions found in natural environments. The meticulous annotation by a diverse group of annotators ensures the dataset's reliability and ecological validity, which are crucial for its significant practical utility and research value in facial expression recognition. Fig. 3 shows an example image of the RAF-DB dataset, whose details are given in TABLE II.

Fig. 3. Example image of the RAF-DB dataset.

TABLE II. Number of Images of each Expression in the RAF-DB Dataset

Expression category	*Number of pictures*
Surprise	1619
Fear	355
Disgust	877
Happy	5957
Sad	2460
Angry	867

To eliminate the potential impact of color on expression recognition, all images within the RAF-DB dataset have been

converted to grayscale. To further validate the robustness of the algorithm, this paper applies a 30% vertical and horizontal flipping to the images in the dataset.

IV. Children's Expression Recognition Network

A. Multi-scale Asymmetric Convolution Layer

In deep learning image recognition tasks, researchers usually use the algorithm of increasing the depth of the network to improve the feature extraction ability of the model, and then improve the recognition effect. For instance, ResNet enhances the depth and performance of the network by stacking residual blocks and incorporating residual connections, which preserves and propagates the original input information [53]. However, this approach also results in an increased number of network parameters and computational costs.

The experimental results of Inception demonstrate that using multiple convolution operations of different scales can achieve good facial expression recognition effects with an appropriate network depth and fewer network parameters. Multi-scale convolution is composed of three branches for feature extraction. By using convolution kernels of different sizes, the input features are convolved in parallel with different-sized kernels, enabling the network to perceive different values at the same layer. The features extracted from the three feature extraction branches are fused through the concatenate (concat) method to output the final feature map. The multi-scale convolution operation can extract feature information of different scales from the input facial expression image data, integrating both local and global feature information.

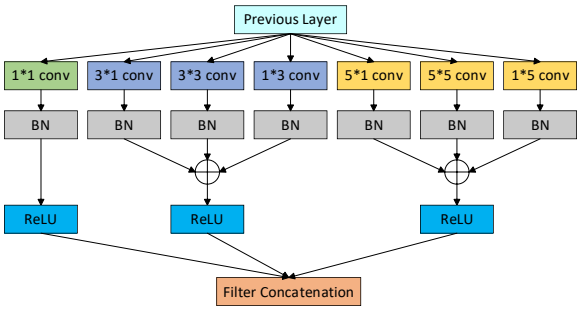

Fig. 4. Multi-scale asymmetric convolutional layer in training mode.

The experimental outcomes of ACNet demonstrate that asymmetric convolution effectively enhances model performance without escalating the count of network parameters. Consequently, this study incorporates the ACNet concept and substitutes the d×d convolution kernel in the multi-scale convolution with parallel convolution kernels of dimensions d×d, 1×d, and d×1. Consequently, a multi-scale asymmetric convolution layer is proposed. In the multi-scale asymmetric convolution layer, multi-scale convolution operations are employed to extract facial expression features of children at different scales. The original square convolution is replaced with asymmetric convolutions. Specifically, the horizontal convolution of size d×1 exhibits horizontal flip invariance, while the vertical convolution of size 1×d exhibits vertical flip invariance. This approach enhances the model's rotational robustness and generalization ability. After training, the asymmetric convolutions can be replaced with equivalent square convolutions to simplify computations. The multi-scale asymmetric convolution layer during training is illustrated in Fig. 4.

The multi-scale asymmetric convolution layer in training mode comprises seven branches. Specifically, three branches with 3×1, 3×3, and 1×3 convolutions replace the original 3×3 convolution kernel, while three branches with 5×1, 5×5, and 1×5 convolutions replace the original 5×5 convolution kernel. This approach allows the network to perceive features of different sizes within the same layer. Due to the horizontal convolution of size d×1 having horizontal flip invariance and the vertical convolution of size 1×d having vertical flip invariance, the use of asymmetric convolutions enhances the model's rotational robustness and generalization capability, thereby further improving the model's performance. Assuming that the input feature map is F, the convolution kernel is K, and the final output feature map is F_{concat}, the calculation formula for the multi-scale asymmetric convolution layer in training mode is presented in Eq. (1) to Eq. (10).

$$F_1 = ReLU(BN(Conv(F, K_{1\times1}) + b_1)) \quad (1)$$

$$F_{3\times1} = BN(Conv(F, K_{3\times1}) + b_2) \quad (2)$$

$$F_{3\times3} = BN(Conv(F, K_{3\times3}) + b_3) \quad (3)$$

$$F_{1\times3} = BN(Conv(F, K_{1\times3}) + b_4) \quad (4)$$

$$F_2 = ReLU(F_{3\times1} \oplus F_{3\times3} \oplus F_{1\times3}) \quad (5)$$

$$F_{5\times1} = BN(Conv(F, K_{5\times1}) + b_5) \quad (6)$$

$$F_{5\times5} = BN(Conv(F, K_{5\times5}) + b_6) \quad (7)$$

$$F_{1\times5} = BN(Conv(F, K_{1\times5}) + b_7) \quad (8)$$

$$F_3 = ReLU(F_{5\times1} \oplus F_{5\times5} \oplus F_{1\times5}) \quad (9)$$

$$F_{concat} = Concat(F_1, F_2, F_3) \quad (10)$$

2Due to the additivity of convolution, several compatible-sized 2D kernels operate on the same input with the same stride to generate outputs of the same resolution. Then, the outputs of these kernels are summed up, and these kernels are added at the corresponding positions, resulting in an equivalent kernel that produces the same output. Assuming that K_1 and K_2 represent two 2D kernels with compatible sizes, respectively, and I represent a matrix, the method is shown in Eq. (11).

$$I \times K_1 + I \times K_2 = I \times (K_1 \oplus K_2) \quad (11)$$

Compatibility among 2D kernels requires that different 2D kernels can produce outputs of the same size with the same input. Therefore, this paper adopts the algorithm of cropping the input feature map to enable d×d, d×1, and 1×d kernels to generate outputs of the same size. For instance, when d=3, to generate outputs of the same size, the image input to the 3×1 convolution kernel needs to be cropped by two rows of pixels, specifically the first and the last rows, while the image input to the 1×3 convolution kernel requires the removal of two columns of pixels, namely the first and the last columns, as shown in Fig. 5.

Similarly, when d=5, a similar approach can be employed to ensure that the 5×5, 5×1, and 1×5 convolution kernels produce outputs of the same size.

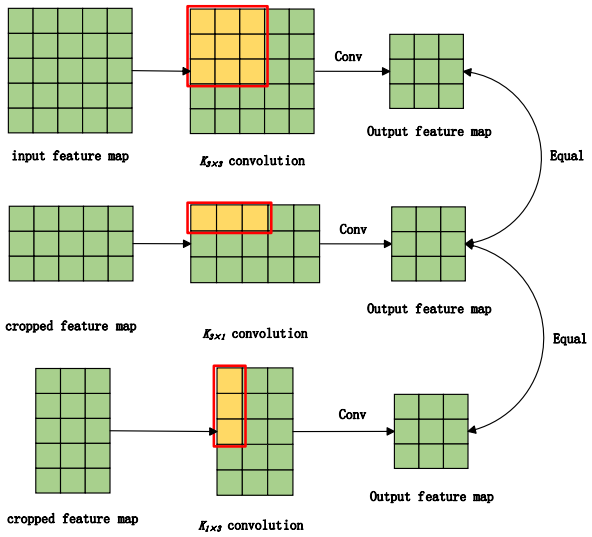

Fig. 5. Schematic diagram of producing outputs of the same size through cropping multi-scale convolution kernels.

Integrating Eq. (11), the aggregation of output feature maps derived from 3×3, 1×3, and 3×1 convolution kernels is analogous to the output feature map generated by a novel convolution kernel, which is constructed by amalgamating the output feature maps of the three kernels at their corresponding spatial locations. Assuming that the 3×3, 1×3, and 3×1 convolution kernels are represented as $K_{3\times3}$, $K_{1\times3}$, and $K_{3\times1}$, respectively, and I represents the input feature map, as shown in Eq. (12).

$$I\times K_{3\times3}+I\times K_{1\times3}+I\times K_{3\times1}=I\times(K_{3\times3}\oplus K_{1\times3}\oplus K_{3\times1}) \quad (12)$$

Where $K_{3\times3}\oplus K_{1\times3}\oplus K_{3\times1}$ represents the new convolution kernel obtained by adding $K_{3\times3}$, $K_{3\times1}$, and $K_{1\times3}$ at corresponding positions, as shown in Fig. 6.

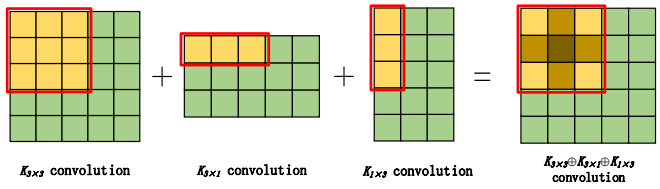

Fig. 6. Schematic diagram of summing convolution kernel output feature maps.

Fig. 6 shows that the $K_{3\times3}\oplus K_{1\times3}\oplus K_{3\times1}$ convolution kernel also has a size of 3×3. After training, since the model ceases parameter updates and gradient calculations and only performs inference, the convolution kernels with the same shape and number of parameters are equivalent. Therefore, the $K_{3\times3}\oplus K_{1\times3}\oplus K_{3\times1}$ convolution kernel and the $K_{3\times3}$ convolution kernel are equivalent, as shown in Fig. 7.

Fig. 7 illustrates that a single $K_{3\times3}$ convolution kernel can be used to replace the $K_{3\times3}\oplus K_{1\times3}\oplus K_{3\times1}$ convolution kernel to simplify calculations. Similarly, the $K_{5\times5}\oplus K_{1\times5}\oplus K_{5\times1}$ convolution kernel can also be replaced with a single $K_{5\times5}$ convolution kernel after training to simplify computations. Fig. 8 depicts the network structure of the multi-scale asymmetric convolution layer after training completion.

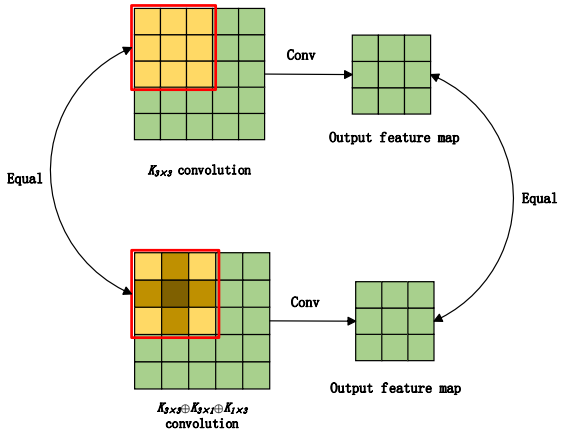

Fig. 7. Schematic diagram of convolution kernel equivalence after training completion.

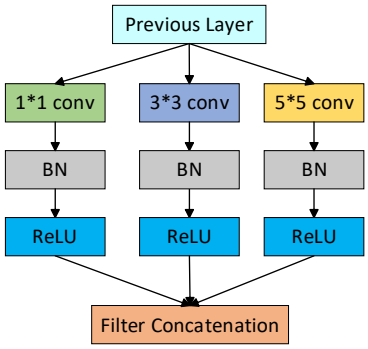

Fig. 8. Multi-scale asymmetric convolution layer after training completion.

Assuming the input feature map is F', K represents the convolution kernel, and the final output feature map is F'_{concat}, the computational formulas for the multi-scale asymmetric convolution layer upon completion of training are shown in Eq. (13) to Eq. (16).

$$F'_1=ReLU(BN(Conv(F',K_{1\times1})+b'_1)) \quad (13)$$

$$F'_2=ReLU(BN(Conv(F',K_{3\times3})+b'_2)) \quad (14)$$

$$F'_3=ReLU(BN(Conv(F',K_{5\times5})+b'_3)) \quad (15)$$

$$F'_{concat}=Concat(F'_1,F'_2,F'_3) \quad (16)$$

B. MACNN Network Structure

Addressing the issues of low recognition accuracy, insufficient generalization ability, and inefficiency in ordinary convolutional neural networks for child facial expression recognition, this paper proposes the MACNN. This network utilizes multi-scale convolution to extract feature information from different scales of images, enhancing the feature extraction capability of the model. The asymmetric convolution further enhances the model's rotational robustness and generalization ability. Finally, the VGG16, serving as the main body of the

network, improves the network depth through the stacking of 3×3 convolution layers, further enhancing the model's feature extraction capability. The network structure of MACNN is illustrated in Fig. 9, where MAConv represents the multi-scale asymmetric convolution layer.

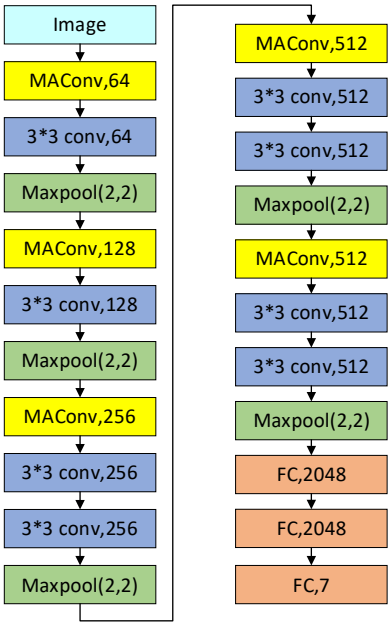

Fig. 9. The network structure of MACNN.

V. Experiment

To assess the performance of the proposed algorithm, a 10-fold cross-validation methodology was implemented on the child facial expression dataset. This approach ensured the algorithm's effectiveness was rigorously tested. To further evaluate the robustness and generalization of the algorithm in natural scene expression recognition, comprehensive experiments were performed on the RAF-DB dataset, a large-scale collection of real-world facial expressions. The model's performance was analyzed using confusion matrices and ROC curves generated from the test set, providing a detailed understanding of its ability to recognize various facial expressions.

Additionally, the efficiency of the proposed algorithm was compared against other state-of-the-art methods to demonstrate its real-time processing capabilities in recognizing child facial expression images.

The experimental setup was established using Python 3.8, with the PyTorch framework (version 1.7.1) and CUDA (version 11.0) for network model construction. The training and testing were conducted on a Linux system (version 3.10.0-1062.9.1.el7.x86_64). The system's hardware configuration included an Intel Xeon Silver 4114 CPU with a 2.20GHz clock speed, 252GB of memory, a Tesla P100 GPU, and 16GB of graphics memory.

For the optimization process, the Adam optimizer was selected with a learning rate set to 0.0001. The training was performed using a batch size of 32 and a total of 100 epochs, ensuring thorough convergence of the model's parameters.

VI. Results

A. Expression Recognition Results for Children's Expression Dataset

To enhance the precision of evaluating the MACNN's capability in discerning various categories of pediatric facial expressions, this manuscript introduces a confusion matrix derived from the experimental outcomes of the MACNN on a pediatric facial expression dataset. The matrix is depicted in Fig. 10. The diagonal elements of the matrix correspond to the true positive rate, indicating the proportion of instances correctly classified within each category. Conversely, the off-diagonal elements signify the misclassification probabilities, reflecting the rate at which instances from one category are incorrectly assigned to another.

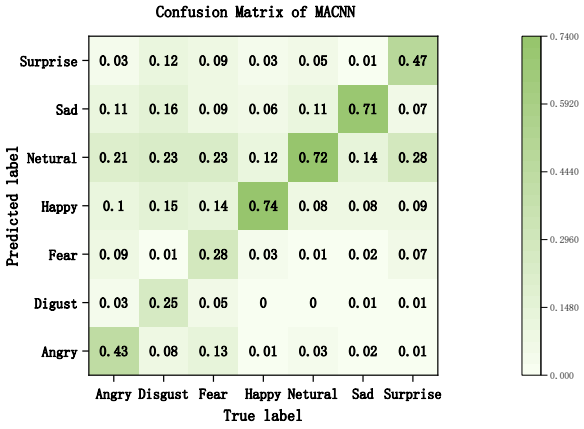

Fig. 10. Confusion matrix of MACNN.

This paper presents ROC curves for the MACNN's recognition of each category of children's facial expressions, as shown in Fig. 11. The AUC values for the seven expressions are 0.74, 0.64, 0.62, 0.95, 0.92, 0.91, and 0.80, respectively.

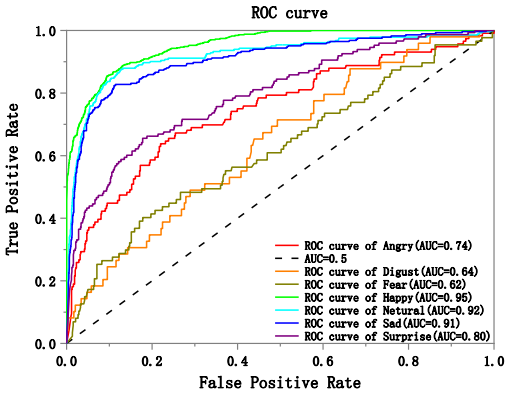

Fig. 11. ROC curve of children's expression dataset.

To substantiate the superiority of the proposed algorithm in the domain of children's facial expression recognition, this manuscript conducts comparative experiments against several classical algorithms. TABLE III. presents the comparative accuracy results across various algorithms on pediatric facial expression datasets.

TABLE III. RECOGNITION ACCURACY OF DIFFERENT ALGORITHMS IN CHILDREN'S EXPRESSION DATASET

Algorithms	*Accuracy (%)*
VGG16	59.96
ResNet-50	60.66
ResNet-152	61.20
Co-ChiLeRFE	60.39
MMANET	63.09
MACNN	63.35

This study aims to demonstrate the efficiency of the proposed algorithm for facial expression recognition by comparing the average recognition time required to process a single image across different algorithms. Specifically, the average recognition time is assessed for each algorithm on a standardized dataset. Table IV provides a summary of the average recognition times recorded for the respective algorithms.

TABLE IV. AVERAGE RECOGNITION TIME OF DIFFERENT ALGORITHMS IN CHILDREN'S EXPRESSION DATASET

Algorithms	*Average Recognition Time (seconds)*
VGG16	0.004
ResNet-50	0.011
ResNet-152	0.031
Co-ChiLeRFE	0.069
MMANET	0.015
MACNN	0.006

B. Expression Recognition Results for RAF-DB Expression Dataset

The confusion matrices depicting expression recognition on the RAF-DB dataset for both the baseline VGG16 model and the proposed network model are presented in Fig. 12 and Fig. 13, respectively. These figures also illustrate the recognition accuracy for various categories of facial expressions.

Fig. 14 presents the ROC curves for evaluating the classification performance of the MACNN on each facial expression category using the RAF-DB dataset.

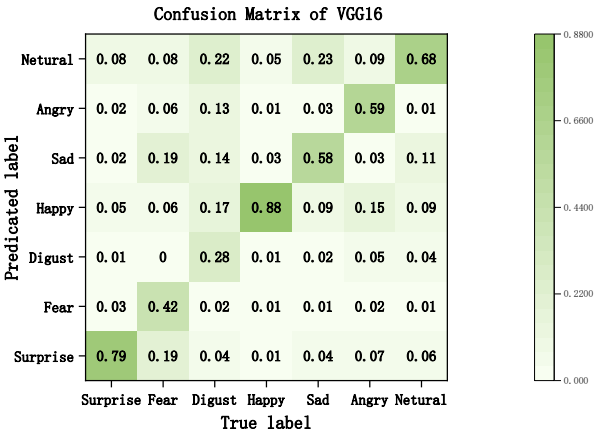

Fig. 12. Confusion matrix of baseline VGG16 under RAF-DB.

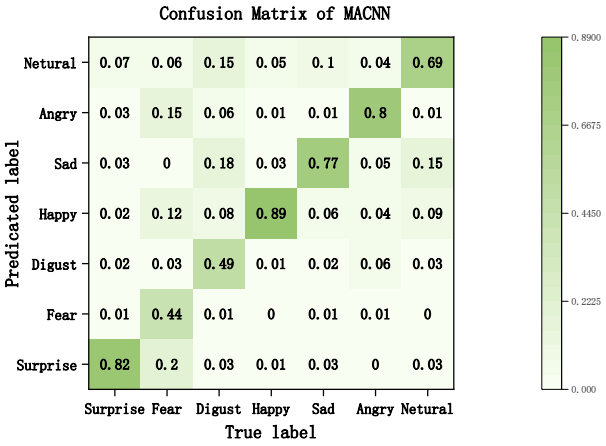

Fig. 13. Confusion matrix of baseline MACNN under RAF-DB.

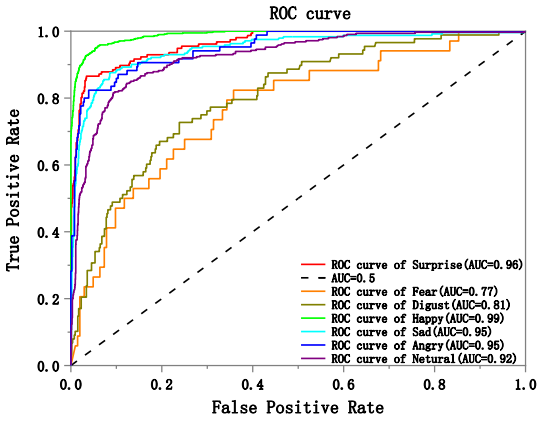

Fig. 14. ROC curve of RAF-DB dataset.

To assess the performance of the proposed algorithm in recognizing natural facial expressions, a comparative analysis has been performed against other prevalent and cutting-edge algorithms using the RAF-DB dataset. The results of these experiments are detailed in Table V.

TABLE V. RECOGNITION ACCURACY OF DIFFERENT ALGORITHMS IN RAF-DB DATASET

Algorithms	*Accuracy (%)*
VGG16	71.94
VGG19	73.76
PCARNet	77.67
ResNet-50	74.53
ResNet-152	75.06
Co-ChiLeRFE	72.58
MMANET	77.92
MACNN	78.26

C. Analysis of Experimental Results

Fig. 10 illustrates that the MACNN achieves higher recognition accuracy for happiness, neutrality, and sadness expressions in the context of pediatric facial expression

recognition. This outcome can be attributed to several factors. Firstly, the subtlety of children's facial features often results in images with minor expression variations being misclassified as neutral expressions. Secondly, the volunteer subjects who provided images were not professionally trained, leading to potential misclassification of some neutral expressions into other categories, which in turn affects the model training. Additionally, an analysis of the distribution of the pediatric facial expression dataset (as shown in Fig. 15) reveals an uneven distribution of sample sizes across different expression categories, which hinders model learning and significantly impacts recognition accuracy.

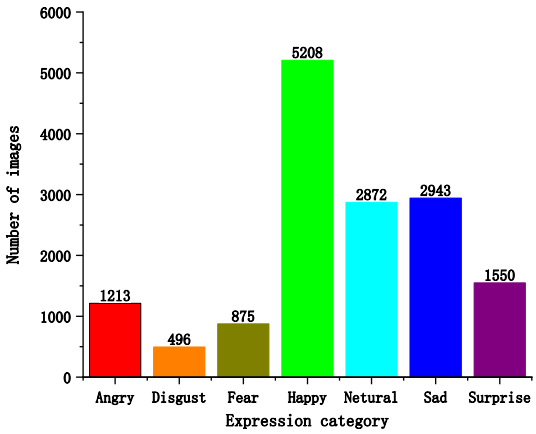

Fig. 15. Number of images of various expressions in the children's expression dataset.

Fig. 15 illustrates the imbalanced distribution of various expressions within the children's facial expression dataset, with happiness being the most frequently represented and disgust the least, resulting in a disparity exceeding an order of magnitude between the two. This imbalance affects the shape of the ROC curves, as shown in Fig. 11. It indicates that the MACNN achieves the highest recognition performance for expressions of happiness, neutrality, and sadness, with respective AUC values of 0.95, 0.92, and 0.91. In contrast, the recognition performance for expressions of disgust and fear is the lowest, with AUC values of 0.64 and 0.62, respectively.

TABLE III. presents a comparative analysis of the accuracy rates achieved by various algorithms on the children's facial expression dataset. The method proposed in this paper attained an accuracy rate of 63.36%, which is the highest among all the algorithms evaluated. TABLE IV. illustrates the average time required by each algorithm to recognize a single image. The method introduced in this study achieved an average recognition time of 0.006 seconds, ranking second among all algorithms, only 0.002 seconds slower than the VGG16. The experimental results demonstrate that while ensuring algorithmic efficiency, the performance of the algorithm has been enhanced.

Fig. 12 and Fig. 13 demonstrate that the MACNN has enhanced the recognition accuracy for the seven facial expressions in the RAF-DB dataset, with the most significant improvements observed in expressions of disgust, sadness, and anger, with respective increases of 0.21, 0.19, and 0.21. Improvements are also noted for the other expressions.

Fig. 14 presents the ROC curves of the proposed algorithm on the RAF-DB dataset, with AUC values for the seven expressions being 0.96, 0.77, 0.81, 0.99, 0.95, 0.95, and 0.92, respectively. All seven curves are positioned above the diagonal line, indicating that the algorithm performs exceptionally well in classifying expressions of surprise, happiness, sadness, anger, and neutrality, and provides good classification for expressions of fear and sadness.

TABLE V. demonstrates that the algorithm presented in this paper achieves the highest performance on the naturalistic expression dataset RAF-DB, with an accuracy rate of 78.26%, indicating excellent robustness and generalization capabilities.

In summary, compared to other algorithms, the MACNN proposed in this paper enhances the network's perceptual capacity by integrating multi-scale convolution to collect information from different receptive fields within the network. Additionally, the introduction of asymmetric convolution has improved the robustness and generalization of the algorithm while ensuring high efficiency, making it well-suited for application in pediatric facial expression recognition scenarios.

VII. Conclusion

This study introduces the Multi-scale Asymmetric Convolutional Neural Network (MACNN), an advanced architecture for recognizing children's facial expressions. It utilizes multi-scale and asymmetric convolution layers to enhance feature extraction and recognition accuracy.

Our experiments, conducted with a GPU Tesla P100 and 16GB of video memory, yielded a 63.35% accuracy on a self-constructed children's expression dataset. This result exceeds that of other benchmarked algorithms, showcasing MACNN's superior performance.

Further testing on the RAF-DB dataset, which features expressions in natural environments, resulted in a 78.26% accuracy. This underscores MACNN's robustness and its ability to generalize across different conditions, a critical aspect for real-world applications.

The high recognition accuracy and computational efficiency of MACNN position it well for practical use in fields such as child psychology, human-computer interaction, and child safety. Its demonstrated adaptability suggests it is well-suited for broader real-world deployment.

The network's performance metrics highlight its potential for real-time expression analysis in various systems, including educational software, telemedicine platforms, and child monitoring systems. Future work may focus on optimizing MACNN for mobile and embedded systems, expanding the diversity of training datasets, and incorporating temporal dynamics for enhanced dynamic expression recognition.

Acknowledgment

This study was supported by Shandong Provincial Undergraduate Teaching Reform Project (Grant Number: Z2021450), Shandong Provincial Natural Science Foundation of P.R. China (Grant Number: ZR2020QF069), National College Students' Innovation and Entrepreneurship Training Program (Grant Number: 202310433069), and Shandong University of

Technology Postgraduate Teaching Reform Project (Grant Number: 4053222063).

REFERENCES

[1] Pise. Anil Audumbar, et al., "Methods for facial expression recognition with applications in challenging situations," Computational intelligence and neuroscience, 9261438, 2022.

[2] X. Wen, J Zhou, J. Gan, and S. Luo, "A Discriminative Multiscale Feature Extraction Network for Facial Expression Recognition in the Wild," Measurement Science and Technology, vol. 35, 2024.

[3] Y. He, Y. Zhang, S. Chen, and Y. Hu, "Facial Expression Recognition Using Hierarchical Features With Three-Channel Convolutional Neural Network," IEEE Access, vol. 11, pp. 84785-84794, 2023.

[4] C. Shi, C. Tan, and L. Wang, "A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network," IEEE Access, vol. 9, pp. 39255-39274, 2021.

[5] J. H. Kim, A. Poulose, and D. S. Han, "CVGG-19: Customized Visual Geometry Group Deep Learning Architecture for Facial Emotion Recognition," IEEE Access, vol. 12, pp. 41557-41578, 2024.

[6] Y. He, "Facial Expression Recognition Using Multi-Branch Attention Convolutional Neural Network," IEEE Access, vol. 11, pp. 1244-1253, 2023.

[7] Q. Dong, W. Ren, Y. Gao, W. Jiang, and H. Liu, "Multi-Scale Attention Learning Network for Facial Expression Recognition," IEEE Signal Processing Letters, vol. 30, pp. 1732-1736, 2023.

[8] W. Aly, A. I. Shahin, and S. Aly, "A Novel Modular Deep Fully Convolutional Network for Efficient Low Resolution Facial Expression Recognition," Journal of Ambient Intelligence and Humanized Computing, vol. 14, pp. 7747–7759, 2023.

[9] B. Chen, J. Zhu, and Y. Dong, "Expression Recognition Based on Residual Rectification Convolution Neural Network," Multimedia Tools and Applications, vol. 81, pp. 9671–9683, 2022.

[10] H. Qi, X. Zhang, Y. Shi, and X. Qi, "A Novel Attention Residual Network Expression Recognition Method," IEEE Access, vol. 12, pp. 24609-24620, 2024.

[11] K. N. Kumar Tataji, M. N. Kartheek, and M. V. N. K. Prasad, "CC-CNN: A cross connected convolutional neural network using feature level fusion for facial expression recognition," Multimedia Tools and Applications, vol. 83, pp. 27619-27645, 2024.

[12] T. Kalsum and Z. Mehmood, "A Novel Lightweight Deep Convolutional Neural Network Model for Human Emotions Recognition in Diverse Environments," Journal of Sensors, 2023.

[13] Y. Liu, W. Dai, F. Fang, Y. Chen, R. Huang, R. Wang, and B. Wan, "Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition," Information Sciences, vol. 578, pp. 195-213, 2021.

[14] M. Mukhopadhyay, A. Dey, R. N. Shaw, and A. Ghosh, "Facial emotion recognition based on Textural pattern and Convolutional Neural Network," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), pp. 1-6, 2021.

[15] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," Neurocomputing, vol. 411, pp. 340-350, 2020.

[16] S. Saeed, A. A. Shah, M. K. Ehsan, M. R. Amirzada, A. Mahmood, and T. Mezgebo, "Automated Facial Expression Recognition Framework Using Deep Learning," J Healthc Eng., vol. 2022, 5707930, 2022.

[17] E. Serrat, A. Amadó, C. Rostan, B. Caparrós, and F. Sidera, "Identifying Emotional Expressions: Children's Reasoning About Pretend Emotions of Sadness and Anger," Front Psychol., vol. 11, 602385, 2020.

[18] Pipicella, Joseph Louis, et al., "Co-design and Consultation Ensure Consumer Needs Are Met: Building an eHealth Platform for Children with Inflammatory Bowel Disease," Digestive Diseases and Sciences pp. 4368-4380, 2023.

[19] M. Singh, S. Nagpal, R. Singh and M. Vatsa, "Dual Directed Capsule Network for Very Low Resolution Image Recognition," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 340-349, 2019.

[20] M. Rathod, C. Dalvi, K. Kaur, S. Patil, S. Gite, P. Kamat, K. Kotecha, A. Abraham, and L. A. Gabralla, "Kids' Emotion Recognition Using Various Deep-Learning Models with Explainable AI," Sensors, vol. 22, 8066, 2022.

[21] A. Lopez-Rincon, "Emotion Recognition using Facial Expressions in Children using the NAO Robot," in Proceedings of the 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP), pp. 146-153, 2019.

[22] A. Rao, S. Ajri, A. Guragol, R. Suresh, S. Tripathi, "Emotion Recognition from Facial Expressions in Children and Adults Using Deep Neural Network," in Proceedings of the Intelligent Systems, Technologies and Applications. Advances in Intelligent Systems and Computing, vol. 1148, 2020.

[23] W. Wang, M. Abisado, "Children's Expression Recognition Based on a Multiscale Mixed Attention Mechanism," International Journal of Sensor Networks, vol. 43, pp. 116-127, 2023.

[24] U. Mahsa Anandiwa, E. Rachmawati, R. Risnandar, "The Co-ChiLeRFE: Couple LBP and LTP Methods of Children-Learning Readiness Using Facial Expression," in Proceedings of the 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), pp. 177-182, 2021.

[25] S. Li, W. Deng, "Deep Facial Expression Recognition: A Survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195-1215, 2022.

[26] Lyons. Michael J, "" Excavating AI," re-excavated: debunking a fallacious account of the JAFFE dataset," arXiv: 2107.13998 , 2021.

[27] Li. Shan, Weihong. Deng, "Deep facial expression recognition: A survey," IEEE transactions on affective computing, pp. 1195-1215, 2020.

[28] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, "Multi-PIE," Image and Vision Computing, vol. 28, no. 5, pp. 807-813, 2010. Wang. Changzhong, et al., "Multiscale collaborative representation for face recognition via class-information fusion," Pattern Recognition, 110586, 2024.

[29] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, T. Gedeon, "From Individual to Group-Level Emotion Recognition: EmotiW 5.0," in Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 524-528, 2017.

[30] S. Li, W. Deng, "Deep Facial Expression Recognition: A Survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195-1215, 2022.

[31] Guo. Runfang, et al. "Development and application of emotion recognition technology—a systematic literature review," BMC psychology, 2024.

[32] Lie. Yang, Haohan. Yang, Bin-Bin. Hu, Yan. Wang, Chen. Lv, "A Robust Driver Emotion Recognition Method Based on High-Purity Feature Separation," IEEE Transactions on Intelligent Transportation Systems, pp.15092-15104, 2023.

[33] C. F. Benitez-Quiroz, R. Srinivasan, A. Martinez, "EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5562-5570, 2016.

[34] Liang, Chengxu, and Jianshe Dong. "A Survey of Deep Learning-based Facial Expression Recognition Research," Frontiers in Computing and Intelligent Systems, pp. 56-60, 2023.

[35] S. Li, W. Deng, J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2584-2593, 2017.

[36] H. Diao, X. Jiang, Y. Fan, M. Li, and H. Wu, "3D Face Reconstruction Based on a Single Image: A Review," in IEEE Access, pp. 59450-59473, 2024.

[37] A. Mollahosseini, B. Hasani, M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18-31, 2019.

[38] Q. Lin, R. He, P. Jiang, M. Xia, "Feature Guided CNN for Baby's Facial Expression Recognition," Complex., 8855885, 2020.

[39] Dawel. Amy, et al., "A systematic survey of face stimuli used in psychological research 2000–2020." Behavior Research Methods pp. 1889-1901, 2022.

[40] Negrão. Juliana. Gioia, et al., "The child emotion facial expression set: a database for emotion recognition in children," Frontiers in psychology, 666245, 2021.

[41] V. LoBue, C. Thrasher, "The child affective facial expression (CAFE) set: validity and reliability from untrained adults," Frontiers in Psychology, vol. 5, 1532, 2015.

[42] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, L. Morency, "EmoReact: a multimodal approach and dataset for recognizing emotional responses in children," in Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16), pp. 137–144, 2016.

[43] R. A. Khan, A. Crenn, A. Meyer, S. Bouakaz, "A novel database of children's spontaneous facial expressions (LIRIS-CSE)," Image and Vision Computing, vol. 83-84, pp. 61-69, 2019.

[44] J. G. Negrão, A. A. C. Osorio, R. F. Siciliano, et al., "The Child Emotion Facial Expression Set: A Database for Emotion Recognition in Children," Front. Psychol., vol. 12, 666245, 2021.

[45] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Computer Science, arXiv:1409.1556, 2014.

[46] Lee. JunKyu, et al., "Resource-efficient convolutional networks: A survey on model-, arithmetic-, and implementation-level techniques," ACM Computing Surveys pp. 1-36, 2023.

[47] Zu. Yueran, et al., "Asymmetric convolution kernel for deep optical flow estimation," 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). IEEE, 2020.

[48] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1-9, 2015.

[49] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation," arXiv: 1606.02147, 2016.

[50] Lo, S. Y., Hang, H. M., Chan, S. W., and Lin, J. J. Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In Proceedings of the 1st ACM International Conference on Multimedia in Asia pp. 1-6, December 2019.

[51] J. Jin, A. Dundar, E. Culurciello, "Flattened Convolutional Neural Networks for Feedforward Acceleration," arXiv:1412.5474, 2014.

[52] X. Ding, X. Zhang, A. Liu, J. Han, "ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks," IEEE International Conference on Computer Vision (ICCV), pp. 1911-1920, 2019.

[53] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.