Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(4): 5007–5031.
DOI: 10.3934/mbe.2024221
Received: 13 December 2023
Revised: 05 February 2024
Accepted: 06 February 2024
Published: 01 March 2024

Research article

Micro-expression recognition based on multi-scale 3D residual convolutional neural network

Hongmei Jin, Ning He*, Zhanli Li and Pengcheng Yang

College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China

* **Correspondence:** Email: 19208088023@stu.xust.edu.cn.

Abstract: In demanding application scenarios such as clinical psychotherapy and criminal interrogation, the accurate recognition of micro-expressions is of utmost importance but poses significant challenges. One of the main difficulties lies in effectively capturing weak and fleeting facial features and improving recognition performance. To address this fundamental issue, this paper proposed a novel architecture based on a multi-scale 3D residual convolutional neural network. The algorithm leveraged a deep 3D-ResNet50 as the skeleton model and utilized the micro-expression optical flow feature map as the input for the network model. Drawing upon the complex spatial and temporal features inherent in micro-expressions, the network incorporated multi-scale convolutional modules of varying sizes to integrate both global and local information. Furthermore, an attention mechanism feature fusion module was introduced to enhance the model's contextual awareness. Finally, to optimize the model's prediction of the optimal solution, a discriminative network structure with multiple output channels was constructed. The algorithm's performance was evaluated using the public datasets SMIC, SAMM, and CASME II. The experimental results demonstrated that the proposed algorithm achieves recognition accuracies of 74.6, 84.77 and 91.35% on these datasets, respectively. This substantial improvement in efficiency compared to existing mainstream methods for extracting micro-expression subtle features effectively enhanced micro-expression recognition performance and increased the accuracy of high-precision micro-expression recognition. Consequently, this paper served as an important reference for researchers working on high-precision micro-expression recognition.

Keywords: micro-expression recognition; attention mechanism; multi-scale feature extraction; 3D residual convolutional neural network; discriminative network

1. Introduction

Micro-expressions are imperceptible facial expressions that individuals exhibit when intentionally concealing or suppressing their genuine emotions in response to external stimuli. These expressions are characterized by their brief duration and subtle changes [1]. Recognizing micro-expressions poses a significant challenge in affective computing, as it involves identifying minute facial movements that 2 are difficult for humans to perceive within a short time frame (0.25 to 0.5 seconds). It should be noted that micro-expressions primarily originate from the subconscious mind and cannot be consciously hidden or suppressed. Consequently, they genuinely reflect an individual's thoughts and attitudes at a specific moment. Due to their authenticity, micro-expressions find extensive applications in various fields, including clinical psychotherapy [2], criminal interrogation [3], and beyond.

Micro-expression recognition systems typically consist of several stages: image preprocessing, micro-expression localization, feature extraction, and classification. During the preprocessing stage, captured images undergo meticulous enhancement techniques such as noise attenuation, spatial zooming, face region identification, geometric correction, motion amplification, and time-series normalization. Among these stages, feature extraction and classification processes are crucial for improving overall recognition accuracy. Three main strategies are employed for feature extraction: those based on static face image analysis, those utilizing optical flow analysis, and those employing deep learning architectures. In the context of face image-based methods specifically, feature computations often involve extracting differential patterns derived from local binary patterns or using 3D gradient histogram operators as fundamental elements for both detection and detailed feature representation. These approaches effectively capture subtle variations indicative of micro-expressions within facial data.

Local binary pattern (LBP) [4], an operator that characterizes the local texture features of an image, has been widely used. Zhao et al. [5] introduced the local binary patterns on the three orthogonal planes (LBP-TOP) model, which extends the LBP model to three dimensions. This model allows for dynamic encoding of temporal changes by incorporating the time axis and extracting features from local spatiotemporal neighborhoods on the three planes. Wang et al. [6] proposed the LBP with six intersection points (LBP-SIP), which integrates the feature dimensions of LBP-TOP into the LBP-SIP model. This reduces the feature dimension of LBP-TOP to 6 binary bits, minimizing redundant information and reducing the computational time required for model operations. The optical flow-based approach is to extract the non-rigid motion changes of subtle expressions such as similar optical flow or light intensity, or use a facial dynamic map (FDM) to model the motion of facial components, and combine the micro-expression recognition results from the multi-scale sliding window with the dataset samples to complete the classification of micro-expressions. Despite the continuous optimization of traditional recognition algorithms, these algorithms still inherently suffer from a disadvantage in automated feature recognition, which hinders their effective enhancement of classification performance.

In recent years, with the rapid development of computer science and technology, especially in the field of computer vision, the study of micro-expression recognition is no longer limited to the scope of psychology, and more and more researchers have begun to use advanced computer vision technology to assist in micro-expression recognition. Polikovsky et al. [7] used 3D histograms to extract micro-expression features and combined machine learning and micro-expression recognition. Zhao Guoying's team, on the other hand, established the first spontaneous micro-expression database (SMIC) [8], which opens up a new path for exploring micro-expression recognition using deep learning methods. Deep

learning methods are able to achieve end-to-end automatic extraction of deep-level features by virtue of neural network architectures and effectively classify and recognize data accordingly. Gan et al. [9] proposed the optical flow features from apex frame network (OFF-ApexNet), which takes optical flow images between the initial frame and the apex frame as inputs for a convolutional neural network (CNN). Spatial features are then extracted from this optical flow field for recognition purposes. Aside from this method of utilizing CNNs solely for spatial feature extraction, there is also joint extraction of temporal and spatial features. Simultaneous extraction of spatial features over time generally outperforms the method of extracting spatial features alone in terms of recognition accuracy. This include approaches based on long short-term memory (LSTM) [10], as well as the 3D-CNN algorithm that extends the spatial domain convolution of CNNs to the temporal domain. Meanwhile, notable advancements have been made in the development of micro-expression datasets. A comprehensive survey [11] delves into the various datasets, features, and algorithms pertinent to micro-expression analysis. The recently introduced a third generation facial spontaneous micro-expression database (CAS (ME)3) database [12] stands out as a cutting edge third-generation resource for facial spontaneous micro-expressions, significantly enhancing the availability of data for recognition tasks. Furthermore, a spontaneous 4D micro-expression dataset (4DME) dataset [13], representing a spontaneous four-dimensional micro-expression repository with multimodal information, serves as a valuable asset for deep exploration of the spatio-temporal characteristics of micro-expressions.

However, compared with macro-expressions, the number of samples in the dataset of micro-expressions is relatively small, and the current micro-expression recognition methods of various types fail to provide ideal solutions when facing problems such as changes in local illumination of the face. At the same time, due to the limited number of samples and the uneven distribution of the samples, the proposed feature extraction method performs poorly in terms of recognition robustness. Addressing the issue of datasets lacking apex frame labels, Liu et al. [14] employed a lightweight network known as SqueezeNet to effectively localize the apex frame for such datasets. Additionally, they utilized 3D convolutional networks for spatio-temporal feature extraction. To overcome the challenge of focusing on the detailed local features of micro-expressions, Zhao et al. [15] proposed a deep prototypical learning framework, namely ME-PLAN. This framework utilizes expression-related knowledge transfer and scenario training to accurately learn micro-expression features.

In the realm of large-scale micro-expression recognition studies, especially under stringent conditions requiring high precision recognition, it is critically important to construct computationally efficient and resilient recognition models leveraging computer technology. Despite the merits that shallow networks possess in managing costs and achieving swift processing, they inherently struggle with extracting intricate and ephemeral micro-expression characteristics, thus hindering their capacity to significantly boost recognition accuracy. Therefore, in the study of high-precision micro-expression recognition, we focus more on how to improve the recognition accuracy of the model by improving the algorithm or adopting the depth structure, compared with the recognition speed. Considering the complex spatial and temporal characteristics of micro-expressions, this study introduces a micro-expression recognition framework based on a multi-scale 3D residual CNN. The proposed framework consists of a micro-expression facial motion feature extraction network, a multi-scale spatiotemporal 3D CNNs (3D-ResNet50) fusion classification network, and a discriminative network. The ResNet50 fusion classification network and discriminative network contribute to the method's exceptional performance in terms of recognition accuracy, as demonstrated through experimental validation. Overall, the

contributions of this research can be summarized as follows:

1) The backbone network employed in this study is the deep 3D-Resnet50, with micro-expression optical flow feature maps serving as input to the network model. This choice aims to simplify the judgment basis for different micro-expressions, thereby reducing complexity.

2) For spatial integration, a multi-scale approach is adopted by incorporating multi-scale convolution modules of various sizes into the network. This integration enables the fusion of both global and local information. Regarding temporal integration, the attention feature module is utilized to combine feature maps obtained from different layers of the model. This integration enhances the model's context-awareness.

3) To ensure optimal performance, a discriminative network is designed to select the most suitable solution among various outputs. The chosen solution is then presented as the final result of this network.

The rest of the paper is organized as follows: Section 2 describes the related work. Section 3 describes the proposed micro-expression recognition algorithm in detail, and Section 4 describes the experimental results and analysis. Finally, the paper is summarized in Section 5.

2. Related work

In the study of micro-expression recognition, feature extraction is a crucial step to improving recognition accuracy. Huang et al. [16] proposed the spatio-temporal complete local quantization pattern (STCLQP), which goes beyond LBP-TOP by enhancing the capture of input information by integrating the pixel differences in sign, magnitude, and orientation, and constructs a compact spatio-temporal domain codebook to optimize the recognition results. Huang et al. [17] further integrated face shape attributes into spatio-temporal texture features and proposed a spatiotemporal local binary pattern (STLBP) to extract recognition information for facial micro-expression recognition. Li et al. [18] proposed a new method to detect the apex frame by estimating pixel-level change rates in the frequency domain, which focuses on the emotional information carried by the peak frames and exploits the pixel-level rate of change in the frequency domain to locate peak frames, and this method performs better in determining key frames. In addition, they combined local and global information under peak frames for joint feature learning, which enables the model to focus on key facial regions rich in emotional information and suppresses the influence of irrelevant regions.

Given that optical flow can infer relative motion information between different frames, some researchers have started using optical flow-based methods to extract motion-related information from micro-expression videos or sequences for micro-expression recognition. Liu et al. [19] introduced the main directional mean optical flow (MDMO) method, which utilizes optical flow to construct a region-of-interest-based feature vector describing the local motion of a micro-expression on the face. This feature vector is then inputted into a support vector machine for micro-expression recognition. Building upon this work, Liong et al. [20] proposed the bi-weighted oriented optical flow (Bi-WOOF) feature descriptor, which represents a subtle micro-expression sequence using only two frames, namely, the start and peak frames. The Bi-WOOF method incorporates both the optical flow magnitude and the optical strain magnitude as weights to generate directional histograms of face region blocks, thereby emphasizing the importance of each optical flow for micro-expression recognition. Furthermore, Ni et al. [21] introduced the LGSNet, a novel dual-stream network that combines optical streams and segment-level features from the video. This fusion of information enhances the recognition capabilities of the local suppression and global enhancement spotting network (LGSNet) for micro-expression analysis. Li et al. [22] proposed a

micro-expression recognition method that utilizes deep multitask learning to localize facial landmarks and segment the facial region into regions of interest (ROIs). Since the movement of facial muscles generates micro-expressions, a robust optical flow approach is combined with histograms of oriented optical flow (HOOF) [23] features to assess the direction of movement of facial muscles. SVMs are then employed as classifiers for micro-expression recognition. The facial action coding system (FACS), an important tool for recognizing micro-expressions, necessitates motion recordings at various facial locations, such as eyebrows and corners of the mouth. In this particular method, ROIs and HOOF features are used in conjunction, ultimately resulting in accurate micro-expression recognition that corresponds to action units (AUs).

With the development of deep learning technology, a series of deep neural network structures have been applied to the field of micro-expression recognition. Zhou et al. [24] used a dual-stream inception network, focusing on acquiring salient and discriminative features of a specific expression and better recognizing micro-expressions by fusing the features of a specific micro-expression. Liong et al. [25] proposed a shallow three-stream 3D CNN (STSTNet) to embed spatial and temporal information into micro-expression video clips, which learns from three optical flow features (i.e., optical strain, horizontal and vertical optical flow fields) computed based on the start and apex frames of each video and extracts discriminative high-level features and micro-expression details through lightweight computation. Li et al. [26] proposed a dual-stream convolutional network that utilizes a multi-scale three-dimensional (M3D) convolutional layer to build temporal streams in a 2D CNN to efficiently extract spatio-temporal features required for video character re-identification. This M3D convolution is designed to be compact and easy to optimize, and it enhances the ability to learn multi-scale temporal features while maintaining a low number of parameters compared to traditional 3D convolutional networks. Li et al. [27] proposed a deep local holistic network to extract locally enriched spatio-temporal and global features for micro-expression recognition through subnetwork fusion. The attention mechanism can effectively improve the network model by assigning higher weights to important features in the image and introducing the attention mechanism in the channel and spatial dimensions of the data to suppress the background interference, thus enhancing the network's ability to perceive key features. Quang et al. [28] proposed a CapsuleNet-based micro-expression recognition method that innovatively utilizes the "capsule network" structure to effectively overcome the problem of information loss (e.g., relative positional information) caused by the maxpooling operation in traditional image processing when dealing with micro-expression sequences of apex frames. (e.g., relative position information). By precisely maintaining the spatial relationship between features, recognition accuracy and robustness are significantly improved. Xia et al. [29] introduced the concept of spatio-temporal transformations for the first time and constructed the spatiotemporal recurrent convolutional networks (STRCN) model, which skillfully integrates spatial features and dynamic changes through the dual processing of the time dimension and strongly enhances the ability to parse the subtle and fast facial muscle movements.

Zhang et al. [30] proposed a spatio-temporal transform network based on the long- and short-term dependence of facial expression sequences and time-dependent spatio-temporal transformer architecture, which includes several key modules such as a spatial encoder, a temporal aggregator, and a classifier. This work is unique in micro-expression recognition research because it completely abandons CNNs and pioneers the application of transformers to the micro-expression recognition task. This novel architectural design allows the model to efficiently capture and analyze long-distance dependencies, solving the challenge of micro-expression recognition due to the transient and imperceptible nature of the expression, but its limitations include the ability to generalize to large and diverse datasets,

computational efficiency, and under-utilization of prior knowledge, as well as the model's interpretive nature and stability under low-quality inputs. Su et al. [31] used a component-aware attention module to highlight relevant regions of micro-expressions to efficiently capture motivational information and non-rigid deformations. Gajjala et al. [32] proposed a 3D residual attention network model based on facial micro-expression recognition using 3D residual attention network (MERANet) to learn deeper, finer-grained subtle features to classify emotions by taking advantage of the joint strengths of spatial-temporal attention and channel attention. Although MERANet performs well on a limited number of datasets, the problem of insufficient sample size prevalent in micro-expression datasets still constrains the model's generalization ability. Zhang et al. [33] proposed a deformation repair network (DINet) to implement a visual dubbing technique for face-to-face in high resolution that performs spatial deformation on the feature map of the reference face image and uses the spatially deformed features to repair the mouth region, but the model may have difficulty accurately capturing and representing the corresponding details of the mouth movement when dealing with complex, variable, or low-quality audio signals. Zhou et al. [34] proposed the micro-expression recognition dual branch attention network (Dual-ATME) to solve the problem of ineffective single-scale features representing ME, but, limited by the reliance on a priori knowledge, the ability to dynamically adjust and adaptively learn in rapidly changing and complexly represented micro-expression scenarios is still a challenge.

Current mainstream micro-expression recognition algorithms have made great progress in micro-expression micro-feature detection through various optimization methods. However, in the design of micro-expression recognition algorithms, in addition to the pursuit of powerful feature extraction, how to deeply understand and effectively utilize these features, especially the robustness and generalization ability in different contexts, is also the key to determining the efficacy of the algorithms. Meanwhile, in the face of challenges such as insufficient samples and rapid changes unique to micro-expression recognition, future research needs to further explore more efficient feature expression and model optimization schemes.

3. Proposed method

3.1. Microexpression recognition framework

The algorithm proposed in this paper comprises three main components: the micro-expression facial motion feature extraction network, the multi-scale 3D-ResNet50 fusion classification network, and the discriminative network. These components collaborate within the framework, as depicted in Figure 1. Initially, the micro-expression facial motion feature extraction network is employed to preprocess the data, minimizing the influence of both extrinsic and intrinsic factors. Subsequently, the micro-expression motion feature maps, obtained from the preprocessing step, are inputted into the multi-scale 3D-ResNet50 fusion classification network. This network generates numerous results. Finally, the discriminative network evaluates the multiple outcomes from the classification network to determine the optimal solution and produce the final results.

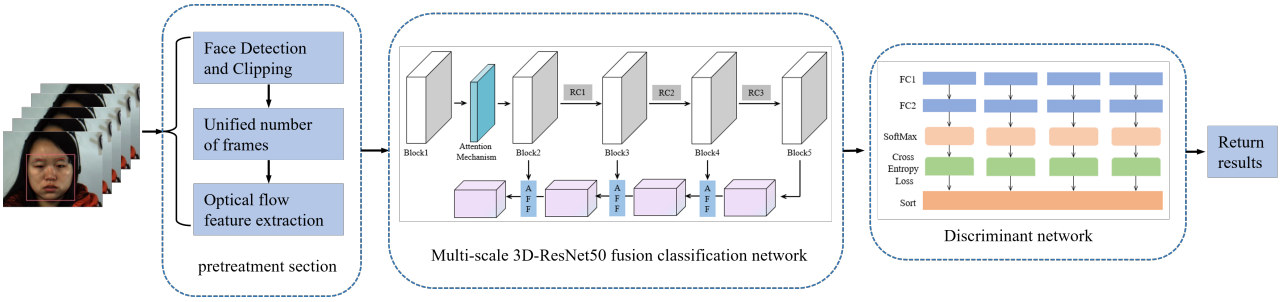


Figure 1. Micro-expression recognition framework based on multi-scale 3D residual CNN.

3.2. Video image preprocessing

(1) Face spotting and cropping

To eliminate irrelevant distractions such as background details, the dlib 68-point face spotting algorithm [35] is utilized for face cropping. This algorithm employs histogram-oriented gradient (HOG) features to initially identify the facial contour, mouth, nose, eyes, and eyebrows using 68 reference points. Subsequently, the OpenCV tool is employed to precisely crop the designated area.

(2) Data enhancement

The original dataset underwent a process of database expansion, wherein the resulting face frame was shifted by 15 pixels in the left, right, upward, and downward directions. Additionally, a vertical flip was applied to further augment the dataset, resulting in a size that is four times larger than its original.

(3) Keyframe extraction

The video sequences of micro-expressions in the datasets have different frame lengths, and some datasets are too long and contain a lot of useless information, which makes the task of micro-expression recognition very difficult. In this paper, we use a key frame extraction technique based on spatio-temporal slices. Based on the feature analysis of the spatio-temporal slice texture, the motion state of the image acquisition device is reflected as the tilt change of the spatio-temporal slice texture, and then the key frames that can accurately describe the motion state of the image acquisition device are determined by measuring the pixel proximity of the adjacent spatio-temporal slices and by using the nearest pixel matching method.

For an image dataset containing frames, the horizontal spatial-temporal slice is a combination of one row of pixels continuously extracted at a fixed coordinate, where the horizontal slice extracted from the ith frame can be expressed as:

$$h_i = (P_i(1, y_k), \cdots P_i(x, y_k), \cdots P_i(m, y_k)) \ (0 \leqslant i \leqslant N, 0 \leqslant k \leqslant n), \tag{3.1}$$

where (x, y) denotes the image dimension, the size of each frame is $m \times n$, and $P_i(x, y_k)$ denotes the pixel value at (x, y_k) in the ith frame. The spatio-temporal slices extracted from the same place in each frame are merged together in frame order to obtain a spatio-temporal slice image.

(4) Local image enhancement

Euler's algorithm amplifies brightness variations instead of magnifying motion. This complex algorithm consists of four steps: spatial filtering, temporal filtering, amplification, and reconstruction. Spatial filtering involves creating a Laplacian or Gaussian pyramid for each frame in the input video.

Each level of the pyramid has different spatial frequencies and signal-to-noise ratios, with higher frequencies and ratios decreasing as you go down the pyramid. Spatial decomposition is important because, beyond a certain threshold of high spatial frequency, substituting motion with brightness changes becomes inaccurate. Therefore, the amplification coefficient for the high spatial frequency level is relatively small. Temporal filtering applies a Fourier transform to the frequency domain of a specific fixed pixel, highlighting brightness fluctuations over time and generating the input signal. Only the frequency band that requires amplification is retained in the frequency domain, while other bands are reduced to zero through band-pass filtering. Amplification is then performed at each level of the pyramid, with the amplification coefficient varying depending on the spatial frequency. The amplified portion is superimposed onto the section before the time-domain filtering. Finally, the pyramid is reconstructed, resulting in the final amplified video.

The mathematical principle of Euler's video magnification algorithm is as follows: Let $I(x, t)$ denote the brightness of the pixel at the position at x and time at t and $\delta(t)$ denotes the displacement of the object and the magnification factor. The following equation can be obtained.

$$\begin{cases} I(x, 0) = f(x), \\ I(x, t) = f(x + \delta(t)). \end{cases} \tag{3.2}$$

The following amplification function can be obtained by bringing the amplification factor α into the above Eq (3.3).

$$\hat{I}(x, t) = f(x + (1 + \alpha)\delta(t)). \tag{3.3}$$

Expanding $I(x, t)$ with the first-order Taylor's formula gives.

$$I(x, t) \approx f(x) + \delta(t)\frac{\partial f(x)}{\partial x}. \tag{3.4}$$

The Taylor series expansion is not applicable for high spatial frequencies, which is obtained after time domain filtering.

$$B(x, t) = \delta(t)\frac{\partial f(x)}{\partial x}. \tag{3.5}$$

It is obtained after zooming and superimposition:

$$\tilde{I}(x, t) = f(x) + (1 + \alpha)\delta(t)\frac{\partial f(x)}{\partial x}. \tag{3.6}$$

It is also obtained from the first order Taylor series expansion:

$$\tilde{I}(x, t) = f(x + (1 + \alpha)\delta(t)) = \hat{I}(x, t). \tag{3.7}$$

Represented as a simple one-dimensional function as shown in Figure 2, the luminance varies in a cosine fashion over the null domain. The meaning represented by each different colored line is given in the figure.

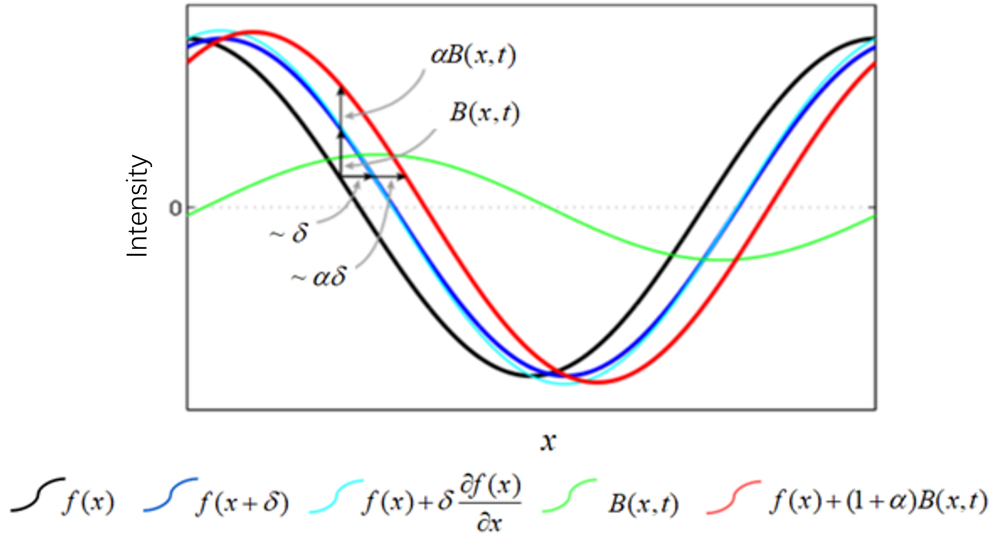


Figure 2. One-dimensional functional representation of the Euler amplification function.

The bottom side of the right triangle in the figure represents the motion $\delta(t)$, the vertical side represents $B(x,t)$, and the hypotenuse represents the gradient $\partial f(x)/\partial x$. The intended amplification motion is $\delta(t)$, but what we are actually amplifying is $B(x,t)$ because there is such a relationship between the two, as $B(x,t) \approx \delta(t)\partial f(x)/\partial x$ allows this approximate amplification to hold.

(5) Optical flow feature extraction

The optical flow method, specifically the total variation (TVL1) optical flow method, is utilized for micro-expression motion feature extraction due to the inherent complexity of micro-expressions. This method helps to improve the signal-to-noise ratio of the original dataset after performing feature extraction.

3.3. Multi-scale 3D-ResNet50 fusion classification network

3.3.1. Overall network framework

The main purpose of the multi-scale 3D-ResNet50 fusion classification network is to perform further feature extraction on the optical flow feature maps. The recognition process of micro-expressions has the correlation continuity of different feature maps in addition to analyzing the spatial features of a single feature map. In this paper, the 3D-ResNet50 network, used as the recognition backbone, has an extra-temporal dimension compared to the general ResNet50 network. The input of the network is $134 \times 134 \times 16$, 134×134 is the feature map's spatial dimension, and 16 is the temporal feature of the feature map. The network is divided into 5 blocks; the 1st block is a $3 \times 7 \times 7$ convolutional layer, and the number of convolutional layers in the 2nd to 5th blocks is 6, 8, 12, and 6, respectively, where every 2 convolutional layers form a residual structure. The residual structure can effectively prevent the deep network model from gradient vanishing and gradient explosion phenomenon, which improves the performance of the network to some extent. In this paper, the temporal and spatial attention mechanism is added after the 1st block of this network to improve the network's attention to the temporal dimension and spatial dimension of micro-expression movement; filters of different sizes (RC1, RC2, RC3) are

added after the 2nd, 3rd, and 4th blocks of the network to extract spatial multi-scale features, which constitutes a multi-scale feature extraction network; and, finally, in order to prevent temporal dispersion, the results obtained from Block2, Block3, Block4, and Block5 are fused with the results of the upper layer after back-convolution to form a multi-scale constitutive time fusion network. The results of each layer are outputted separately and the discriminative network is used to decide the best output. The network structure is shown in Figure 3.

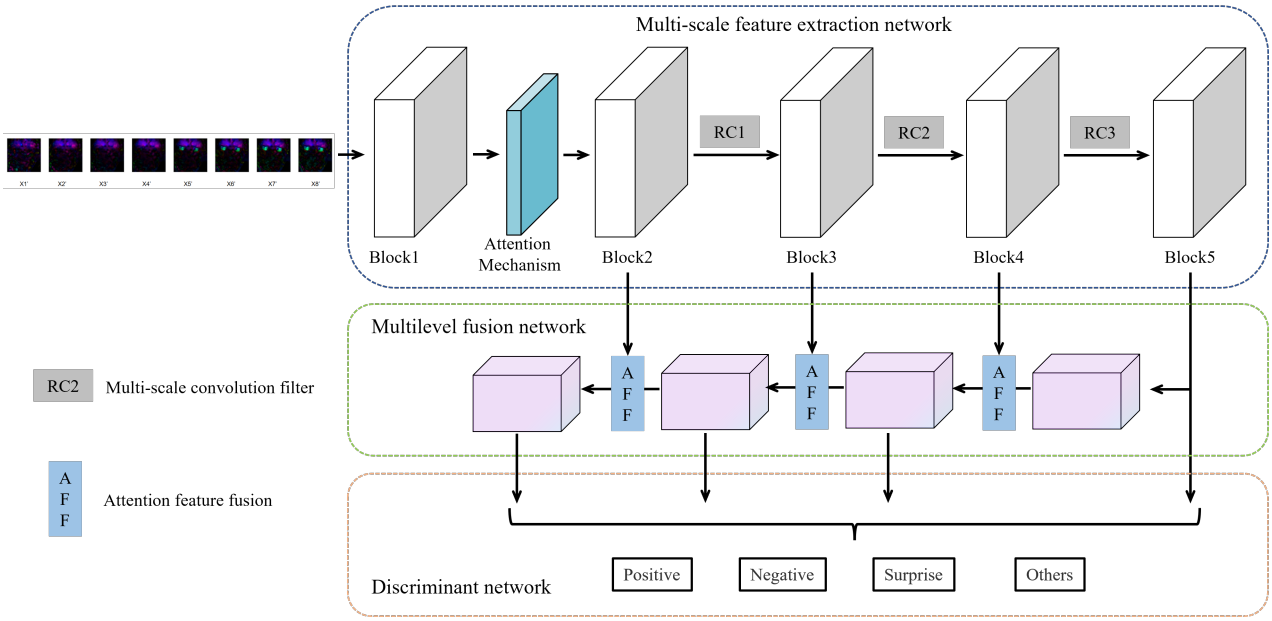


Figure 3. Multi-scale 3D-ResNet50 fusion classification network.

3.3.2. Multi-scale extraction module

In this study, additional multi-scale extraction modules are integrated into the 2nd, 3rd, and 4th blocks of the 3D-ResNet50 model. These modules consist of convolutional kernels of varying sizes, as shown in Figures 4, 5, and 6. Three unique multi-scale filters, denoted as RC1, RC2, and RC3, are designed in this paper. It is worth noting that as the depth increases, the complexity of the filter's structure also increases to effectively capture smaller features.

As shown in Figure 4, RC1 has 4 branches. The first branch consists of a $1 \times 1 \times 1$ convolutional layer for extracting smaller-sized features. The second branch includes a $1 \times 1 \times 1$ convolutional layer and a $3 \times 3 \times 3$ convolutional layer. The $1 \times 1 \times 1$ convolutional layer downscales the feature maps, while the $3 \times 3 \times 3$ convolutional layer extracts larger-sized features. The third branch is similar to the first branch but with a double-layer substitution. Instead of using a single $5 \times 5 \times 5$ convolutional layer, two stacked $3 \times 3 \times 3$ convolutional layers are used. This not only reduces computation but also increases the depth of convolution. The fourth branch integrates global spatial information using a global average pooling layer.

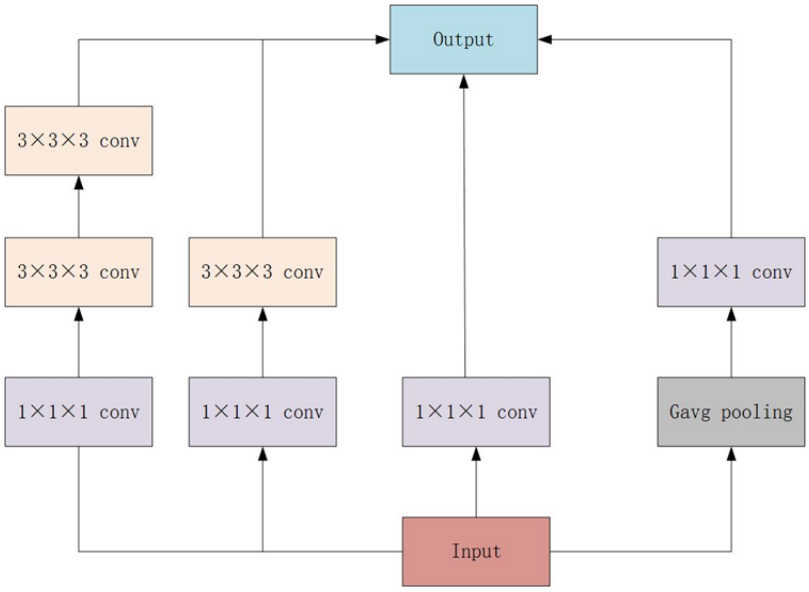


Figure 4. Multi-scale RC1 module.

Finally, the multi-scale features are obtained by adaptive weighted feature fusion of the feature maps from the four branches, as defined by Eq (3.8).

$$K = \alpha \bullet X1 + \beta \bullet X2 + \gamma \bullet X3 + \delta \bullet X4, \tag{3.8}$$

where K denotes the fusion feature map, $\alpha, \beta, \gamma, \delta$ denote the feature weights of the four branches, and the weights are updated by network adaptive learning. $X1$, $X2$, $X3$, and $X4$ denote the output features of the four branches.

As shown in Figure 5, RC2 has 4 branches. Two branches use larger convolution kernels, while the other two keep the $1 \times 1 \times 1$ convolution kernel and global average pooling layer unchanged to extract smaller features and integrate global spatial information. The third branch selects a $7 \times 7 \times 7$ convolution kernel with a larger size. To reduce the amount of computation, we use the strategy of depth-separable convolution by replacing the original $7 \times 7 \times 7$ convolution with $1 \times 1 \times 7$, $1 \times 7 \times 1$, and $7 \times 1 \times 1$ convolution, through which the amount of computation can reduce the computation to $1/3$ of the original without affecting the accuracy. RC2 uses a larger convolution kernel at a deeper level, which pays more attention to the overall feature layout, and the multi-scale idea of the model is also more prominent. The last branch takes the idea of using double layer replacement by replacing the $9 \times 9 \times 9$ convolutional kernel with two $1 \times 1 \times 7$, $1 \times 7 \times 1$, and $7 \times 1 \times 1$ convolutional stacks to capture larger size features.

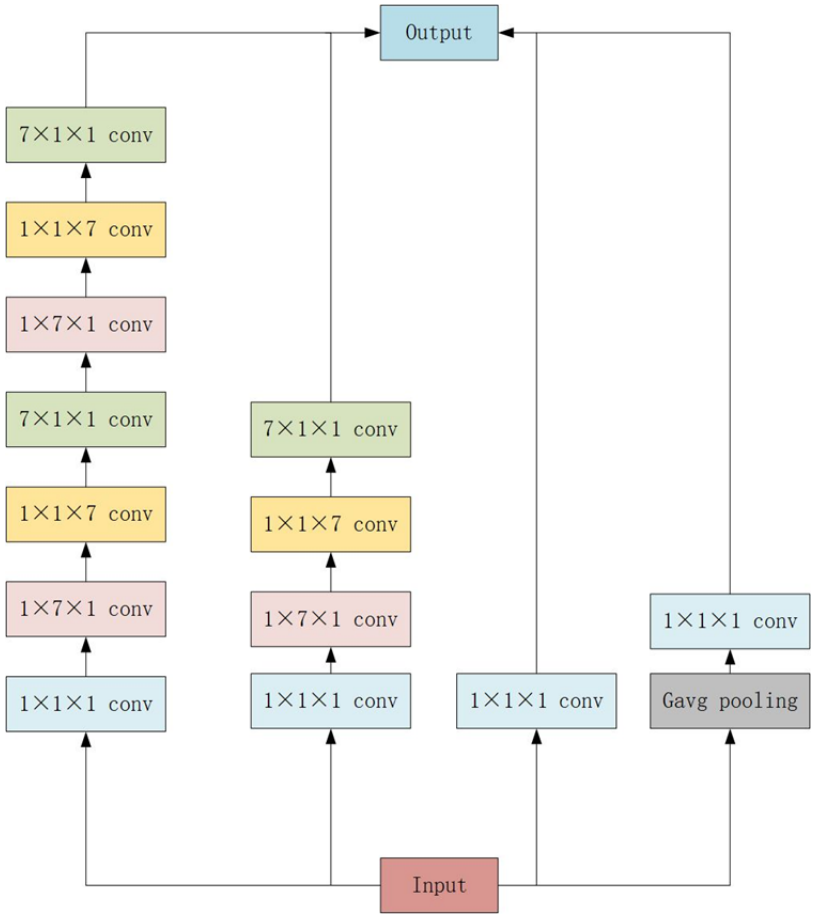


Figure 5. Multi-scale RC2 module.

Finally, the final multi-scale feature map is obtained by adaptively weighted feature fusion of the feature maps of the four branches, according to Eq (3.8) above.

As shown in Figure 6, RC3 also has 4 branches, but it is more complicated; the first branch is a large-size, different-direction multi-scale convolutional branch. It undergoes a $1 \times 1 \times 1$ convolutional kernel for dimensionality reduction and then passes through depth-separable convolutional layers of $1 \times 1 \times 3$, $1 \times 3 \times 1$, and $3 \times 1 \times 1$. The result obtained from this branch is then subjected to transverse convolution ($1 \times 1 \times 3$), longitudinal convolution ($1 \times 3 \times 1$), and radial convolution ($3 \times 1 \times 1$) for feature subdivision. The obtained feature maps in different directions are outputted by feature fusion. The second branch is a smaller-size, different-direction multi-scale convolutional branch. It is first downscaled by a $1 \times 1 \times 1$ convolution kernel and then directly downscaled by horizontal convolution ($1 \times 1 \times 3$), vertical convolution ($1 \times 3 \times 1$), and radial convolution ($3 \times 1 \times 1$) for feature subdivision. The obtained feature maps in different directions are subjected to feature fusion for output. The 3rd and 4th branches are the same as RC1 and RC2, consisting of a $1 \times 1 \times 1$ point-by-point convolution layer and a global average pooling layer.

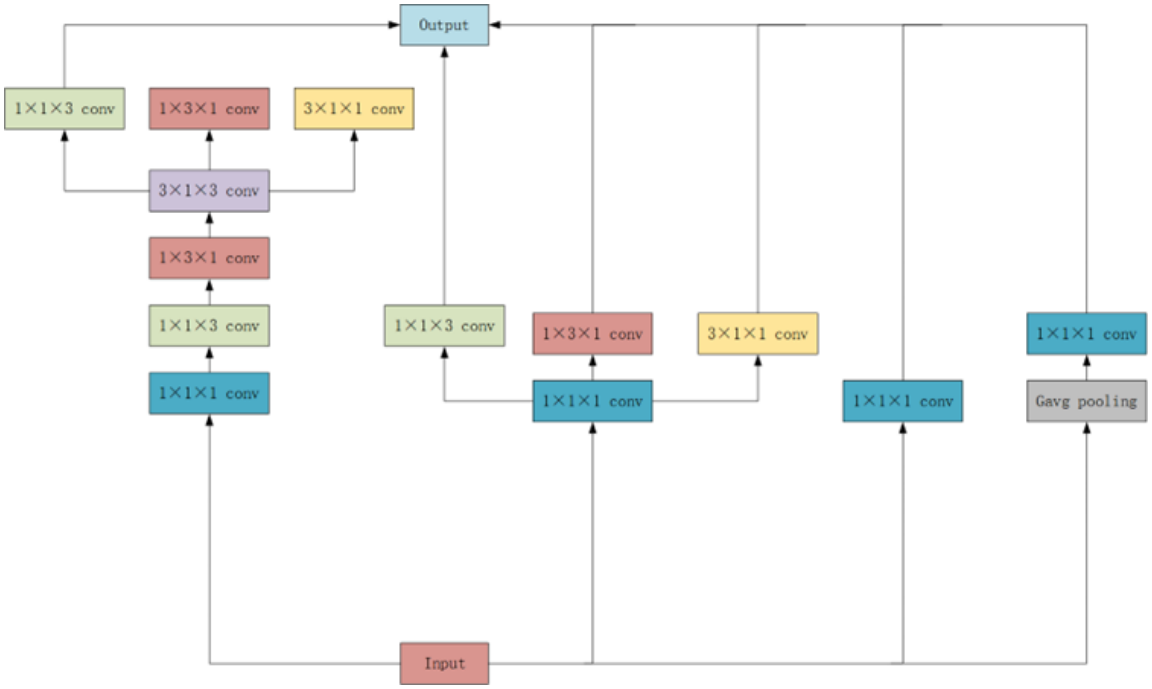


Figure 6. Multi-scale RC3 module.

Finally, the feature maps of 8 out of the 4 branches are adaptively weighted according to Eq (3.9) for feature fusion to obtain multi-scale features. The feature maps in branch 1 and branch 2 are realized by the padding operation to achieve equal size with the output feature maps of other branches.

$$K = \alpha \bullet X1 + \beta \bullet X2 + \gamma \bullet X3 + \delta \bullet X4 + \varepsilon \bullet X5 + \varphi \bullet X6 + \phi \bullet X7 + \partial \bullet X8, \tag{3.9}$$

where K denotes the fusion feature map, $\alpha, \beta, \gamma, \delta, \varepsilon, \varphi, \phi, \partial$ denotes the feature weights of the 8 outputs in the 4 branches, and the weights are updated by adaptive learning of the network. $X1$, $X2$, $X3$, $X4$, $X5$, $X6$, $X7$, and $X8$ denote the 8 output feature maps in the 4 branches.

The standard for recognizing micro-expressions involves not only detecting changes in individual facial features such as the mouth, eyes, and nose, but also considering the overall change in the person's expression. This combination of local and overall features is crucial. The use of different sizes of convolution kernels, such as $1 \times 1 \times 1$ and $5 \times 5 \times 5$ compared to the standard $3 \times 3 \times 3$ kernel, allows for a more comprehensive extraction of features, resulting in a richer feature map. Meanwhile, at the layer of RC2, as the network deepens, the size of the feature map decreases, and a $9 \times 9 \times 9$ kernel is sufficient to capture most facial features. This layer also incorporates feature extraction by linking the mouth and nose or the nose and eyes, along with global average pooling for integrated analysis. Multiple layers are used in the final output because some layers can already make reasonable predictions for classification. The final layer, RC3, uses a $5 \times 5 \times 5$ convolution kernel to extract features in all directions, combined with global features, enabling accurate micro-expression classification.

3.3.3. Multi-scale temporal feature fusion network

In this paper, a multi-scale temporal feature fusion network is devised using the outputs of various blocks of the 3D-ResNet50. These features are combined across different scenes. The outputs of the 2nd, 3rd, 4th, and 5th blocks are denoted as $X2$, $X3$, $X4$, and $X5$. Among these, $X5$ is chosen as the first preselected output, and its size is adjusted to match that of $X4$ through inverse convolution. This adjusted $X5$ is then used as the second preselected output through attentional feature fusion. Similarly, $X4$ is fused with $X3$ through attentional feature fusion after inverse convolution, resulting in the third preselected output. The fourth preselected output is obtained by fusing $X3$ after inverse convolution with $X2$ through attentional fusion. These four preselected outputs form the multi-scale temporal feature fusion network. Finally, these outputs are fed into the final discriminative network to identify the optimal solution. The fusion method employed here is the attention feature fusion mechanism (AFFA), which will be elaborated upon in the subsequent sections.

The attention feature fusion module incorporates the channel attention mechanism (CAM) by utilizing two branches with varying scales to extract channel attention weights. One of the branches employs global average pooling to extract the attention of the global average feature, as computed in Eq (3.10).

$$L(X) = B(f^{1\times1\times1}(\delta(B(f^{1\times1\times1}(AvgPool(X)))))), \tag{3.10}$$

where $f^{1\times1\times1}$ point-by-point convolution reduces the number of channels of the input feature X to the original $1/r$, B denotes the BachNorm layer, δ denotes the Relu activation function; the number of channels is restored to the original value by point-by-point convolution in the second layer, and r denotes the channel scaling ratio.

Another branch uses global max pooling to extract the attention of the global maximal features, also through 2-layer point-wise conv for channel attention scaling. Finally, the two are added together to calculate the weights through the Sigmoid function and multiplied with the original X to get the final X', which constitutes the final attention feature network that is structured as shown in Figure 7 below.

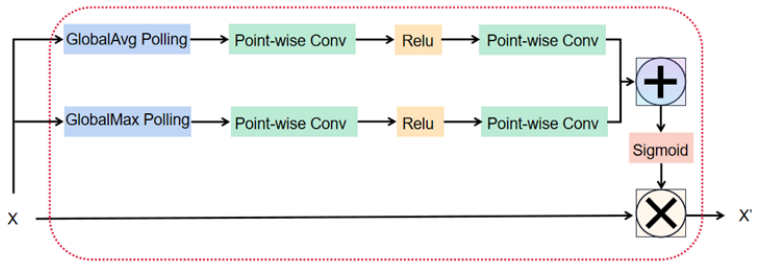


Figure 7. Attention feature network.

In this paper, two attentional feature networks are used to combine into an attentional feature fusion mechanism AFFA. One AFFA corresponds to 2 fusion operations, and the fusion operation is computed as shown in Eq (3.11).

$$Z = CAM(X + Y) \times X + (1 - CAM(X + Y)) \times Y, \tag{3.11}$$

where X and Y denote the feature maps in different scenarios, corresponding to the outputs of different block blocks in the multi-scale 3DResNet50 network, respectively. The final result of Z is the output after two fusion operations. The structure of the attention feature fusion mechanism is shown in Figure 8.

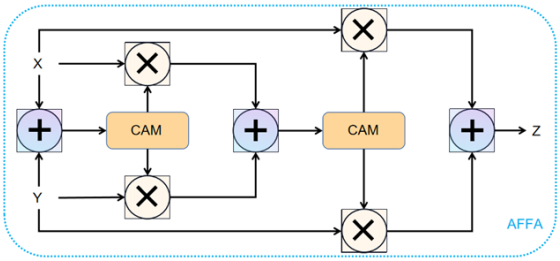


Figure 8. Mechanisms for fusion of attentional features.

By employing a multi-scale temporal feature fusion network, the model integrates the outputs of the 5th block with the outputs of the 4th block. Similarly, it fuses the outputs of the 4th block with the outputs of the 3rd block, and the outputs of the 3rd block with the outputs of the 2nd block. This fusion process results in three temporal fusion outputs. Additionally, the outputs of the 5th block, obtained through the modeling routine, contribute to a total of four outputs, which are subsequently fed into the final discriminative network.

3.4. Discriminant networks

The primary objective of the discriminative output network is to determine the most optimal solution among the four obtained results. The multi-scale time fusion network generates four 5-dimensional tensors, each dimension representing batch size, channel, depth, height, and width. These tensors cannot be directly compared to identify the optimal solution. Therefore, it is necessary to pass the four tensors through two fully connected layers to obtain probability distribution vectors. The SoftMax function is then applied to map these vectors to the range $[0-1]$. Finally, the cross-entropy loss function is utilized to calculate the loss value for each of the four results.

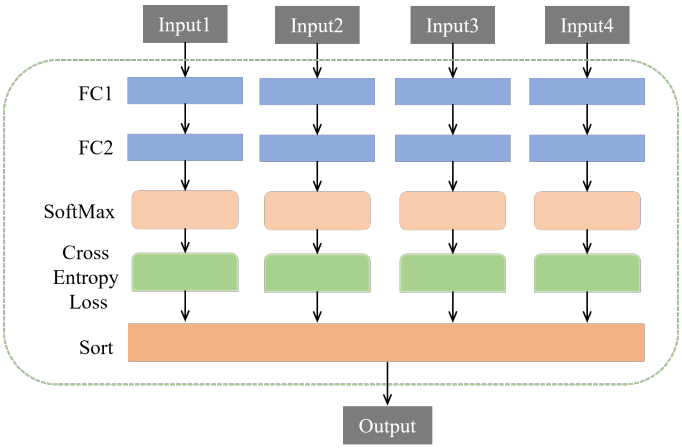


Figure 9. Discriminant network.

4. Experiments

4.1. Dataset

In this study, we utilize the SMIC dataset, SAMM dataset, and the CASME II dataset for conducting our experiments. The SMIC dataset is constructed using a high frame rate camera and includes a subset of hyperspectral (HS) data. It consists of a total of 164 micro-expression video segments, which are further divided into three categories: negative, surprise, and positive. The negative category comprises 70 video segments, while the surprise category contains 43 segments. The positive category, on the other hand, encompasses 51 video segments. The SAMM dataset consists of 159 micro-expression videos with 32 participants and is categorized into 8 classes: happiness contains 26 videos, fear contains 8 videos, surprise contains 15 videos, anger contains 57 videos, disgust contains 9 videos, sadness contains 6 videos, contempt contains 12 videos, and other contains 26 videos. The happiness category is categorized as positive in this paper's experiment. The angry, fear, disgust, sadness and contempt categories are uniformly classified as negative categories. The surprise category is categorized as surprise. The other category is categorized as the other category. In contrast, the CASME II dataset consists of 246 micro-expression video segments, which are categorized into five distinct categories: happiness, surprise, disgust, repression, and others. The happiness category includes 32 video segments, whereas the surprise and disgust categories comprise 25 and 63 segments, respectively. The repression category consists of 27 segments, and the other category contains 99 video segments. Notably, in our experiment, the happiness category is classified as positive, the disgust category as negative, and the surprise category is labeled as surprise. Furthermore, the repression and other categories are categorized as other.

4.2. Experimental setup

The testing environment for this experiment is Windows 11 (a 64-bit operating system), AMD Ryzen 7-5800H 3060 @ 1.90 GHz, and 16 GB of memory. Pytorch was utilized to construct the model in this study, with the programming language being Python 3.6. The initial learning rate was set to 0.00001, the training epoch was set to 100, and the optimizer employed was Adam. The loss function used was cross-entropy loss. The evaluation of the experimental results was performed with accuracy. Let N represent the total number of samples. Accuracy is determined by two values: true positives (TP), which represents the number of dataset labels that are true and the model's recognition results that are also true, and true negatives (TN), which represents the number of dataset labels that are false and the model's recognition results that are also false. By obtaining these four values, the accuracy rate of the experimental model can be calculated using Eq (4.1).

$$Accuracy = \frac{TP + TN}{N}, \tag{4.1}$$

The initial precision and recall are calculated based on TP, false negatives (FN), and false positives (FP) as shown in Eqs (4.2) and (4.3), respectively:

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP}, \tag{4.2}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN}. \tag{4.3}$$

The value of $F1$ is calculated based on precision and recall as shown in Eq (4.4):

$$F1 = \frac{2 \times \mathrm{Pr}\,ecision \times \mathrm{Re}call}{\mathrm{Pr}\,ecision + \mathrm{Re}call}. \tag{4.4}$$

Finally, the final results need to be validated. This paper adopts the leave-one-out cross validation method to prevent bias, using only one as the test set each time and all the rest as the training set. The formula is shown in Eq (4.5), discussing the impact of each improvement point on the experiments in this paper, respectively, and repeating to complete all the training. After that, the average of the experimental accuracy is calculated and, finally, we compare the method of this paper with the existing mainstream methods, including traditional methods and deep learning methods.

$$\frac{1}{N}\sum_{n=1}^{N} e_n = \frac{1}{N}\sum_{n=1}^{N} err\left(g_n^-(x_n), y_n\right), \tag{4.5}$$

where N represents the total number of samples. $n = 1$ denotes the sum of all samples from the 1st sample to the nth sample. e_n denotes the error term for the nth sample. $g_n^-(x_n)$ denotes the model prediction for the nth sample.

4.3. Experimental evaluation

4.3.1. Results of the experiment

Table 1 shows the results obtained from the experiments of the proposed method in this paper with various methods on the publicly available datasets SMIC, SAMM, and CASME II. These include LBP-based methods, optical flow-based methods, and deep learning-based methods. All the compared methods run the relevant experiments in the same environment.

Table 1. Comparison results of SMIC, SAMM, and CASME II datasets.

Method	SMIC		SAMM		CASME II	
	Accuracy%	F1	Accuracy%	F1	Accuracy%	F1
LBP-TOP	54.44	0.567	46.22	0.488	65.09	0.674
STCLQP	57.90	0.591	54.65	0.515	72.63	0.708
DisLBP-STIP	59.94	0.611	60.2	0.619	78.81	0.714
LGCcon	61.16	0.60	62.77	0.608	77.4	0.737
Bi-WOOF	62.56	0.564	61.88	0.583	78.46	0.709
C3D	61.28	0.610	73.71	0.692	69.71	0.711
OFF-ApexNet	71.79	0.712	59.66	0.566	87.69	0.822
DINet	68.66	0.644	75.29	0.739	78.43	0.763
CapsuleNet	65.06	0.643	68.3	0.644	70.18	0.706
RCN	70.65	0.711	70.2	0.691	80.39	0.826
SLSTT	72.40	0.707	76.5	0.683	90.12	0.885
Our	**74.6**	**0.751**	**84.77**	**0.793**	**91.35**	**0.889**

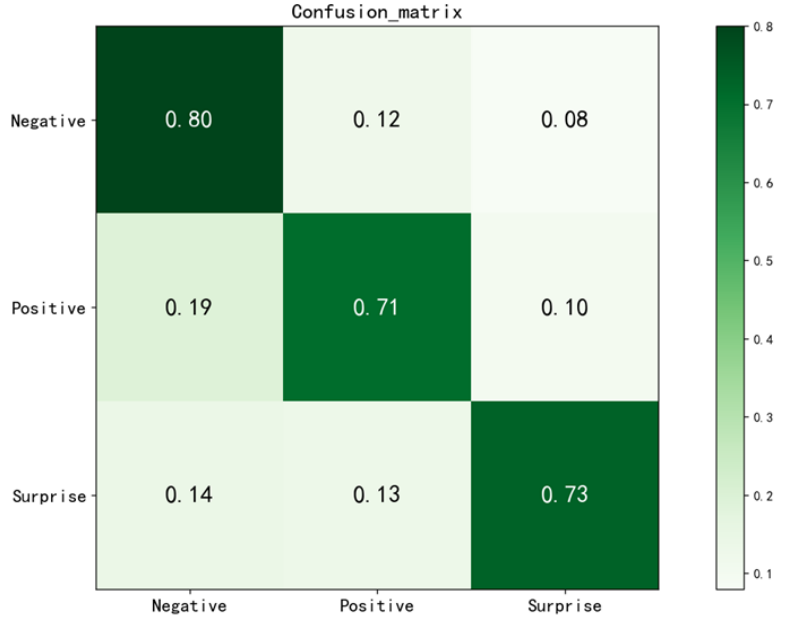
Note:Bold font is the best value for each column.

As can be seen from Table 1, the micro-expression recognition method based on the multi-scale 3D residual network architecture proposed in this paper shows significant competitive advantages on different benchmark datasets. Specifically, on the SMIC dataset, the proposed method improves the recognition accuracy by about 2.81% compared to the state of the art OFF-ApexNet model and also improves the accuracy by 5.94% compared to the DINet model. This strongly verifies the critical role of the spatial and temporal multi-scale fusion strategy that we adopt and the design of the 3D-ResNet backbone network in improving the micro-expression recognition performance. Comparing with traditional methods such as LBP-TOP and LBP-SIP, the limitations of their simple structure and under-utilization of spatio-temporal information lead to a bottleneck in improving recognition accuracy. In contrast, in the SMIC dataset, the method in this paper has a significant performance leap over LBP-TOP, with an accuracy improvement of about 18.16%. On the more complex and demanding SAMM dataset, the method in this paper shows excellent generalization ability and adaptability, with a recognition accuracy as high as 84.77%, which is significantly better than that of deep learning competitors such as Bi-WOOF, CapsuleNet, etc., and further highlights the advantages of the method in dealing with difficult micro-expression recognition scenarios. When applied to the better-quality CASME II dataset, the micro-expression recognition method proposed in this paper not only maintains its leading position, but also achieves an accuracy improvement of about 21.64% in comparison with deep learning benchmark algorithms such as convolutional 3D(C3D). This empirical result strongly proves that the method in this study has excellent accuracy and robustness in processing high-quality micro-expression data, thus providing a more scientific and effective solution to achieve the high-precision micro-expression recognition task.

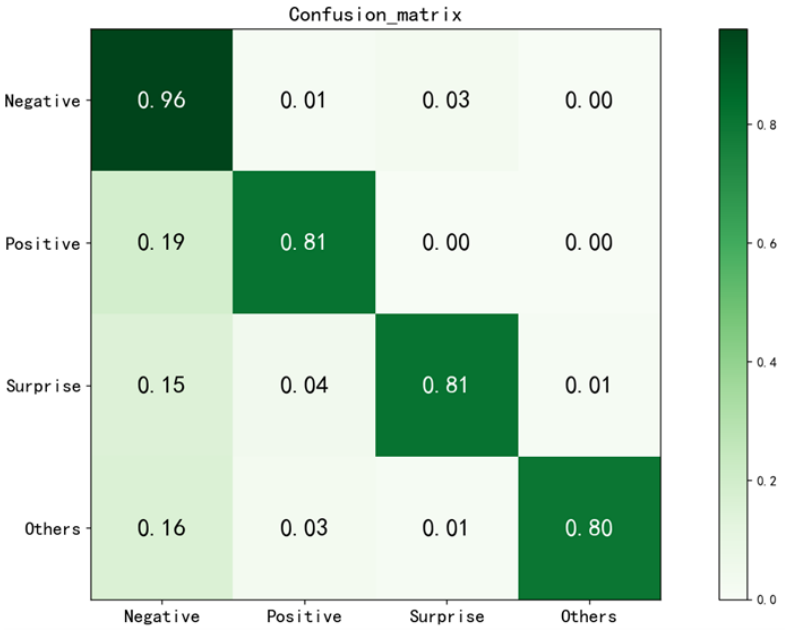
4.3.2. Analysis of the confusion matrix

The classification performance of the micro-expression recognition method proposed in this paper on three datasets, SMIC, SAMM, and CASME II, is visually demonstrated by the confusion matrix, as shown in Figure 10. From the confusion matrix of the SMIC dataset in Figure 10(a), it is observed that the algorithm has the highest recognition accuracy for the negative category, while the positive category has a relatively low recognition accuracy. This phenomenon can be attributed to the uneven distribution of the number of samples for each category in the dataset. In the SMIC dataset, the proportion of samples in the negative category reaches 42.7%, which is close to half of the total, so the recognition performance under this category is relatively good. Figure 10(b) demonstrates the performance of the algorithm on the SAMM dataset, which also presents the highest recognition accuracy for the negative category, which is closely related to the phenomenon that the sample size of the negative category occupies a high proportion in this dataset. In the SAMM dataset, the sample share of the negative category is further elevated, leading to a more unbalanced recognition result. Figure 10(c) exhibits the performance of the algorithm on the CASME II dataset, and we can see that the prediction accuracy of the OTHER category reaches 96%, which is the highest among all categories. This remarkable recognition effect also stems from the unbalanced nature of category distribution within the dataset. In the CASME II dataset, the number of samples in the OTHER category occupies 51.2% of the dataset. The model's ability to learn and recognize this category is optimized to the greatest extent possible during the training and testing process, thus achieving high prediction accuracy. The differences in classification performance exhibited by this paper's method on different datasets have a direct correlation with the distribution of the number of samples in each category within the dataset, especially in the

category with a larger proportion of samples. The micro-expression recognition method proposed in this study demonstrates better recognition efficacy.

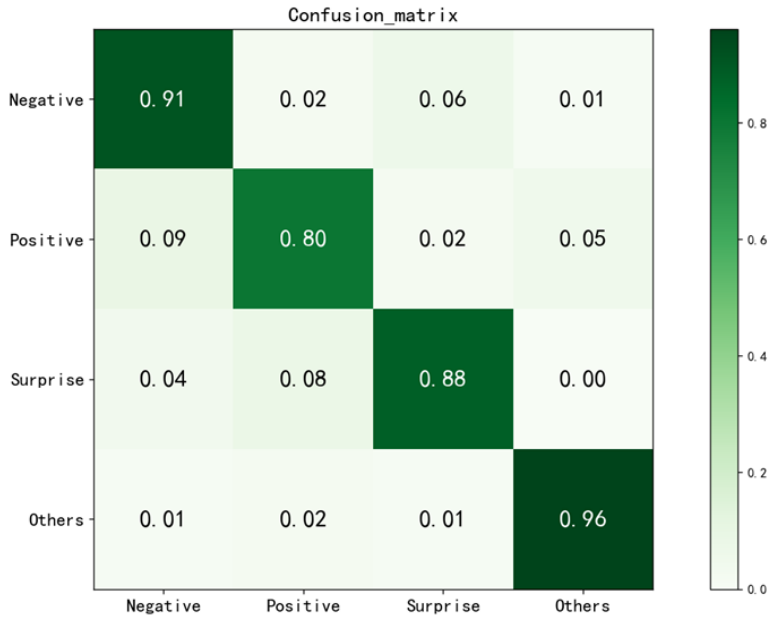


(a) Confusion matrix on the SMIC dataset



(b) Confusion matrix on the SAMM dataset (continued)

Continued on next page



(c) Confusion matrix on the CASME II dataset

Figure 10. Confusion matrix results on different datasets.

4.3.3. Analysis of ablation experiments

To demonstrate the effectiveness of the network model proposed in this paper in enhancing micro-expression classification, a series of ablation experiments were conducted. These experiments involved incorporating various enhancements into the basic 3D-ResNet network, including the optical flow method, attention mechanism, spatial multi-scale, and temporal multi-scale discriminative networks. The experiments were then performed on the SMIC, SAMM, and CASME II datasets.

By analyzing the experimental results in Figure 11, we can clearly see that the optical flow method and the temporal multi-scale module play a decisive role in improving the performance of micro-expression recognition. The optical flow method, as a key technology for dynamic visual information processing, has demonstrated excellent ability to capture extremely subtle facial motion changes. Whether on the SMIC, CASME II, or SAMM datasets, the motion features extracted using the optical flow method brought about a 4% or so accuracy improvement, and this remarkable effect fully verifies the advantages of the optical flow method in capturing and analyzing dynamic features, which is crucial for achieving high-precision micro-expression recognition. On the other hand, the introduced temporal multi-scale module also plays a key role in optimizing the model's performance. This module not only strengthens the model's ability to integrate contextual information from different time scales, but also takes full advantage of the deep neural network structure. Experimental results on three benchmark datasets show that the recognition accuracy is improved by 3.08% (SMIC), 3.7% (CASME II), and 1.86% (SAMM) with the addition of the temporal multi-scale module, which is strong evidence that the temporal multi-scale information greatly liberates the model's performance in providing multiple possible outputs and selecting the optimal solution. In addition, the attentional mechanism and the spatial multi-scale module likewise played an active role in this study. The attention mechanism enables the model to

focus more on the specific time nodes at which micro-expressions occur and their associated key facial regions (e.g., mouth, eyes, nose, etc.), thus improving the model's accuracy in the recognition task. The spatial multi-scale module, on the other hand, effectively strengthens the connection between local features and the overall scene by giving the model multiple sizes of receptive fields, further optimizing micro-expression recognition. Taken together, these modules collectively drive the performance of this paper's method in the micro-expression recognition task.

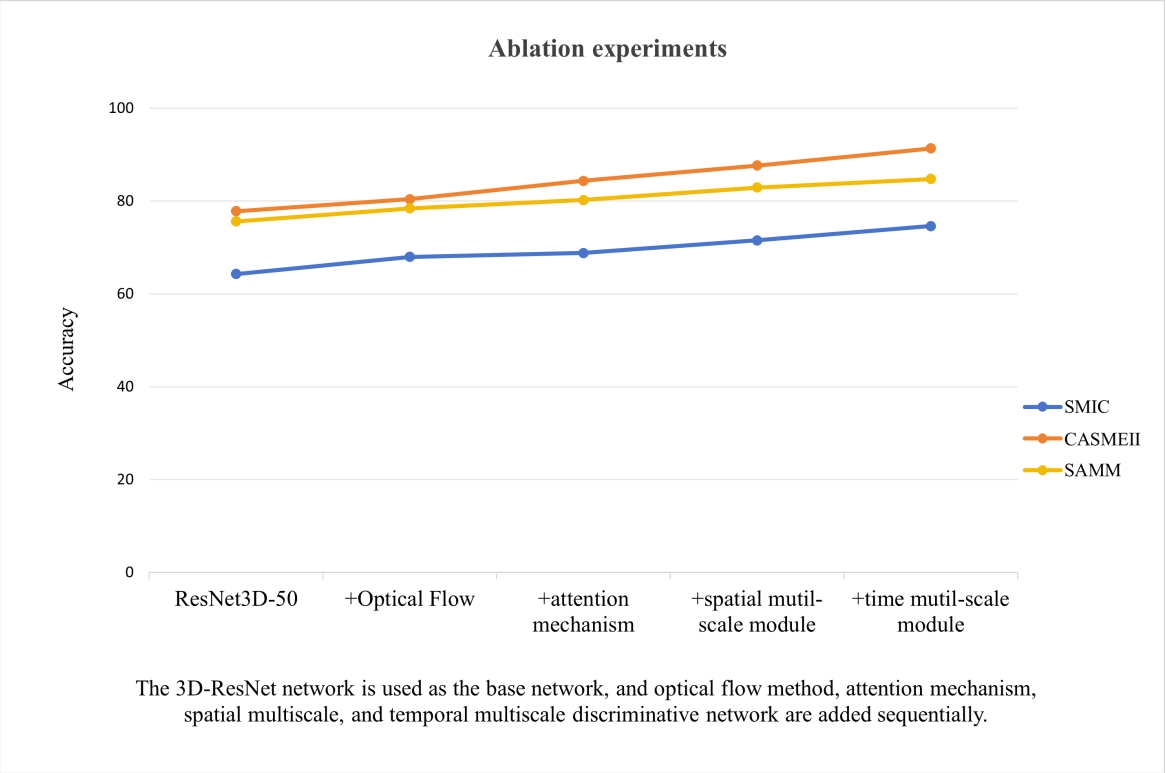


Figure 11. Experimental results of ablation experiment.

5. Conclusions

In this study, we propose a micro-expression recognition algorithm based on a multi-scale 3D residual CNN. Current mainstream micro-expression recognition algorithms such as MERANet, Transformers, and Dual-ATME have made substantial progress in micro-expression micro-feature detection through various optimization methods. Although the core improvement strategies of these algorithms generally tend to be how to more fully mine and extract features with significant distinguishing ability and have already improved the recognition rate to a certain extent, it is worth noting that an efficient recognition algorithm should not only be reflected in the powerful feature extraction function but also that the in-depth understanding and effective utilization of the acquired features are also crucial.

The algorithm presented herein builds upon existing research and adopts the deep 3D-ResNet50 as the foundational network model, with preprocessed micro-expression optical flow feature maps serving as input data. To bolster the model's capacity for expressive feature representation, we integrate a spatial

multi-scale convolution module along with an attention-driven feature module. This integration enables the system to glean valuable information from various hierarchical levels and dimensions. Furthermore, this research introduces an innovative multi-scale temporal feature fusion mechanism that seamlessly concatenates global contextual information with local fine-grained details, thereby enhancing the model's contextual awareness of micro-expressions. Ultimately, the algorithm employs a hierarchical output mechanism complemented by a discriminative network to pinpoint the most discriminant recognition results. Experimental evaluations show that our algorithm achieves remarkable recognition accuracies of 74.6, 84.77, and 91.35% on the SMIC, SAMM, and CASME II datasets, respectively, outperforming other mainstream approaches significantly. These outcomes validate the effectiveness of the proposed method in extracting subtle micro-expression features and improving recognition performance, thereby providing a rigorous academic reference for high-precision micro-expression recognition studies.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was partly supported by the Natural Science Basis Research Plan in Shaanxi Province of China under Grant 2023–JC–YB–517, and the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University under Grant VR-LAB2023B08.

Conflict of interest

All of the authors declare that there is no conflict of interest regarding the publication of this article and would like to thank the anonymous referees for their valuable comments and suggestions.

References

1. X. Shen, Q. Wu, X. Fu, Effects of the duration of expressions on the recognition of microexpressions, *J. Zhejiang Univ. Sci. B*, **13** (2012), 221–230. https://doi.org/10.1631/jzus.B1100063
2. C. Zhu, X. Chen, J. Zhang, Z. Liu, Z. Tang, Y. Xu, et al., Comparison of ecological micro-expression recognition in patients with depression and healthy individuals, *Front. Behav. Neurosci.*, **11** (2017), 199. https://doi.org/10.3389/fnbeh.2017.00199
3. X. Ben, Y. Ren, J. Zhang, S. J. Wang, K. Kpalma, W. Meng, et al., Video-based facial micro-expression analysis: A survey of datasets, features and algorithms, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 5826–5846. https://doi.org/10.1109/TPAMI.2021.3067464
4. T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, *IEEE Trans. Pattern Anal. Mach. Intell.*, **24** (2002), 971–987. https://doi.org/10.1109/TPAMI.2002.1017623
5. G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, *IEEE Trans. Pattern Anal. Mach. Intell.*, **29** (2007), 915–928. https://doi.org/10.1109/TPAMI.2007.1110

6. Y. Wang, J. See, R. C. W. Phan, Y. H. Oh, Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition, in *Computer Vision–ACCV 2014*, (2015), 525–537. https://doi.org/10.1007/978-3-319-16865-4_34

7. T. Pfister, X. Li, G. Zhao, M. Pietikäinen, Recognising spontaneous facial micro-expressions, in *2011 International Conference on Computer Vision*, (2011), 1449–1456. https://doi.org/10.1109/ICCV.2011.6126401

8. X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: Inducement, collection and baseline, in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, (2013), 1–6. https://doi.org/10.1109/FG.2013.6553717

9. Y. S. Gan, S. T. Liong, W. C. Yau, Y. C. Huang, L. k. Tan, OFF-ApexNet on micro-expression recognition system, *Signal Process.-Image Commun.*, **74** (2019), 129–139. https://doi.org/10.1016/j.image.2019.02.005

10. H. Q. Khor, J. See, R. C. W. Phan, W. Lin, Enriched long-term recurrent convolutional network for facial micro-expression recognition, in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, (2018), 667–674. https://doi.org/10.1109/FG.2018.00105

11. X. Ben, Y. Ren, J. Zhang, S. J. Wang, K. Kpalma, W. Meng, et al., Video-Based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 5826–5846. https://doi.org/10.1109/TPAMI.2021.3067464

12. J. Li, Z. Dong, S. Lu, S. J. Wang, W. J. Yan, Y. Ma, et al., CAS(ME)3: A Third Generation Facial Spontaneous Micro-Expression Database With Depth Information and High Ecological Validity, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2022), 2782–2800. https://doi.org/10.1109/TPAMI.2022.3174895

13. X. Li, S. Cheng, Y. Li, M. Behzad, J. Shen, S. Zafeiriou, et al., 4DME: A spontaneous 4D micro-expression dataset with multimodalities, *IEEE Trans. Affect. Comput.*, **14** (2022), 3031–3047. https://doi.org/10.1109/TAFFC.2022.3182342

14. S. Liu, Y. Ren, L. Li, X. Sun, Y. Song, C. C. Hung, Micro-expression recognition based on SqueezeNet and C3D, *Multimedia Syst.*, **28** (2022), 2227–2236. https://doi.org/10.1007/s00530-022-00949-z

15. S. Zhao, H. Tang, S. Liu, Y. Zhang, H. Wang, T. Xu, et al., ME-PLAN: A deep prototypical learning with local attention network for dynamic micro-expression recognition, *Neural Netw.*, **153** (2022), 427–443. https://doi.org/10.1016/j.neunet.2022.06.024

16. X. Huang, G. Zhao, X. Hong, W. Zheng, M. Pietikäinen, Spontaneous facial micro-expression analysis using Spatiotemporal Completed Local Quantized Patterns, *Neurocomputing*, **175** (2016), 564–578. https://doi.org/10.1016/j.neucom.2015.10.096

17. X. Huang, S. J. Wang, X. Liu, G. Zhao, X. Feng, M. Pietikäinen, Discriminative Spatiotemporal Local Binary Pattern with Revisited Integral Projection for Spontaneous Facial Micro-Expression Recognition, *IEEE Trans. Affect. Comput.*, **10** (2017), 32–47. https://doi.org/10.1109/TAFFC.2017.2713359

18. Y. Li, X. Huang, G. Zhao, Joint Local and Global Information Learning With Single Apex Frame Detection for Micro-Expression Recognition, *IEEE Trans. Image Process.*, **30** (2020), 249–263. https://doi.org/10.1109/TIP.2020.3035042

19. Y. J. Liu, B. J. Li, Y. K. Lai, Sparse mdmo: Learning a discriminative feature for spontaneous micro-expression recognition, *IEEE Transactions on Affective computing*, **12** (2021), 254–261. https://doi.org/10.1109/TAFFC.2018.2854166

20. S. T. Liong, J. See, K. S. Wong, R. C. W. Phan, Less is more: Micro-expression recognition from video using apex frame, *Signal Process. Image Commun.*, **62** (2018), 82–92. https://doi.org/10.1016/j.image.2017.11.006

21. R. Ni, B. Yang, X. Zhou, S. Song, X. Liu, Diverse local facial behaviors learning from enhanced expression flow for microexpression recognition, *Knowl.-Based Syst.*, **275** (2023), 110729. https://doi.org/10.1016/j.knosys.2023.110729

22. X. Li, J. Yu, S. Zhan, Spontaneous facial micro-expression detection based on deep learning, in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, (2016), 1130–1134. https://doi.org/10.1109/ICSP.2016.7878004

23. R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 1932–1939. https://doi.org/10.1109/CVPR.2009.5206821

24. L. Zhou, Q. Mao, X. Huang, F. Zhang, Z. Zhang, Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition, *Pattern Recogn.*, **122** (2022), 108275. https://doi.org/10.1016/j.patcog.2021.108275

25. S. T. Liong, Y. S. Gan, J. See, H. Q. Khor, Y. C. Huang, Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition, in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, (2019), 1–5. https://doi.org/10.1109/FG.2019.8756567

26. J. Li, S. Zhang, T. Huang, Shallow Triple Stream Three-dimensional CNN (STSTNet) for Micro-expression Recognition, *Computer Vision and Pattern Recognition*, (2018), 8618—8625. https://doi.org/10.1109/FG.2019.8756567

27. J. Li, T. Wang, S. J. Wang, Facial micro-expression recognition based on deep local-holistic network, *Appl. Sci.*, **12** (2022), 4643. https://doi.org/10.3390/app12094643

28. N. Van Quang, J. Chun, T. Tokuyama, CapsuleNet for micro-expression recognition, in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, (2019), 1–7. https://doi.org/10.1109/FG.2019.8756544

29. Z. Xia, X. Hong, X. Gao, X. Feng, G. Zhao, Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions, *IEEE Trans. Multimedia*, **22** (2019), 626–640. https://doi.org/10.1109/TMM.2019.2931351

30. L. Zhang, X. Hong, O. Arandjelović, G. Zhao, Short and long range relation based spatio-temporal transformer for micro-expression recognition, *IEEE Trans. Affect. Comput.*, **13** (2022), 1973–1985. https://doi.org/10.1109/TAFFC.2022.3213509

31. Y. Su, J. Zhang, J. Liu, G. Zhai, Key facial components guided micro-expression recognition based on first & second-order motion, in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, (2021), 1–6. https://doi.org/10.1109/ICME51207.2021.9428407

32. V. R. Gajjala, S. P. T. Reddy, S. Mukherjee, S. R. Dubey, MERANet: Facial micro-expression recognition using 3D residual attention network, in *Proceedings of the twelfth Indian conference on computer vision, graphics and image processing*, (2021), 1–10. https://doi.org/10.1145/3490035.3490260

33. Z. Zhang, Z. Hu, W. Deng, C. Fan, T. Lv, Y. Ding, DINet: Deformation inpainting network for realistic face visually dubbing on high resolution video, preprint, arXiv:2303.03988.

34. H. Zhou, S. Huang, J. Li, S. J. Wang, Dual-ATME: Dual-branch attention network for micro-expression recognition, *Entropy*, **25** (2023), 460. https://doi.org/10.3390/e25030460

35. V. Kazemi, J. Sullivan, One Millisecond Face Alignment with an Ensemble of Regression Trees, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 1867–1874. https://doi.org/10.1109/CVPR.2014.241

AIMS Press