ISSN: 2688-8645

Research on Micro Expression Emotion Recognition Algorithm Based on Improved YOLOV8 and ResNet

Lei Guo [a], Zhiyu Lin [b]

School of Electronics and Information, Southwest Minzu University, Chengdu Sichuan, 610000, China

[a] 17695089252@163.com, [b] 15238507626@163.com

Abstract

With the rapid advancement of artificial intelligence, micro-expression recognition has demonstrated significant application value in fields such as criminal investigation and lie detection. However, the extremely short duration of micro-expressions and the scarcity of annotated data pose substantial challenges to accurate recognition. This study proposes a two-stage micro-expression recognition method based on an improved YOLOv8 and ResNet framework. In the first stage, a YOLOv8 model enhanced with StarNet multi-scale feature fusion and the CBAM attention mechanism is employed to achieve high-precision face detection. In the second stage, the detected static facial frames are fed into a ResNet50 network embedded with an SE module to strengthen channel-wise feature representation, thereby enabling micro-expression classification. For experimental validation, the improved models were trained and tested on the WiderFace and RAF-DB datasets. Results indicate that the improved YOLOv8 achieved an mAP of 0.693 on WiderFace, representing an increase of 0.109 over the baseline, while the SE-ResNet50 attained an F1-score of 0.7548 on RAF-DB, with an improvement of 0.2. These findings confirm the effectiveness of the proposed approach in both face detection and micro-expression classification. Moreover, this study provides a feasible solution to address the challenges of capturing micro-expressions and data insufficiency, while also laying the groundwork for future research on end-to-end joint modeling and lightweight network optimization.

Keywords

Micro-expression Recognition; YOLOv8; ResNet; Attention Mechanism; Two-stage Model.

1. Introduction

Facial expressions are a primary form of nonverbal communication that directly convey an individual's emotions and psychological states. Unlike macro-expressions, micro-expressions are involuntary, extremely brief (typically less than 500 ms)[1], and therefore provide deeper insights into genuine emotions. With increasing demands in fields such as criminal investigation, lie detection, and mental health assessment, automatic micro-expression recognition has attracted growing attention. Nevertheless, the transient nature of micro-expressions, coupled with the scarcity of annotated datasets, makes accurate and efficient recognition particularly challenging.

Early research in micro-expression recognition relied heavily on handcrafted features such as Local Binary Pattern on Three Orthogonal Planes (LBP-TOP)[2] and optical flow analysis. While these approaches capture certain local facial dynamics, they often suffer from computational complexity, limited robustness, and poor generalization. More recently, deep learning methods, especially convolutional neural networks (CNNs), have become the mainstream due to their capacity for end-to-end feature learning. However, existing deep learning approaches still face

ISSN: 2688-8645

limitations in accuracy and real-time performance, largely constrained by small dataset sizes and the subtle, complex nature of micro-expressions.

To address these issues, this study proposes a two-stage micro-expression recognition framework integrating an improved YOLOv8[3] and ResNet50. In the first stage, YOLOv8 is enhanced with StarNet multi-scale feature fusion and the Contextual Anchor Attention (CAA) mechanism to achieve more accurate face detection, particularly for small targets. In the second stage, a Squeeze-and-Excitation (SE) module is incorporated into ResNet50 to strengthen channel-wise feature representation and improve classification. Experiments on the WiderFace and RAF-DB datasets demonstrate that the proposed approach significantly outperforms baseline models, achieving higher detection accuracy and classification performance.

The contributions of this work are threefold: (1) a two-stage recognition framework that effectively combines face detection and expression classification; (2) integration of attention mechanisms and feature fusion strategies to enhance subtle micro-expression representation; and (3) empirical validation through extensive experiments and ablation studies. These findings provide a promising direction for advancing micro-expression recognition and lay the groundwork for future research on end-to-end modeling and lightweight network optimization.

2. Methods

This section presents the design of a micro-expression recognition framework based on an improved YOLOv8 integrated with ResNet. First, for YOLOv8, StarNet and the Convolutional Block Attention Module (CBAM)[4] were combined and incorporated into the C2f module, thereby enhancing the detection accuracy of both general faces and small facial targets. Second, for the ResNet network, the Squeeze-and-Excitation (SE)[5] module was introduced to enable the network to focus on channels with high information content, which plays a crucial role in the subsequent classification of expression patterns. Finally, the advantages and limitations of a two-stage recognition strategy are discussed, and the two-stage approach was adopted as an efficient solution for micro-expression recognition in this study.

2.1. Improvement of YOLOv8-Based Face Detection Algorithm

2.1.1. StarNet

StarNet is a model based on the Star Operation, constructed by stacking multiple Star Operation layers[6]. Its core design philosophy is to achieve efficient feature representation and task processing through a minimalist architecture and reduced manual intervention. The key concepts of StarNet are as follows:

1. **Star Operation:** The Star Operation refers to element-wise multiplication used to integrate features from different subspaces. This operation maps input features into a high-dimensional nonlinear feature space, similar to kernel techniques, without increasing the network width.
2. **High-Dimensional Nonlinear Feature Mapping:** Unlike traditional neural networks that achieve high-dimensional features by widening the network, the Star Operation achieves this in a manner analogous to polynomial kernel functions.
3. **Efficient and Compact Network Structure:** By stacking multiple Star Operation layers, each layer significantly increases the implicit dimensional complexity. Even within a compact feature space, the Star Operation can exploit high-dimensional latent features effectively.

ISSN: 2688-8645

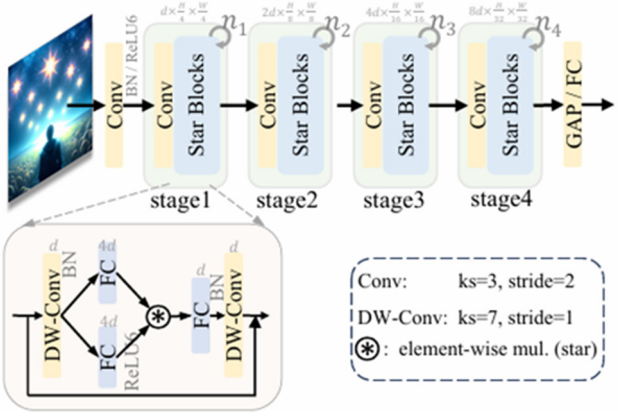

Figure 1. Structure of StarNet

2.1.2. CBAM

The Convolutional Block Attention Module (CBAM) is referred to as a hybrid attention mechanism because, compared to the SE channel attention mechanism, it not only retains the original channel attention but also incorporates spatial attention. By optimizing the network from both channel and spatial perspectives, CBAM enables the model to capture more informative features along both dimensions, thereby further enhancing feature extraction effectiveness. The structure of CBAM is illustrated in Figure 2.

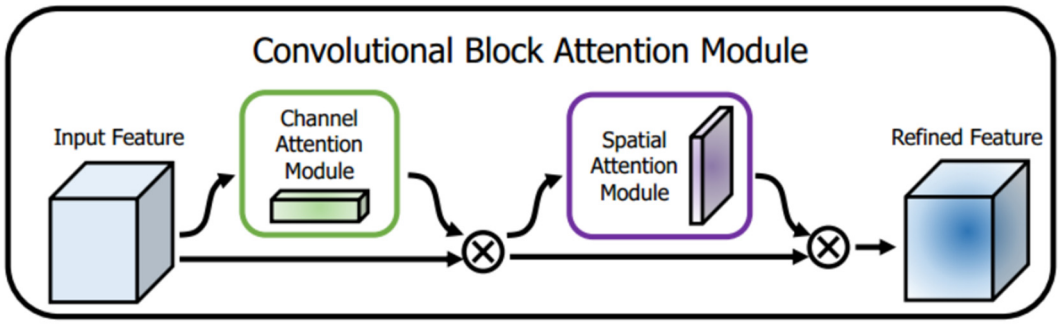

Figure 2. Structure of CBAM

The channel attention (CAM) and spatial attention (SAM) components are shown in Figure 3.

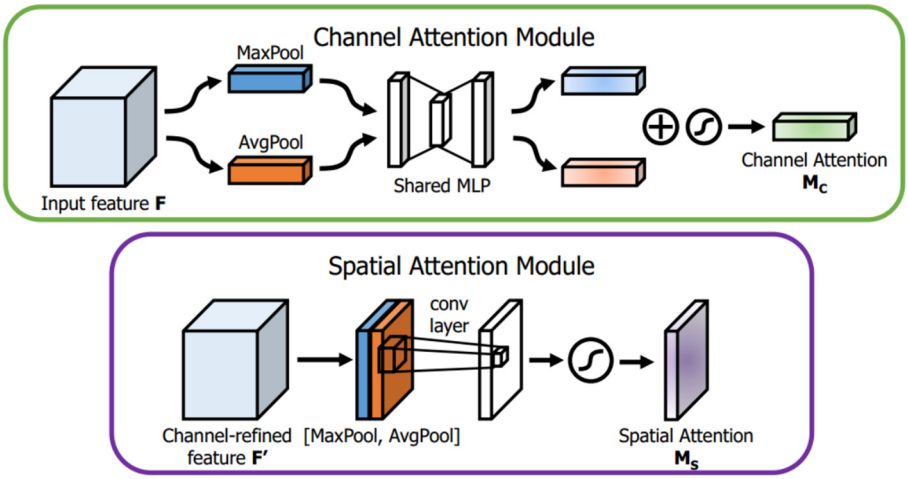

Figure 3. Channel Attention (CAM) and Spatial Attention (SAM)

ISSN: 2688-8645

2.1.3. C2f_Star_CBAM Module

The C2f_Star_CBAM module inherits from the C2f module and implements a mechanism for feature aggregation and processing across multiple layers. This class defines the model's structure, including parameters such as the number of channels (c1, c2), the number of layers (n), and the presence of shortcut connections. While the specific behavior of C2f depends on its internal implementation, it provides a foundational framework for stacking multiple modules. The core processing unit of C2f_Star_CBAM consists of multiple Star_Block_CBAM instances, which integrate convolution operations, the CBAM attention mechanism, and dropout regularization (implemented via DropPath).

The design of C2f_Star_CBAM comprises three key components:

1. **Star-Shaped Dual-Branch Topology:** This combines depthwise separable convolutions with a feature gating mechanism. The main branch employs a 7×7 depthwise separable convolution as the base operator, using grouped convolutions to reduce the number of parameters. The auxiliary branch consists of two 1×1 pointwise convolutions forming a feature transformation path, which dynamically adjusts feature dimensions through an MLP expansion ratio (mlp_ratio=3). A ReLU6-activated gating mechanism then enables feature interaction via element-wise multiplication. This dual-branch structure enhances the detection accuracy for small targets, while the star-shaped topology shortens gradient propagation paths and accelerates model convergence.
2. **CBAM Attention Mechanism:** The CBAM module optimizes the network through joint spatial and channel attention, enabling the model to capture more informative features from both channel and spatial dimensions.
3. **Dynamic Feature Calibration:** Skip connections are introduced to mitigate the vanishing gradient problem, and DropPath regularization is employed to improve model robustness.

In summary, this study replaces the original C2f module in YOLOv8s with the C2f_Star_CBAM module. This modification enhances small-target detection performance while maintaining a balance between model accuracy and computational complexity.

2.2. ResNet-Based Classification Enhancement Module

Considering the characteristics of facial micro-expression recognition—namely, subtle expression textures, susceptibility to illumination variations, and small inter-class differences—this study adopts ResNet50 as the baseline model and introduces the Squeeze-and-Excitation (SE) attention mechanism for improvement. Traditional ResNet50 performs indiscriminate channel-wise feature aggregation, which limits its ability to focus on subtle changes in key facial regions such as the eyebrows and mouth corners during micro-expression recognition. Moreover, it exhibits insufficient robustness to illumination changes and local occlusions. To address these issues, the SE module is incorporated to dynamically enhance the weights of expression-sensitive feature channels while suppressing irrelevant background noise.

The SE-ResNet50 incorporates the following improvements over the original ResNet50:

1. **Integration of SE Modules:** Lightweight SE attention modules are embedded after each residual block. Global average pooling is used to obtain channel-wise statistics, followed by a two-layer fully connected network to model nonlinear inter-channel dependencies.
2. **Adaptive Feature Calibration:** This mechanism enables the network to automatically amplify convolutional responses in key expression regions, such as the eyes and

ISSN: 2688-8645

mouth, significantly improving the ability to capture subtle micro-expression variations.

3. **Efficient Implementation:** The addition of SE modules increases computational cost by only approximately 2%, thereby enhancing feature discriminability while maintaining model efficiency.

3. Experimental Results

3.1. Datasets

(1) WiderFace Dataset

The WiderFace[7] dataset is a subset of the WIDER dataset, released by MMLab in 2016, and is primarily used as a benchmark for face detection algorithms. This dataset contains 32,203 images with approximately 393,703 annotated faces, averaging 12 faces per image. A key feature of the WiderFace dataset is its coverage of diverse scenarios, including both indoor and outdoor environments under various lighting conditions. Additionally, each annotated face is accompanied by detailed attributes such as blur, expression, illumination, occlusion, and pose, as illustrated in Figure 4.

Figure 4. Detailed Attributes Annotated for Each Face

(2) RAF-DB Dataset

In this study, micro-expression recognition is approximated by identifying static frames. Therefore, the micro-expression classification algorithm utilizes the RAF-DB dataset[8], which has been preprocessed and made available on PaddlePaddle AI. The dataset contains 10,749 images spanning seven expression categories. Compared with datasets such as FER2013 and MMAFEDB, RAF-DB offers higher image quality. An example is shown in Figure 5.

Figure 5. Sample Images of the "Angry" Category from RAF-DB

3.2. Experimental Setup

All experiments in this study were conducted on the PyCharm platform, using Python version 3.13.2 and PyTorch framework version 2.6.0, with CUDA version 12.6. The hardware configuration used in this study is summarized in Table 1.

ISSN: 2688-8645

Table 1. Hardware Configuration

Component	Model
CPU	Intel(R)_Core(TM)_i5-10200H_CPU
Memory	16GB
GPU	NVIDIA GeForce RTX 3060 Laptop GPU
GPU Memory	6GB

3.3. Evaluation Metrics

Evaluating model performance during training is crucial. In this study, multiple metrics were used to assess model performance, including mean Average Precision (mAP), F1 score, recall, precision, and ROC-AUC. These metrics comprehensively reflect the model's performance across different aspects, as detailed below:

(1) **mAP (mean Average Precision):** mAP is the average of the Average Precision (AP) across multiple classes (or a single class), reflecting the model's overall detection capability across different object localization accuracies.

(2) **Recall:** Recall measures the proportion of correctly identified positive samples out of all actual positive samples. It quantifies the model's ability to identify true positives and is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

(3) **Precision:** Precision measures the proportion of correctly identified positive samples out of all samples predicted as positive by the model. It reflects the accuracy of the model's positive predictions, defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

(4) **ROC-AUC:** The Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) evaluates the model's overall ability to discriminate between classes by analyzing the true positive rate (TPR) and false positive rate (FPR) across different thresholds. The closer the AUC value is to 1, the better the model distinguishes between different classes.

3.4. Experimental Results

In this study, YOLOv8s was used as the baseline model to conduct comparative experiments on the WiderFace dataset. The specific experimental parameters are listed in Table 4-7, and the experimental results are presented in Table 2.

Table 2. Ablation Experiment Results of YOLOv8s

Model	Dataset	Precision	Recall	mAP@0.5
Basline	WiderFace	0.8117	0.5118	0.584
Baseline+Star_Block	WiderFace	0.8033	0.5103	0.575
Baseline+Star_Block+CBAM	WiderFace	0.8619	0.5780	0.681

The ablation results indicate that introducing only the Star_Block module led to a slight decline in Precision, Recall, and mAP@0.5, suggesting that this module alone may introduce parameter redundancy or feature interference in the original feature extraction network and thus requires complementary optimization strategies. By jointly incorporating Star_Block and CBAM, model performance improved significantly: Precision increased by 5.02%, Recall by 8.45%, and mAP@0.5 by 8.06%. This demonstrates that the CBAM module plays an important role in enhancing contextual awareness and attention weight allocation, effectively mitigating the

ISSN: 2688-8645

negative effects observed when using Star_Block alone. Therefore, the experiments validate the effectiveness of the Star_Block + CBAM approach in face detection tasks.

Additionally, ResNet50 was used as the baseline model for comparative experiments on the RAF-DB dataset. The results are summarized in Table 3.

Table 3. Ablation Experiment Results of ResNet50

Model	Dataset	Precision	Recall	F1	ROC-AUC
Basline	RAF-DB	0.7506	0.7097	0.7269	0.9233
Basline+CBAM	RAF-DB	0.7373	0.6889	0.7083	0.9204
Basline+SE	RAF-DB	0.8739	0.8399	0.8548	0.9209

The ablation results indicate that combining ResNet50 with the CBAM module led to a noticeable decline in F1 score (−1.86%), suggesting that CBAM's effectiveness may have been weakened due to preprocessing steps applied to the RAF-DB images (e.g., center cropping) and the introduction of background noise affecting classification confidence. In contrast, incorporating the SE module significantly improved F1 score, demonstrating that channel-wise feature weighting is well-suited for static facial expressions, and global pooling effectively compresses spatial dimensions to capture subtle cross-region expression patterns. These results validate the effectiveness of the SE module for micro-expression classification tasks.

The conclusions from the ablation experiments support the effectiveness of the proposed improvements. Based on this, the models were further trained to obtain optimal results, as shown in Table 4.

Table 4. Optimal Training Results of the Improved Models

Model	Dataset	Precision	Recall	F1	mAP@0.5
yolov8s-C2f-Star-CAA	WiderFace	0.9033	0.7249	-	0.693
SEResNet50	RAF-DB	0.8739	0.8399	0.8548	-

Overall, the improved models demonstrate strong performance in micro-expression recognition. They exhibit robust feature representation capabilities, effectively capturing subtle facial changes and improving recognition accuracy. Compared with existing methods, the experimental results show significant improvements. Specifically, the YOLOv8s-C2f-Star-CBAM model outperformed the YOLOv8 baseline across all evaluation metrics, particularly in Precision and mAP@0.5, indicating enhanced detection accuracy for facial targets. Similarly, the SE-ResNet50 model outperformed the baseline ResNet50 in all metrics, especially in F1 score, demonstrating superior classification performance. These findings confirm the effectiveness of the proposed model structure enhancements in improving both detection and classification accuracy.

4. Summary

This study addresses the challenges of micro-expression recognition, including data scarcity, rapid temporal variations, and high detection difficulty, by proposing a two-stage recognition framework based on an improved YOLOv8 and ResNet. In the face detection stage, StarNet multi-scale feature fusion and the CBAM attention mechanism are introduced to enhance the accuracy and robustness of small-target detection. In the expression classification stage, the SE attention module is embedded into the ResNet50 network to strengthen the model's focus on key regional features. Experiments conducted on the WiderFace and RAF-DB datasets validate

ISSN: 2688-8645

the effectiveness of the proposed method, demonstrating its advantages in both detection accuracy and classification performance.

Nevertheless, several limitations remain. On the one hand, the two-stage recognition introduces additional inference overhead, which restricts real-time applicability. On the other hand, class imbalance in the RAF-DB dataset hinders the discrimination of certain similar expressions. Moreover, employing static frames as an approximation for micro-expression recognition alleviates data scarcity but fails to fully exploit temporal features. Future research may explore: (1) end-to-end joint training frameworks to mitigate efficiency loss caused by staged processing; (2) lightweight design and model compression techniques to improve real-time applicability; and (3) temporal modeling approaches, such as 3D convolutions or spatio-temporal Transformers, to better capture dynamic characteristics of micro-expressions.

In summary, the proposed improvements achieve notable performance gains in micro-expression recognition, providing valuable insights for further research and practical applications in this field.

References

[1] Saeed U. Facial micro-expressions as a soft biometric for person recognition[J]. Pattern Recognition Letters, 2021, 143: 95-103.

[2] Huang XH, Zhao GY, Hong XP, et al. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns[J]. Neurocomputing, 2016, 175: 564-578.

[3] T. Pfister, X. Li, G. Zhao, et al. Recognising Spontaneous Facial Micro-Expressions[C]. 2011 International Conference on Computer Vision. IEEE, 2011, 1449-1456.

[4] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2.

[5] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[6] Liu W, Chen C, Wong K Y K, et al. Star-net: a spatial attention residue network for scene text recognition[C]//BMVC. 2016, 2: 7.

[7] Yang S, Luo P, Loy C C, et al. Wider face: A face detection benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5525-5533.

[8] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.