2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

CP-Xception: A Lightweight Facial Expression Recognition Model For AI Companions

Xiaomeng Zhang[1] and Jianhua Cao[1(✉)]

[1] College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China
caojh@tust.edu.cn

Abstract. In this paper, we propose a lightweight facial expression recognition model: CP-Xception for AI companions, which is based on the Mini-Xception and features a small number of parameters, fast inference speed, and high recognition accuracy. Specifically, we integrate the feature segmentation concept from CSPNet into the model, splitting the input features into primary and secondary paths to extract deep and shallow features, respectively. We also incorporate the ParC module into the last two feature extraction stages of the backbone network. This enhancement enables the model to effectively capture both local details and global contextual information. The CP-Xception model is trained and evaluated on four public datasets: FER2013, FER2013Plus, CK+, and JAFFE. The results show that the CP-Xception model achieves recognition accuracy improvements of 2.16%, 3.37%, and 4.31% over the Mini-Xception model on the FER2013, CK+, and JAFFE datasets, respectively. And CP-Xception has only 30,149 parameters and 3.527 MFLOPs, which are approximately 50% of those of the Mini-Xception model, which makes the model more lightweight while also ensuring fast inference speed. We have deployed the model to the companion terminal for practical testing and observed satisfactory performance.

Keywords: facial expression recognition, CP-Xception model, AI companions, deep learning, lightweight model

1 Introduction

Over the past two years, the rapid advancement of artificial intelligence(AI), especially the large language model (LLM) technology has led to the emergence of various emotional companion electronic devices. These innovative products have delivered a novel experience to many consumers, particularly in stress relief and mental well-being improvement. Some devices have intelligent sensors and cameras, and the on-site images could be fetched. By using facial expression recognition(FER) technology, the companions directly give the required friendly and warm feedback to users, achieving a deeper understanding of users and providing genuine emotional companionship. Particularly for individuals with tendencies toward loneliness or depression, the device's ability to accurately identify negative emotions such as tension and anger enables it to

improve their psychological state through friendly dialogue and playing soothing music.

Facial expression recognition reveals the true emotional state by analyzing the changing features of the face. The detection of this capability relies on extracting salient features from the datasets. However, due to varying image acquisition conditions, datasets often exhibit significant feature fluctuations caused by factors such as illumination changes, head pose variations, and facial occlusion. Traditional machine learning methods mainly adopt feature engineering strategies, such as HOG [1], SVM[2], SURF[3], and SIFT[4]. Although these methods are simple, they often produce unsatisfactory results. In recent years, deep learning methods have achieved significant breakthroughs in many fields, including the FER task. The VGG model has demonstrated high accuracy on classical facial expression recognition datasets such as FER2013 and CK+. However, its deep architecture with 19 layers is prone to overfitting and gradient vanishing during training. Yexiu Zhong et al.[5] proposed a depth-FER model inspired by ResNet and SENet. Their results demonstrate that this model achieves higher accuracy compared to the VGG model. Yingjian Li[6] proposed an attention-based SPWFA-SE model capable of simultaneously perceiving local and global features, thereby achieving effective automatic facial expression recognition. Chang[7] introduced a Patch Attention Convective Vision Transformer (PACVT) model, which addressed the challenge of occluded facial expression recognition by adaptively calculating the horizontal attention weight of local features. However, the aforementioned models have problems of large parameter sizes and high computational complexity, which makes them unsuitable for deployment on such consumer-level electronics. Therefore, lightweight models that balance accuracy and parameter efficiency are essential to meet the demands of terminal deployment.

Mini-Xception[8] is a lightweight convolutional neural network model that achieves efficient feature extraction and classification through depthwise separable convolution and residual connections. It has been extensively applied in the field of facial expression recognition, particularly in resource-constrained scenarios[9]. Nevertheless, the operational efficiency of the Mini-Xception model still needs to be improved on certain resource-constrained devices, and its capabilities in feature extraction and recognition accuracy also have limitations[10]. To enhance the accuracy of facial expression recognition in companion robots, we propose an improved lightweight FER algorithm named CP-Xception, which is based on the Mini-Xception model but has less parameters and higher accuracy. The model has been deployed on the AI companion and shows impressive performance.

2 Methodology

2.1 Mini-Xception model

Mini-Xception is a lightweight neural network architecture based on the Xception[11] network, optimized for devices with limited resources. The Xception(Extreme Inception) network, proposed by Google, combines the strengths of traditional Inception modules with Depthwise Separable Convolutions. Depthwise Separable Convolutions

split the standard convolution operation into two stages: depthwise convolution and pointwise convolution. This design significantly reduces the computational complexity and the number of parameters while retaining the network's representation ability. Building on the Xception network, Mini-Xception further simplifies the model structure by reducing the number of layers and channels. These improvements substantially lower the demands on computing resources and memory, making it well-suited for deployment on mobile devices and other resource-constrained environments. However, despite its reduced parameter count and suitability for edge devices, Mini-Xception still exhibits limitations in real-time performance and recognition accuracy.

2.2 CP-Xception model

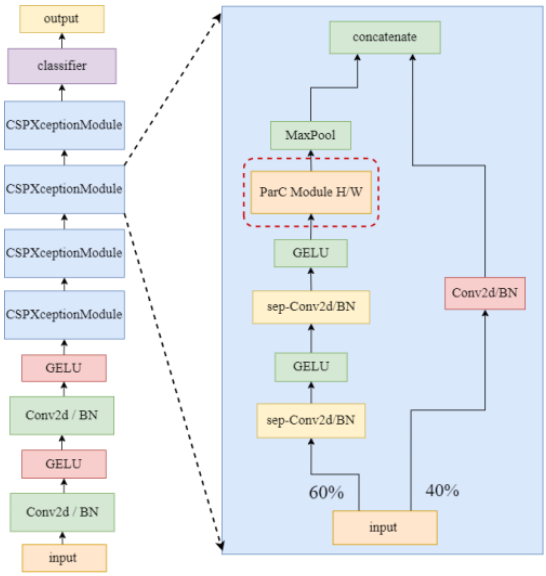


Fig. 1 Overall architecture of CP-Xception model

To enhance the real-time performance and accuracy of expression recognition, we propose CP-Xception model. While reducing the parameters, we enhance the model's ability to capture key facial feature areas, thereby improving accuracy in expression recognition tasks. In the backbone network design, we integrated the feature segmentation concept from CSPNet, splitting the basic features extracted from the initial features into two paths: the deep path is mainly responsible for deep feature extraction, while the shallow path retains the original feature details through the lightweight structure. After feature extraction by four modules, multi-level feature fusion is achieved via channel concatenation. Additionally, the ParC module is integrated into the last two feature extraction modules for further feature extraction. Finally, while maintaining classification performance, we lightened the original classification head to reduce the model's parameters and computational complexity. Given that our model builds on the Mini-Xception architecture and incorporates CSPNet and ParC concepts, we named it as CP-Xception. The overall architecture is shown in **Fig. 1**.

ParC module ParC-Net（Position-aware Circular Convolution Network)[12] is an innovative convolutional neural network that combines traditional convolution with global modeling capabilities through position-aware circular convolution (ParC). Its core design includes Circular Padding: the input feature map is spliced to form a circular boundary, enabling the convolution kernel to capture global context across the image **edges**; Position Embedding: it enhances the model's perception of key local areas by injecting spatial structure information; Grouped Convolution: it reduces the calculation amount through the lightweight design of independent channels, while retaining the translational invariance of convolution. The FER task mainly depends on the characteristics of local key areas (such as eye and mouth shapes) and the overall relevance of the face. The ability of circular padding and large kernel convolution to capture the global context can improve the discrimination of complex expressions (such as "surprise" and "fear"); Meanwhile, position embedding enhances the model's perception of key areas, and the complexity of grouping convolution and linear calculation makes it more suitable for real-time expression recognition. Thus, the ParC module is integrated in the last two modules of the backbone network and performs convolution operations in vertical and horizontal directions, respectively. In this way, the ParC module can help the model to improve the comprehensive processing of global information and local features. The structure of the ParC module in the network is shown in **Fig. 2**.

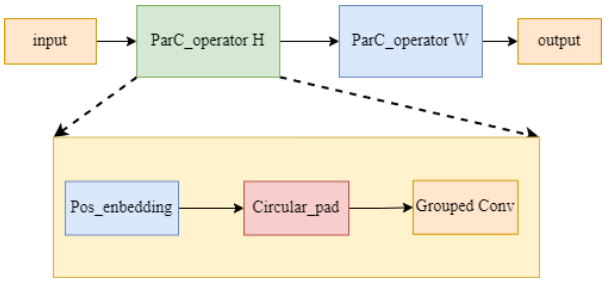


Fig. 2 ParC module

CSPNet CSPNet (Cross Stage Partial Network)[13] is based on optimizing the processing mode of the feature map, reducing the calculation and memory occupation, and enhancing the learning ability of the convolutional neural network (CNN). Its core lies in cross-stage partial connection. Specifically, the feature map is divided into multiple parts, processed separately, and then reassembled in the subsequent stage. After the initial feature extraction of the model, inspired by the concept of CSPNet, we divide the input feature map into two parts with a ratio of 6:4. The main path is responsible for deep feature extraction and is the main module of the model. It includes two Depthwise Separable Convolution layers and ParC modules added to the last two modules to further enhance the expression ability of the features. The secondary path is responsible for shallow feature extraction, where basic features can be quickly extracted through a simple 1×1 convolution and batch normalization. Finally, the features extracted from the main path and the secondary path are merged in the channel dimension and activated by the GELU activation function to obtain the final output. The model can strike a balance between deep and shallow features. At the same time, due to the segmentation

of feature maps, only 60% of features enter the main path for deep processing, which significantly reduces the model parameters.

In addition, to address the problem of gradient vanishing caused by zero gradient in the negative interval of the traditional ReLU activation function, we replace the activation function in the network with the GELU activation function. Compared with the unidirectional suppression characteristic of ReLU, GELU exhibits a smooth, differentiable, and non-monotonic activation characteristic. Its smooth and non-monotonic nature is conducive to the model capturing more nonlinear feature interactions.

It is worth noting that after improving the Mini-Xception model by incorporating the CSPNet feature segmentation concept and adding the ParC module, the CP-Xception model's parameters were reduced to approximately 53% of the original. This significantly decreased the parameter count while maintaining model performance.

3 Experiments and Analysis

3.1 Datasets

Facial expression recognition(FER) datasets are crucial resources in the fields of affective computing and computer vision, serving as the foundation for training and evaluating facial expression recognition models. In this experiment, we utilized four common datasets as follows.

FER2013 Dataset. The FER2013 dataset is a widely used benchmark in the field of facial expression recognition. The dataset comprises 35,887 grayscale images of size 48×48 pixels, labeled with seven basic expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral.

FER2013+ Dataset. This dataset is an extension of the original FER2013 dataset. The annotations were refined by the Microsoft Research Institute, with each image relabeled by 10 crowdsourcing annotators. The dataset contains the same number of images as the original FER2013 dataset, comprising 35,887 facial expression images, each with a size of 48×48 pixels.

CK+ Dataset. The Extended Cohn-Kanade Dataset (CK+) is a widely used benchmark for facial expression recognition. It comprises 593 image sequences from 123 subjects, capturing the transition from a neutral expression to a target expression. The target expressions include six basic emotions—anger, disgust, fear, happiness, sadness, and surprise—as well as neutral. By extracting the last three frames of each sequence, a total of approximately 981 images were obtained for the dataset.

JAFFE Dataset. The dataset consists of facial expression images captured from 10 Japanese female participants who posed various expressions according to experimental

instructions. The dataset contains a total of 213 images, with each participant displaying seven distinct expressions: anger, disgust, fear, happiness, sadness, surprise and neutral. Each expression category includes approximately 20 samples.

Fig. 3 shows sample images of seven basic expressions in these four public datasets.

Fig. 3 Facial expression recognition datasets

There is a serious imbalance in the distribution of FER2013 and FER2013+ datasets, which leads to the fact that the model can not learn fewer types of data when learning. In the training process, we implement the following strategies to solve the problem of unbalanced datasets: first, we apply more extensive data augmentation to the classes with fewer samples in order to increase the volume of data in these categories; Second, we assign appropriate weights to each category. This ensures that the model focuses more on categories with less data while still paying attention to other categories, thereby achieving a relative balance across the entire dataset.

The original datasets are divided into training, validation, and test sets according to a 7:2:1 ratio. The model is trained on the training set and subsequently evaluated on the validation and test sets. For datasets with limited data, we employ 10-fold cross-validation to train the proposed model.

3.2 Experimental conditions and Evaluation metrics

The experimental environment is configured as follows: The Python version is 3.9, with an NVIDIA GeForce RTX 3090 GPU capability. The training parameters are set as follows: The batch size is 64, and the initial learning rate is set to 0.001. The AdamW optimizer is employed to optimize the training process, and Focal Loss is used as the loss function.

In the task of facial expression classification, Confusion Matrix, Accuracy, and ROC curve are used as evaluation metrics, K-fold cross-validation is as the validation methods. The following are the specific introductions：

Confusion **Matrix** In facial expression recognition, the confusion matrix evaluates classification performance. For a binary case ("happy"=positive, "unhappy"=negative):

TP(True Positive): "happy" expressions correctly classified as "happy"
TN(True Negative): "unhappy" expressions correctly classified as "unhappy"
FP(False Positive): "unhappy" expressions mistakenly classified as "happy"
FN(False Negative): "happy" expressions mistakenly classified as "unhappy".

Accuracy In facial expression recognition, accuracy measures the proportion of correctly classified emotional instances relative to the total number of samples. The calculation formula is:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \quad (1)$$

ROC Curve The ROC curve helps evaluate model performance in expression recognition tasks. The closer the ROC curve is to the upper-left corner, the better the model's performance, indicating a higher true positive rate at a lower false positive rate.

K-Fold Cross-Validation In scenarios with small-scale datasets, K-Fold Cross-Validation can significantly enhance model evaluation robustness by maximizing the use of limited data resources. In this study, we apply a 10-Fold Cross-Validation strategy to the CK+ and JAFFE datasets. The datasets are randomly divided into 10 mutually exclusive and balanced subsets via stratified sampling, ensuring that the category distribution of each subset mirrored the original dataset. During each iteration, one subset serves as the test set, while the remaining nine subsets are combined into the training set. A model is trained on the current training set, and its expression recognition accuracy is evaluated on the corresponding test set. Finally, the average accuracy across the 10 iterations is used as the comprehensive performance metric.

3.3 Experiment results

In this study, systematic experiments have been conducted on four public datasets: FER2013, FER2013Plus, CK+, and JAFFE. The comprehensive performance of our model is evaluated by comparing it with other deep neural network models, such as Xception, MobilenetV2[14], and Mini-Xception. The comparisons have been made in two key dimensions: recognition accuracy and parameter Count.

Table 1 Experimental comparison results of the model

Model	FER2013/%	FER2013+/%	CK+/%	JAFFE/%	Parameters/M
Xception	68.12	77.32	97.13	90.47	22.8
MobileNetV2	67.90	78.35	90.82	89.48	3.47
Mini-Xception	66.20	77.76	93.88	86.82	0.06
CP-Xception(**ours**)	68.36	76.58	97.25	91.13	0.03

The experimental results in **Table 1** indicate that our model outperforms the Mini-Xception model on CK+ and JAFFE datasets, particularly on the CK+ dataset, and the recognition accuracy reaches 97.25%. The CK+ dataset consists of high-precision labeled images captured in controlled laboratory environments, with sequential transitions from neutral to peak expressions. The ParC module in CP-Xception enhances global context modeling to capture holistic correlations in facial muscle movements. Meanwhile, the dual-path structure of CSPNet improves sensitivity to subtle expression variations. Remarkably, CP-Xception achieves this with only half of the parameters while maintaining competitive accuracy on FER2013 (68.36% vs. 66.20%). The model demonstrates an optimal balance between lightweight efficiency and robust performance across diverse scenarios.

To comprehensively evaluate the model's performance across various datasets, we present the confusion matrix for each dataset, the average accuracy of the CK+ and JAFFE datasets under 10-fold cross-validation, and the corresponding ROC curves in **Fig. 4**, **Fig. 5**, and **Fig. 6**.

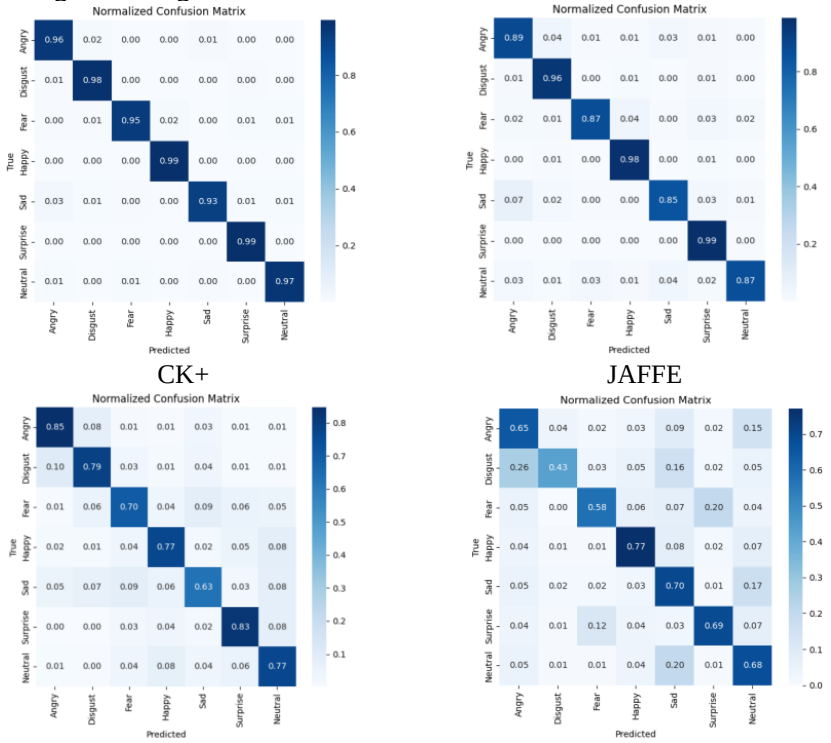
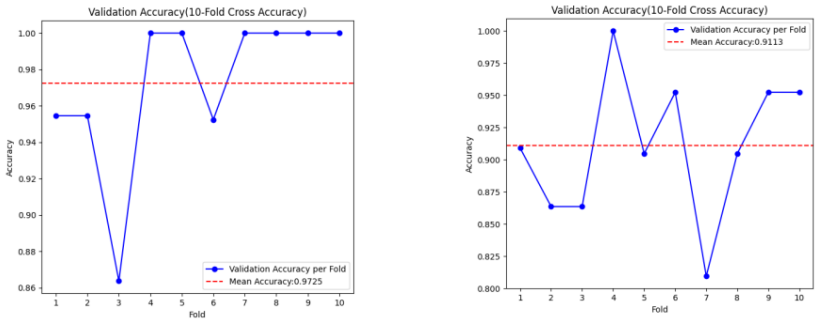
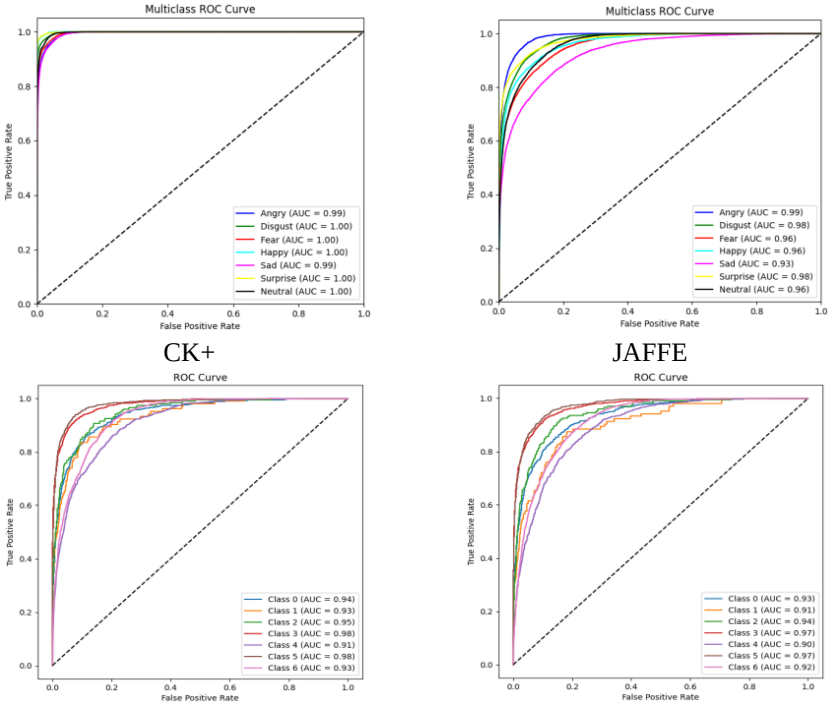


Fig. 4 The confusion matrix of each dataset. The color shading indicates the classification performance. The darker the color, the higher the accuracy of prediction.



CK+ JAFFE

Fig. 5 The average accuracy of 10-fold cross-validation on the validation set for the CK+ and JAFFE datasets



CK+ JAFFE

FER2013Plus FER2013

Fig. 6 ROC curve of each dataset

From the confusion matrix (as shown in **Fig. 4**), it can be seen that the CP-Xception model shows high recognition accuracy across most expression categories. **Fig. 5** presents the results of 10-fold cross-validation on CK+ and JAFFE datasets, and the average accuracy of the improved model on the two datasets reaches 97.25% and 91.13% respectively. The ROC curves in **Fig. 6** demonstrate that the improved model's ROC

curves are close to the upper-left corner for both datasets, indicating high classification accuracy and strong discriminative ability.

3.4 Ablation study

The CP-Xception is based on the Mini-Xception architecture, enhanced by incorporating the CSPNet concept and the ParC module. To thoroughly investigate the impact of these components on the network, we conduct ablation studies to analyze their specific contributions in reducing parameter count and improving performance.

In this ablation study, we systematically test various module combinations in the model to evaluate the specific impact of each module on model performance. The results clearly demonstrate that both the ParC module and the CSPNet module significantly enhance model performance, particularly on the CK+ dataset. Moreover, when the CSPNet and ParC modules are integrated simultaneously, the model achieves optimal performance on both the FER2013 and CK+ datasets. Comparisons of the ablation study results confirm that the CSPNet and ParC modules are highly effective in improving model performance in this experiment.

Table 2 Ablation experiment results

Model	FER2013/%	CK+/%
Mini-Xception	66.20	93.88
Mini-Xception +ParC	67.31	94.42
Mini-Xception +CSPNet	65.87	95.80
Mini-Xception +CSPNet+ParC	68.36	97.25

Table 3 Comparison of parameters and complexity

Model	MFLOPs	Parameters
Mini-Xception	7.102	56951
CP-miniXception(ours)	3.527	30149

By comparing the performance of the Mini-Xception model and the CP-Xception model in terms of parameters and computational complexity, our model has achieved significant optimization in both aspects. Specifically, the MFLOPs of our model is 3.527, approximately 50% lower than that of the Mini-Xception model. This indicates that our model requires fewer computing resources and is more suitable for deployment in resource-constrained environments. Additionally, the parameter count of our model is 30,149, which is about 47% less than the Mini-Xception model's parameters. Therefore, while maintaining high performance, the CP-Xception model significantly reduces computational complexity and parameter count, enhancing its practicality and efficiency. This is particularly advantageous in scenarios requiring rapid inference and limited resources. These results demonstrate that optimizing the model structure can significantly improve efficiency and practicality without sacrificing performance.

4 Applications

To verify the actual operational efficiency of the CP-Xception model, we specially developed an AI companion toy named Judy. Judy is designed to look like a monkey, which is quite aesthetically pleasing. It is equipped with a Raspberry Pi 5B, a high-definition camera, a microphone, speakers, and other hardware components. Additionally, it integrates an open-source DeepSeek large language model through network API, enabling the device to receive voice inputs, respond appropriately to questions, and play back responses through the speaker. The high-definition camera is intended for capturing facial images, analyzing facial expressions to determine the user's emotional state at the time, and providing positive feedback in conjunction with the user's queries, thereby achieving emotional companionship.

Fig. 7 illustrates the technical application process. The file size of the CP-Xception model weights deployed on the terminal device is only 297 KB, representing a significant 65% reduction compared to the Mini-Xception model's 853 KB. This substantially alleviates the storage requirements of hardware devices. Actual measurements indicate that the loading speed of the optimized model on terminal equipment is 70% faster, with an inference time of just 0.021 seconds.We have conducted a practical test with a group of five people. In front of the AI companion Judy, due to factors such as facial angles, hair occlusion, and dynamic movements, the accuracy of emotion recognition reached 72.5%, and the device also provided relatively ideal emotional feedback. It should be noted that these facial expressions were intentionally made during the test and did not reflect genuine emotions. In the future, we will continue to evaluate Judy's performance in real-world usage scenarios, with a particular focus on improving emotion recognition capabilities by fully considering various influencing factors.

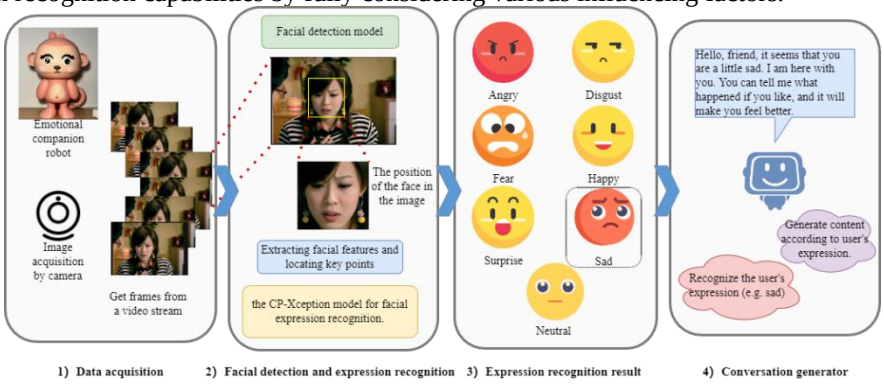


Fig. 7 Overall process of emotional AI companion interaction

5 Conclusion

In this paper, we propose a lightweight deep-learning model CP-Xception for FER study on resource-constrained devices. The model is based on the Mini-Xception and has improvements in both network architecture and feature extraction mechanisms. The

backbone network is divided into primary and secondary paths for feature extraction at different depths, and the ParC module is integrated into the last two feature extraction modules of the backbone network to improve feature extraction capabilities. Experiments on typical datasets demonstrate that the CP-Xception model achieves recognition accuracies of 95.25% on the CK+ dataset, 91.13% on JAFFE, 68.36% on FER2013, and 76.58% on FER2013Plus. Moreover, it maintains a significantly reduced parameter size and computational complexity, approximately 50% lower than that of the benchmark model. We finally deploy the model on a special AI companion toy with camera, demonstrating impressive performance. In future work, we will further optimize the model to adapt to diverse environmental conditions for better emotional feedback from the AI companion.

Acknowledgments. The authors would like to thank all the anonymous reviewers for their valuable comments to improve this paper.

References

1. Pierluigi, C., Marco Del, C., Marco, L., Cosimo, D.: Facial Expression Recognition and Histograms of Oriented Gradients: a Comprehensive Study. SpringerPlus 4(1) (2015)
2. Liyuan, C., Changjun, Z., Liping, S.: Facial Expression Recognition Based on SVM in E-learning. IERI Procedia 2, 781–787 (2012)
3. Qiyu, R., Xing, Q., Qirong, M., Yongzhao, Z.: Multi-pose Facial Expression Recognition Based on SURF Boosting. In: Proceedings of Affective Computing and Intelligent Interaction, pp. 630–635 (2015)
4. Hamit, S., Hasan, D.: Improved SIFT matching for pose robust facial expression recognition. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp. 585–590 (2011)
5. Yexiu, Z., Senhui, Q., Xiaoshu, L., Zhiming, M., Junxiu, L.: Facial Expression Recognition Based on Optimized ResNet. In: Proceedings of the 2nd World Symposium on Artificial Intelligence (WSAI), pp. 84–91 (2020)
6. Yingjian, L., Guangming, L., Jinxing, L., Zheng, Z., David, Z.: Facial Expression Recognition in the Wild Using Multi-Level Features and Attention Mechanisms. IEEE Transactions on Affective Computing 14(1), 451–462 (2023)
7. Chang, L., Kaoru, H., Yaping, D.: Patch Attention Convolutional Vision Transformer for Facial Expression Recognition with Occlusion. Information Sciences 619, 781–794 (2023)
8. Octavio, A., Matias, V., Paul, P.: Real-time Convolutional Neural Networks for Emotion and Gender Classification. CoRR, abs/1710.07557 (2017)
9. Yinghui, K., Zhaohan, R., Ke, Z., Shuaitong, Z., Qiang, N., Jungong, H.: Lightweight Facial Expression Recognition Method Based on Attention Mechanism and Key Region Fusion. Journal of Electronic Imaging 30(6) (2021)
10. Yu, C., Ding, X., Yang, L., Zhou, G., Yan, C.: Study on Mini-Xception-Based Improved Lightweight Expression Detection Model. In: Proceedings of the 2nd International Conference on Control and Intelligent Robotics (ICCIR), pp. 207–215 (2022)
11. Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800 –1807 (2016)

12. Zhang, H., Hu, W., Wang, X.: ParC-Net: Position Aware Circular Convolution with Merits from ConvNets and Transformer. In: Proceedings of the European Conference on Computer Vision (ECCV), LNCS, vol. 13686, pp. 613–630. Springer, Heidelberg (2022)
13. Wang, C., Liao, H. M., Yeh, I., Wu, Y., Chen, P., Hsieh, J.: CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1571–1580 (2020)
14. Mark, S., Andrew, H., Menglong, Z., Andrey, Z., Liang-Chieh, C.: Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520 (2018)