TYPE Original Research
PUBLISHED 27 July 2023
DOI 10.3389/fnins.2023.1216181

Check for updates

OPEN ACCESS

EDITED BY
Cunhang Fan,
Anhui University, China

REVIEWED BY
Jingting Li,
Chinese Academy of Sciences (CAS), China
Cheng Lu,
Southeast University, China

*CORRESPONDENCE
Hang Pan
✉ panhang@czc.edu.cn

RECEIVED 03 May 2023
ACCEPTED 26 June 2023
PUBLISHED 27 July 2023

CITATION
Pan H, Yang H, Xie L and Wang Z (2023) Multi-scale fusion visual attention network for facial micro-expression recognition.
*Front. Neurosci.* 17:1216181.
doi: 10.3389/fnins.2023.1216181

# Multi-scale fusion visual attention network for facial micro-expression recognition

Hang Pan[1]*, Hongling Yang[1], Lun Xie[2] and Zhiliang Wang[2]

[1]Department of Computer Science, Changzhi University, Changzhi, China, [2]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China

**Introduction:** Micro-expressions are facial muscle movements that hide genuine emotions. In response to the challenge of micro-expression low-intensity, recent studies have attempted to locate localized areas of facial muscle movement. However, this ignores the feature redundancy caused by the inaccurate locating of the regions of interest.

**Methods:** This paper proposes a novel multi-scale fusion visual attention network (MFVAN), which learns multi-scale local attention weights to mask regions of redundancy features. Specifically, this model extracts the multi-scale features of the apex frame in the micro-expression video clips by convolutional neural networks. The attention mechanism focuses on the weights of local region features in the multi-scale feature maps. Then, we mask operate redundancy regions in multi-scale features and fuse local features with high attention weights for micro-expression recognition. The self-supervision and transfer learning reduce the influence of individual identity attributes and increase the robustness of multi-scale feature maps. Finally, the multi-scale classification loss, mask loss, and removing individual identity attributes loss joint to optimize the model.

**Results:** The proposed MFVAN method is evaluated on SMIC, CASME II, SAMM, and 3DB-Combined datasets that achieve state-of-the-art performance. The experimental results show that focusing on local at the multi-scale contributes to micro-expression recognition.

**Discussion:** This paper proposed MFVAN model is the first to combine image generation with visual attention mechanisms to solve the combination challenge problem of individual identity attribute interference and low-intensity facial muscle movements. Meanwhile, the MFVAN model reveal the impact of individual attributes on the localization of local ROIs. The experimental results show that a multi-scale fusion visual attention network contributes to micro-expression recognition.

KEYWORDS
**micro-expression recognition, attention mechanism, mask operate, multi-scale feature, feature fusion**

## 1. Introduction

Human-computer interaction not only requires machines to complete specified tasks, but also requires machines to have emotional cognition, communication, and feedback capabilities like humans during the interaction process (Ahmad et al., 2019). Human emotions can be expressed through speech (Zhao et al., 2021), text (Wang et al., 2020b), gestures (Li Y. -K. et al., 2022), and physiological signals (Wu et al., 2023), but facial expressions most intuitively reflect human emotions. The research has found that people would intentionally display certain facial expressions in certain situations, yet when people try to hide their facial expressions in high-stakes situations, it is necessary to interpret facial micro-expressions to determine their true emotional state (Ekman and Friesen, 1969).

Facial micro-expressions (hereinafter referred to as micro-expressions) are short-duration and low-intensity facial muscle movements. Since it is usually caused by suppressed emotions and can result from genuine motivations and emotions (Ekman, 2009). If people are not professionally trained, it is impossible to hide the appearance of micro-expressions (Holler and Levinson, 2019). Researchers found that micro-expressions are often present in lie detection scenarios. Thus, it has major implications when it comes to high-risk situations including criminal investigation, social interactions, national security, and business negotiations (O'Sullivan et al., 2009). It is less accurate to recognize micro-expressions (Ben et al., 2021; Tran et al., 2021; Zhou et al., 2021; Bisogni et al., 2022; Wang et al., 2022; Zhu et al., 2022; Wei et al., 2022a) than facial expressions (Shao and Qian, 2019; Li and Deng, 2020; Chowdary et al., 2021; Bisogni et al., 2022). The current research on micro-expression recognition is still correlating datasets collected the laboratory environments, which is not enough for application in high-risk scenes.

With the rapid development of image acquisition equipment, researchers use high-speed cameras to collect micro-expression images to create the Spontaneous Micro-Expression Database (SMIC) (Li et al., 2013), Chinese Academy of Sciences Micro-expression database (CASME II) (Yan et al., 2014), Spontaneous Actions, and Micro-Movements (SAMM) (Davison et al., 2016), Micro-and-Macro Expression Warehouse (MMEW) (Ben et al., 2021), and Third Generation Facial Spontaneous Micro-Expression Database (CAS (ME)$^3$) (Li J. et al., 2022). These datasets are collected by high-speed cameras to alleviate the problem of short duration. However, facial muscle movements low-intensity are still an important factor inhibiting the enhancement accuracy of micro-expression recognition. For the low-intensity challenge of micro-expressions, researchers extract efficient local features for micro-expression recognition through regions-of-interest (ROIs) localization prior knowledge (Xu et al., 2017; Niu et al., 2019; Yu et al., 2019; Merghani and Yap, 2020) or local ROIs localization based on deep learning (DL) (Bai, 2020; Chen et al., 2020; Xia et al., 2020; Xie et al., 2020; Wang et al., 2020a; Li et al., 2021; Zhao et al., 2022). Although the local features extraction after local ROIs locating through prior knowledge or deep learning methods is helpful for micro-expression recognition, but these methods ignore the feature redundancy caused by the inaccuracy of ROIs. Moreover, psychological research has shown that muscle movement changes during facial expression did not correlate with individual identity attributes such as gender, age, and ethnicity (Ekman and Friesen, 1971). However, these micro-expression recognition methods do not consider the effect of individual identity attributes on the localization of ROIs. To overcome the challenge of low-intensity facial muscle movements in the micro-expression recognition task, this paper proposes a novel multi-scale fusion visual attention network (MFVAN). This model explores the effect of reducing individual identity attributes on emotional change ROIs localization and learns multi-scale local attention weights to mask regions of redundancy features. The framework of the MFVAN model is shown in Figure 1.

The micro-expression image (apex frame) is input into the convolutional neural network (CNN) that is mapped to a calm state image (onset frame) to reduce the effect of individual identity attributes and obtain multi-scale feature maps. The multi-head self-attention (MSA) extracts the local features of the multi-scale feature maps and obtains the attention weights of these features. We reduce feature redundancy by dropping out local features irrelevant to micro-expressions according to attention weights. At the same time, local features with higher attention weights in multi-scale are fused to improve the robustness features. Finally, the multi-scale classification loss, mask loss, and removing individual identity attributes loss joint to optimize the MFVAN model. The experimental results on SMIC, CASME II, SAMM, and their combined datasets (See et al., 2019) demonstrate that the MFVAN can achieve state-of-the-art performance by fusing multi-scale local attention features. Overall, our proposed MFVAN model is the first to combine image generation with visual attention mechanisms to solve the combination challenge problem of individual identity attribute interference and low-intensity facial muscle movements. In summary, the main contributions of this paper can be summarized as follows:

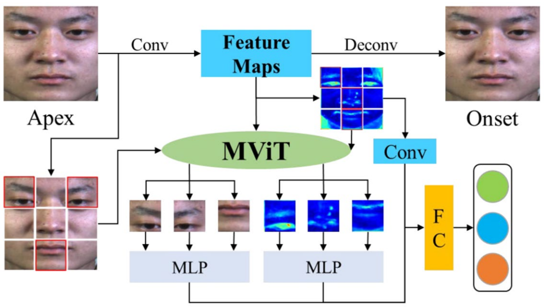

FIGURE 1
The framework of MFVAN model to explores the effect of reducing individual identity attributes on emotional change ROIs localization and learns multi-scale local attention weights to mask regions of redundancy features.

1. This paper analyzed the combined effects of facial identity attributes on micro-expression recognition. This paper analyzes the combined effects of low-intensity of facial muscle movement changes and individual identity attributes in micro-expression recognition.
2. This paper is the first study that combines image generation with visual attention mechanisms and proposes an MFVAN framework. The self-supervised and transfer learning is jointly trained to remove individual identity attributes.
3. Meanwhile, the MFVAN model utilizes the global and multi-scale local attention weights connected for micro-expression recognition. The focal loss, removing identity attributes loss, and the marked loss are used to optimize the MFVAN model. The experimental results on SMIC, CASME II, SAMM, and their combined datasets demonstrate that the MFVAN can achieve state-of-the-art performance that focuses on local at the multi-scale and contributes to micro-expression recognition.

The rest of this paper is structured as follows: In Section 2, we review the micro-expression recognition datasets, handcrafted features, and deep learning micro-expression recognition methods. Section 3 presents the proposed MFVAN framework. Section 4 presents the experimental dataset, evaluation metrics, quantitative analysis, and analysis of ablation experiments. Finally, Section 5 summarizes the proposed algorithm and discusses future research trends.

## 2. Related research

This section introduces the facial micro-expression recognition dataset used in the experiments part. Then, by comparing the research status of micro-expression recognition based on handcrafted features and deep learning, the shortcomings of existing research are obtained that laying the groundwork for the research method.

### 2.1. Datasets description

The premise of micro-expression recognition must have sufficient data with emotional labels. However, the research on facial micro-expression recognition through computer vision has just started. At present, there are very few micro-expression datasets, mainly including imitation and spontaneous datasets. The most important difference between the two is the correlation of facial micro-expression manifestations with underlying emotional states. Among them, spontaneous micro-expressions are facial movements shown through external stimuli, which are consistent with underlying emotional states. Therefore, this paper adopts three spontaneous facial micro-expression datasets and their combined datasets to verify the proposed MFVAN method.

The SMIC is the first public dataset used for micro-expression recognition. This dataset includes 328 videos collected from 20 subjects. During the experiment, a high-speed camera with a resolution of 640 × 480 pixels and a transmission rate of 100 frames per second (FPS) was used to collect the facial images of each subject throughout the process. The researchers screened 164 video clips (negative, positive, surprise) inspired by all 16 subjects participating in the experiment to form SMIC for facial micro-expression recognition. The CASME II was recorded with a camera with 640 × 480 pixels and 200 FPS from 26 subjects. The researchers screened 246 video clips (happiness, surprise, disgust, repression, others) that were selected from more than 3000 facial actions. The SAMM is a high-speed and high-resolution dataset that uses a camera with a frame rate of 200 and a resolution of 2040 × 1088 to capture images. The database collected video data from 32 subjects with an average age of 33.24 years by spontaneous elicitation, with an even and rich ethnic distribution. The 159 samples were labeled including happiness, surprise, contempt, anger, others, disgust, fear, and sadness. The 3DB-combined dataset is a reclassification and combination of CASME II and SAMM based on SMIC dataset labels. Among them, depression, sadness, contempt, and disgust are divided into negative categories in the SMIC dataset, and happiness is divided into positive categories. The recombined 3DB-combined dataset contains 442 samples among which 109 positive, 250 negative, and 83 surprised, including all 164 samples in SMIC, 145 samples in CASME II, and 133 samples in SAMM.

### 2.2. Handcrafted features methods

Facial micro-expression recognition methods are generally divided into two categories, one is to extract the manual change features of facial images of micro-expression video sequences for micro-expression recognition, and the other is to first use deep learning methods for micro-expression recognition. In the previous facial micro-expression recognition, to describe the changes of micro-expressions, many research works have used manual feature-based methods to extract the changes in texture, color, and optical flow characteristics of image sequences, splicing them into a compact feature vector and outputting it to the classifier identifies micro-expressions.

The Local Binary Pattern from Three Orthogonal Planes (LBP-TOP) (Pfister et al., 2011) is the representative handcrafted feature extraction method applied to micro-expression recognition. The follow-up research work is the improvement of LBP-TOP. The Local Binary Pattern Six Interception Points (LBP-SIP) (Wang et al., 2014) and Local Binary Pattern from Mean Orthogonal Planes (LBP-MOP) (Wang et al., 2015) are used to reduce the redundancy problem. The Kernelized Two-Groups Sparse Learning (KTGSL) (Wei et al., 2022b) automatically learns more discriminative features from Local Binary Pattern with Single Direction Gradient (LBP-SDG) (Wei et al., 2021) and Local Binary Pattern from Five Intersecting Planes (LBP-FIP) (Wei et al., 2022a) two sets of features to improve micro-expression recognition performance. The Discriminative Spatiotemporal Local Radon Binary Pattern Based on Revisited Integral Projection (DiSTLBP-RIP) (Huang et al., 2019) fuses shape features into LBP-TOP to improve the ability to discriminate micro-expressions.

In addition to texture and shape features, optical flow features are also manual features commonly used in micro-expression recognition (Li et al., 2020). The Fuzzy Histogram of Optical Flow Orientation (FHOFO) (Happy and Routray, 2017) is employed the Facial Action Coding System (FACS) to locate 36 facial ROIs to extract the subtle changes in these regions for micro-expression recognition. The Weighted Oriented Optical Flow (BI-WOOF) (Liong et al., 2018) weighted average of the overall and local histogram of oriented optical flow features. The Sparse Main Directional Mean Optical Flow (SMDMO) (Liu et al., 2018) averages the optical flow features of the region of interest of 36 motion units in the face area to reduce the noise effect caused by head movement in micro-expression recognition. Although the method based on handcrafted features can achieve good performance in micro-expression recognition, it requires a lot of preprocessing such as face detection alignment and video frame insertion in the early stage. With the application of end-to-end deep learning methods in the field of image recognition, more and more researchers have begun to consider how to use deep learning methods to solve micro-expression recognition tasks.

### 2.3. Deep learning methods

With the development of DL, especially the proposal of the CNN model which includes AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015), ResNet (He et al., 2016), and SENet (Hu et al., 2018) for image recognition. With the application of DL in face detection, face recognition, face editing, and expression recognition, more researchers have begun to pay attention to micro-expression recognition through DL methods. Kim et al. (2016) is the first attempt to use the combination of CNN and Long Short-Term Memory (LSTM) to extract spatiotemporal features of micro-expression video sequences for MER.

Gan et al. feed apex optical flow (OFF-Apex) (Gan et al., 2019) images into CNN for micro-expression recognition. The Dual-Stream Shallow Network (DSSN) (Khor et al., 2019)model reduces the complexity of the model by pruning the CNN model while improving the recognition performance of micro-expressions. The Stacked Hybrid Convolution Feature Network (SHCFNet) (Huang et al., 2020) enhances the CNN network with optical flow features of different scales. The GEender-based Micro-Expression (GEME) (Nie et al., 2021) reduces the impact of gender on CNN models through a two-stream multi-task framework. The Local and Global information joint learning module (LGCcon) (Li et al., 2021) localizes the main emotional information local area while suppressing the negative impact of irrelevant local areas on micro-expression recognition. The Action Unit – Graph Convolution Network (AU-GCN) (Lei et al., 2021) enhances the feature representation of nodes and graph edges extracted by the graph convolutional network by fusing AU coding features. The Two-Stream Graph Attention Convolutional Network (TSGACN) (Kumar and Bhanu, 2021) encodes features and optical flow features by fusing facial key points. The (Feature Refinement, FeatRef) (Zhou et al., 2022) framework uses the attention model to select the obvious discriminant features for micro-expression recognition. The Prototypical Learning with Local Attention Network (PLAN) (Zhao et al., 2022) learns local facial action change features through local attention modules.

These deep learning techniques can improve recognition performance by learning more efficient depth features than hand-crafted features from video sequences or the apex frame of micro-expression samples. However, these works rarely consider the impact of local information. Although the AU-GCN, TSGACN, and PLAN methods use AU information to assist micro-expression recognition, they do not consider the influence of individual identity attributes on local information.

## 3. Methods

In this paper, our proposed network extracts global features by removing individual identity attributes. Then multi-scale attention mechanism is used to capture local information of the global image and different convolutional layers. Finally, the multi-scale local features with high weights are fused to classify the apex frame. The architecture of MFVAN is illustrated in Figure 2, our proposed method framework MFVAN is a multi-scale joint network.

### 3.1. Removing identity attributes

Psychologists believe that facial micro-expressions are not related to individual identity attributes. Meanwhile, Valstar and Pantic (2011) consider that the onset frame in the micro-expression video sequence represents the moment when the appearance of the face is enhanced. So, the onset frame can be considered as an identity image, which contains the identity attributes of the individual. The apex frame is the coupling of individual identity attributes and emotional representation. In the process of decoupling identity information, we map the apex frame to the onset frame through the autoencoder model. The encoder module is used to map apex frames to identity features. Then the identity features are used to generate an identity image through the decoder module. Therefore, we map the apex frame (facial micro-expression images) into CNN to the onset frame (neutral states image) to remove the identity information and obtain the global change features.

The problem of a small dataset of micro-expressions severely constrains the training of the mapping model for removing identity attributes. To better learn the CNN mapping model, we use self-supervised and transfer learning to train the mapping model. We train the teacher model by a deep image self-supervision approach. The image is mapped to a high-dimensional feature space using multiple residual modules at the image encoder and then returned to the original image by deconvolution. The teacher networks are complex with superior performance. Then a shallow network student network

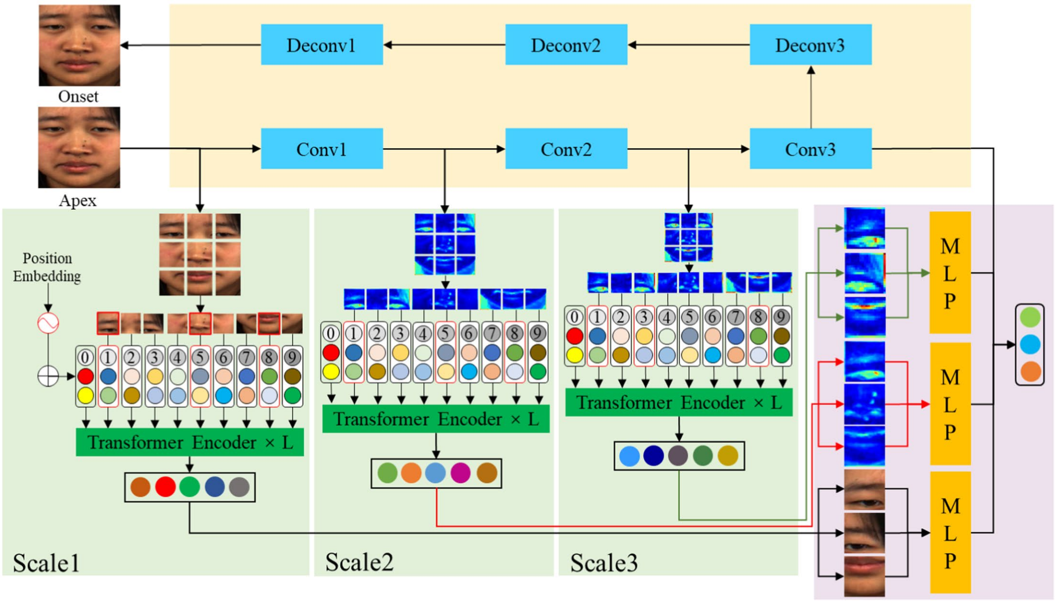

FIGURE 2
The MFVAN model contains three components. The first is a de-identity attribute model based on Convolutional Neural Network (CNN) to extract global facial features. Then multi-scale vision transformer (MViT) model is used to capture local information about the apex frame and different convolutional layers where the red boxes represent regions with higher weights. Finally, multi-scale local features with high attention weight are fused for micro-expression recognition.

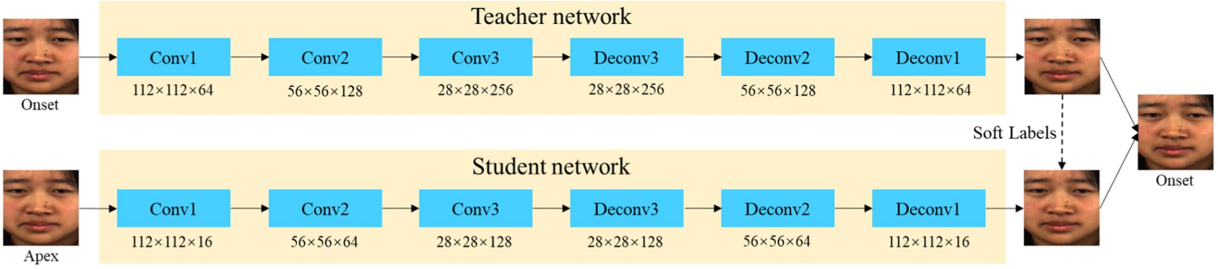

FIGURE 3
Illustration of the transfer learning framework for removing identity attributes.

is designed to learn the mapping relationship of the apex to the onset frame. This teacher network is used as a soft target to guide shallow student networks so that a simpler student model with fewer parameters can have a similar performance as the teacher network. The network structure is shown in Figure 3.

In the training process, we first pre-trained the teacher model. Image self-supervised training is performed by feeding all images from the micro-expression video samples into the teacher model. Then the teacher and student network are jointly trained. The apex frame input generates a network to output the onset frame. The model parameters are optimized by the mean squared error (MSE) loss function of the two models. The loss functions for the pre-training model and the removing identity attribute model are $L_{pre}$ and $L_{remove}$ , which are computed as follows:

$$L_{pre} = \frac{1}{N*M}\sum_{i=1}^{N}\sum_{j=1}^{M}\left|x_{i,j} - x'_{i,j}\right|^2 \tag{1}$$

$$L_{Teacher} = \frac{1}{N*M}\sum_{i=1}^{N}\sum_{j=1}^{M}\left|x_{onseti,j} - x'_{onseti,j}\right|^2 \tag{2}$$

$$L_{student} = \frac{1}{N*M}\sum_{i=1}^{N}\sum_{j=1}^{M}\left|x_{onseti,j} - x''_{onseti,j}\right|^2 \tag{3}$$

$$L_{remove} = L_{Teacher} + L_{student} \tag{4}$$

where $x$ is the image of micro-expression video samples. $x'$ is the image generated by the pre-trained model. $x_{onset}$ is the onset frame. $x'_{onset}$ and $x''_{onset}$ is the image generated by the teacher and student network.

## 3.2. Multi-scale fusion visual attention network

The low-intensity characteristic of micro-expressions represent as muscle movement changes in localized regions of facial images. However, inaccurate localization of local regions can lead to feature redundancy and thus affect recognition performance. In this paper, we propose an MFVAN model for improving micro-expression recognition performance by extracting the local features of the multi-scale feature map. The feature weights of the apex frame and the patch token of the feature map at multiple scales are learned by MSA in the visual transformer model (Dosovitskiy et al., 2021). Then the patch token with a high weight at each scale is input into multi-layer perceptron (MLP) fusion to recognize micro-expressions. The MFVAN structure is shown in Figure 4.

The MFVAN model flattens the apex frame and multi-scale feature maps are split into $s \times s$ patches and flattened to generate image sequences $x_i$. These image patches sequence is mapping to a feature vector $f_i$ with convolution operating and weighting the positional embedding $f_{pos}$ to generate a new feature vector. The dimension of $f_i$ is $k \times q$. The parameter $k$ is $s \times s$ which is the image patches token length. The dimension $q$ is determined by the convolution mapping performed on each block to generate the feature dimension. For each scale feature vector, we add a class token. The calculation process of the new feature vector $f_i'$ of the $i-th$ scale is shown in Eq. (5).

$$f_i' = \left[f_{c,i}, f_i + f_{pos}\right], i \in (1,2,3) \tag{5}$$

$$f_i = F_{conv}\left(\sum x_i w_i + b_i\right) \tag{6}$$

where $f_{c,i}$ is the class token of the $i-th$ scale, $f_i$ is the feature by convolution mapped, $f_{pos}$ is the positional embedding, $i$ is the scale, $w_i$ and $b_i$ are the weights of the convolution mapped of the $i-th$ scale.

In the transformer encoder model, the class token is a learnable classification parameter. But in the MFVAN model, the class token is not only used for classification, but also used to learn the attention weight of each patch, and perform mask operation on the patch with low weight. The transformer encoder module in the MFVAN model contains L layers of MSA and MLP blocks. The feature vectors corresponding to patches with attention weights greater than $\theta$ are fused by the MLP module for classification. The classification process is shown in Eq. (7).

$$p = SoftMax\left(F_{MLP-f}\left(\frac{1}{3}\sum_{i=1}^{3} f'_{i,L}\right)\right), f_c \geq \theta \tag{7}$$

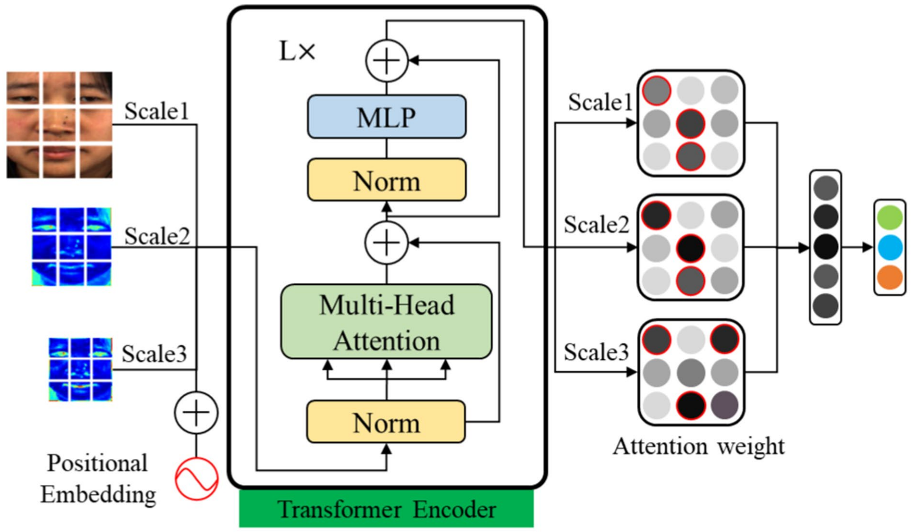

FIGURE 4
The architecture of MViT model. The apex frame and multi-scale feature are input transformer encoders. Then the local feature with high weights on each scale is fused.

$$f'_{i,l} = F_{MLP,l}\left(F_{MSA,l}\left(f'_{i,l-1}\right)\right), l \in 1,2,\ldots,L \qquad (8)$$

where $p$ the prediction result of the MFVAN model, $F_{MLP-f}$ is the fused function, $f'_{i,0}$ is the input embedding vector, $f'_{i,L}$ is the output of the transformer encoder module of the $i-th$ scale, $f_c$ is the attention weight, $\theta$ is the threshold for dividing the attention weights, $L$ is the number of MSA and MLP blocks, $F_{MSA,l}$ and $F_{MLP,l}$ are the $l$-$th$ layer block.

## 3.3. Loss function optimization based on global and local

In this paper, we use removing identity attributes loss, global classification loss, multi-scales classification loss, and multi-scales mask loss function joint to optimize the MFVAN model and learn the local patch attention weight when the micro-expression occurs.

$$L_{all} = L_{remove} + L_{class\_global}(p,y) + \sum_{i=1}^{3} L_{class\_scale,i}\left(p_{i,scale}, y\right) + L_{mask} \qquad (9)$$

$$L_{class} = -\alpha_t \left(1-p_t\right)^{r} \log\left(p_t\right) \qquad (10)$$

$$L_{mask} = \begin{cases} 0, & if\ f_c < \theta \\ f_c, & otherwise \end{cases} \qquad (11)$$

where $L_{remove}$ is the removing identity attributes loss, $L_{class\_global}$ is the global level of classification loss, $L_{class\_scale,i}$ is the classification loss of the $i-th$ scale, $p$ is the probability of the global prediction, $p_{i,scale}$, is the probability of $i-th$ scale, $y$ is the ground truth, $L_{class}$ is the classification loss, $L_{mask}$ is mask loss, $\alpha_t$, $r$ are hyperparameters. $\alpha_t$ represents the weight of the t-th class sample, and $p_t$ represents the probability value of the $t$-th class output by Softmax.

# 4. Experimental analysis

In this section, the evaluation metrics, comparative analysis of experimental results, ablation experiments, and visualization analysis will be introduced in detail. The proposed MFVAN method is evaluated on SMIC, CASME II, SAMM, and 3DB-Combined datasets.

## 4.1. Evaluation metric

The evaluation metric for micro-expression recognition is the accuracy and F1-score on the single dataset by the Leave-One-Subject-Out (LOSO) cross-validation. The Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) on the combined datasets. The evaluation metric is computed using:

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (12)$$

$$Recall = \frac{TP}{TP + FN} \qquad (13)$$

$$Precision = \frac{TP}{TP + FP} \qquad (14)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

$$UF1 = \frac{1}{C}\sum_{c=1}^{C} F1_c \tag{16}$$

$$UAR = \frac{1}{C}\sum Acc_c \tag{17}$$

where TP is true positive, TN is true negative, FP is the false positive, and FN is false negative, $C$ is the total number of categories.

## 4.2. Result analysis of a single datasets

In this part, we evaluate the effectiveness of the MFVAN model by comparing it with two types of baseline methods based on handcrafted features and deep learning on the single dataset. In the comparison experiment of handcrafted feature methods, this paper selects the most representative LBP-TOP, BI-WOOF, DiSTLBP-RIP, LBP-SDG, and KTGSL with our proposed MFVAN model contrasted. In the deep learning method comparison experiment, we chose the OFF-Apex, DSSN, LGCcon, GEME, AU-GCN, and FeatRef models. The Bi-WOOF method based on handcrafted features and the AU-GCN model based on deep learning adopt the method of ROIs positioning. The experimental results are shown in Table 1.

The experimental comparison in the SMIC dataset found that the accuracy of MFVAN was 4.23 and 12.19% higher than the best KTGSL in handcrafted features and the best OFF-Apex in deep learning. The performance of the F1-Score is 0.1109 and 0.13 higher, respectively. The MFVAN model also achieves state-of-the-art performance on two other single CASME II and SAMM datasets. In all comparative experimental analyses, almost all methods input video sequences of micro-expression samples, only OFF-Apex, DSSN, GEME, and AU-GCN methods use apex frame (or apex frame and onset frame) for micro-expression recognition.

The OFF-Apex model is one of the representative methods that only use peak frame information for deep learning training in the early stage of micro-expression recognition. Most of the subsequent methods are based on it to improve and improve the model or method. For example, DSSN compresses the model by pruning, and GEME eliminates the influence of individual gender. Although GEME has considered the interference of gender, their limitation is that it only considers the interference of gender, and the individual identity attribute has the influence of other attributes such as skin color and age in addition to gender. Therefore, the MFVAN model self-supervision and transfer learning reduce the influence of individual identity attributes and increase the robustness of multi-scale feature maps to improve the performance of micro-expression recognition.

## 4.3. Result analysis of a combined dataset

This section also further verifies the effectiveness of the MFVAN model on the 3DB-Combined dataset of the MEGC 2019. Since DiSTLBP-RIP, LBP-SDG, KTGSL, and DSSN do not report experimental results on combined datasets, we conduct comparative experiments with the remaining methods. It is worth noting that since the SMIC dataset does not provide the marker of the peak frame, in the comparison experiment of the composite dataset, the LGCcon model only reports the experimental results of the adjusted dataset. The experimental results are shown in Table 2.

The first 6 columns are the experimental results of the adjusted three-category dataset. Similar to the original dataset, the MFVAN model can achieve competitive results, and the UF1 and UAR indicators are 0.0794/0.0684, 0.0263/0.0186, and 0.0186/0.0263 higher than the optimal AU-GCN model on the two datasets. The MFVAN can achieve state-of-the-art performance in the combined dataset.

## 4.4. Ablation experiment analysis

To evaluate the effectiveness of the MFVAN model, we conducted ablation experiments analysis comparison of attention models at different scales on SMIC, CASME II, and SAMM. The detailed

**TABLE 1** Micro-expression recognition performance comparison on the SMIC (3 categories), CASME II (5 categories), and SAMM (5 categories).

| Methods | SMIC (3) | | CASME II (5) | | SAMM (5) | |
|---|---|---|---|---|---|---|
| | Accuracy | F1-Score | Accuracy | F1-Score | Accuracy | F1-Score |
| LBP-TOP (2011) | 48.78 | 0.4600 | 39.68 | 0.3589 | 35.56 | 0.3589 |
| Bi-WOOF (2018) | 61.59 | 0.6110 | 57.89 | 0.6125 | – | – |
| DiSTLBP-RIP (2019) | 63.41 | – | 64.78 | – | – | – |
| LBP-SDG (2021) | 69.68 | 0.6200 | 71.32 | 0.6700 | – | – |
| KTGSL (2022) | **75.64** | **0.6900** | **72.58** | **0.6800** | **56.11** | **0.4900** |
| OFF-Apex (2019) | **67.68** | **0.6709** | 68.94 | 0.6967 | – | – |
| DSSN (2019) | 63.41 | 0.6462 | 70.78 | 0.7297 | 57.35 | 0.4644 |
| LGCcon (2021) | – | – | 65.02 | 0.6400 | 40.90 | 0.3400 |
| GEME (2021) | 64.63 | 0.6158 | **75.20** | **0.7354** | 55.88 | 0.4538 |
| AU-GCN (2021) | – | – | 74.27 | 0.7047 | **74.26** | **0.7045** |
| FeatRef (2022) | 57.90 | – | 62.85 | – | 60.13 | – |
| MFVAN | **79.87** | **0.8009** | **78.45** | **0.7616** | **76.47** | **0.7325** |

The bold values represent the state-of-the-art performance of Handcrafted features and deep learning methods, as well as the performance of our proposed MFVAN method.

TABLE 2 Micro-expression recognition performance comparison on the 3DB-combined datasets.

| Methods | SMIC | | CASME II | | SAMM | | 3DB-combined | |
|---|---|---|---|---|---|---|---|---|
| | UF1 | UAR | UF1 | UF1 | UF1 | UF1 | UF1 | UAR |
| LBP-TOP (2011) | 0.2000 | 0.5280 | 0.7026 | 0.5882 | 0.5882 | 0.7026 | 0.3954 | 0.4102 |
| Bi-WOOF (2018) | 0.5727 | 0.5829 | 0.7805 | 0.6296 | 0.6296 | 0.7805 | 0.5211 | 0.5139 |
| OFF-Apex (2019) | 0.6817 | 0.6695 | 0.8764 | 0.7196 | 0.7196 | 0.8764 | 0.5409 | 0.5409 |
| GEME (2021) | 0.6288 | 0.6570 | 0.8401 | 0.7395 | 0.7395 | 0.8401 | 0.6868 | 0.6541 |
| LGCcon (2021) | 0.6195 | 0.6066 | 0.7762 | 0.7499 | 0.4924 | 0.4711 | – | – |
| AU-GCN (2021) | **0.7192** | **0.7215** | 0.8798 | **0.7914** | **0.7914** | 0.8798 | **0.7751** | **0.7890** |
| FeatRef (2022) | 0.7011 | 0.7083 | **0.8915** | 0.7838 | 0.7838 | **0.8915** | 0.7372 | 0.7155 |
| MFVAN | **0.7986** | **0.7899** | **0.9061** | **0.8100** | **0.8100** | **0.9061** | **0.8322** | **0.8289** |

The bold values represent the state-of-the-art performance of the deep learning method and our proposed MFVAN method.

TABLE 3 Evaluation for global and local features on the SMIC (3 categories), CASME II (5 categories), and SAMM (5 categories).

| Methods | SMIC (3) | | CASME II (5) | | SAMM (5) | |
|---|---|---|---|---|---|---|
| | Accuracy | F1-Score | Accuracy | F1-Score | Accuracy | F1-Score |
| Global | 59.88 | 0.5964 | 62.29 | 0.5860 | 40.35 | 0.4056 |
| Global+RI | 62.67 | 0.6274 | 63.23 | 0.6288 | 43.01 | 0.4240 |
| Global+scale1 | 64.03 | 0.6394 | 76.55 | 0.7614 | 71.72 | 0.7061 |
| Global+scale1 + RI | 66.41 | 0.6661 | **79.45** | **0.7816** | 74.32 | 0.7164 |
| Global+scale1,2 | 67.47 | 0.6851 | 75.39 | 0.7427 | 75.11 | 0.7137 |
| Global+scale1,2 + RI | 72.15 | 0.7155 | 75.71 | 0.7500 | 75.89 | 0.7267 |
| Global+scale1,2,3 | 70.10 | 0.6941 | 75.92 | 0.7482 | 75.32 | 0.7297 |
| Global+scale1,2,3 + RI | **79.87** | **0.8009** | 78.45 | 0.7616 | **76.47** | **0.7325** |

The bold values represent the optimal performance compared to ablation experiments.

experimental comparisons of global features (Global), global features with removed identity attributes (Global+RI), and fusion with local features at different scales (Global+scale). The results are shown in Table 3. The experiment found that removing the interference of identity information by the apex to the onset frame mapping method can improve the performance of Accuracy and F1-Score, both in global features and global features fused with multi-scale local features. This situation also illustrates that removing the identity attributes can optimize the recognition performance of micro-expressions on SMIC, CASME II, and SAMM.

At the same time, we found that the performance of micro-expression recognition increases accordingly with the fusion of local features at multiple scales. In the SMIC and SAMM, multi-scale local feature fusion can better capture the local detail changes of the apex to the onset frame and improve the recognition performance of micro-expressions. However, we achieve the best performance by fusing Accuracy and F1-Score with the original scales in CASME II, which are 79.45 and 0.7816, respectively. Therefore, for the consistency of experimental results across all datasets, we use multi-scale fusion to obtain the final experimental results in the experimental validation process.

## 4.5. Visualization analysis

This section further uses the confusion matrix of the MFVAN model on the SMIC, CASME II, SAMM, and 3DB-Combined datasets to visually analyze the recognition performance of different types of micro-expressions. The experimental results are shown in Figure 5. In general, whether it is a single data set or a combined data set, the performance of the negative type is higher than the emotional performance of the positive type. The experimental results mainly include two reasons. First, in the process of constructing the micro-expression dataset. The negative emotion category of the subject is more likely to be stimulated, making the samples of negative emotions in the data set higher than the samples of positive emotions; on the other hand, the reason is that the MFVAN model tends to be more inclined to negative emotional types, this is exactly one of the problems that need to be solved in the follow-up.

We visualized analysis of the effect of multi-scale features on the micro-expression recognition through the Grad-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017). The experimental results are shown in Figure 6. The first column is the original image. The second column is the category activation map corresponding to the local features at the scale of the original image. The third and fourth columns are the category activation maps corresponding to the first and second convolutional feature map scales. In terms of the performance on the multi-scale local feature of the class activation maps corresponding to single-scale features, the local attention weights focused are more dispersed in the original image scale. And the local features at small scales are relatively more concentrated. Therefore, it is necessary to fuse local features from multiple scales to recognize micro-expressions.

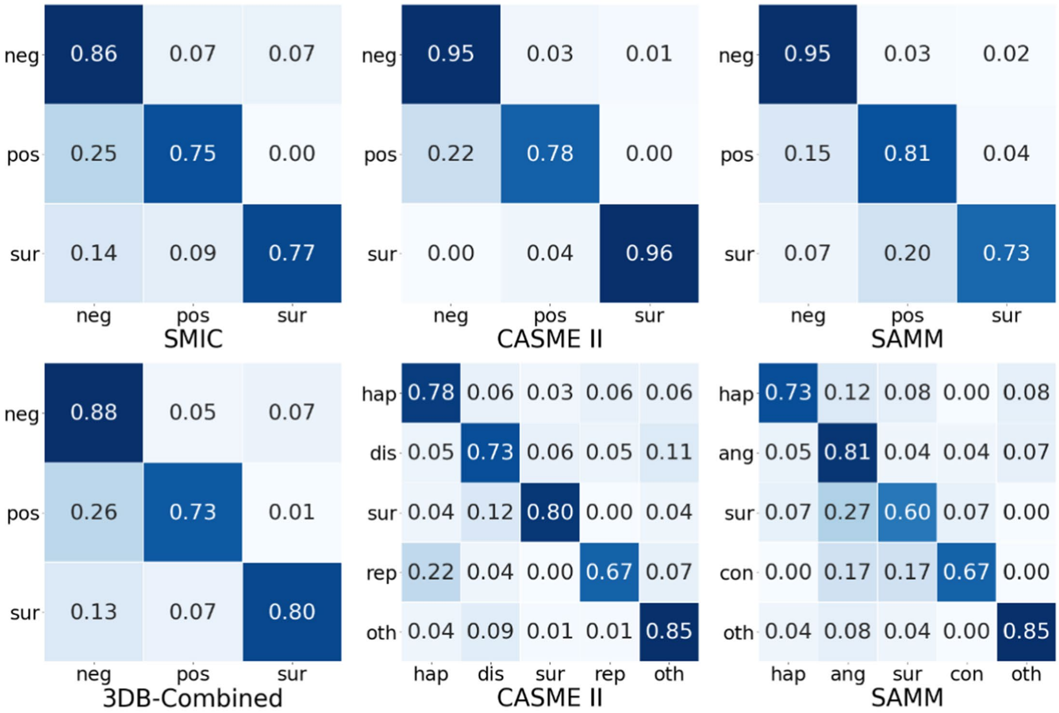

FIGURE 5
The confusion matrices on of MFVAN model on micro-expression datasets.

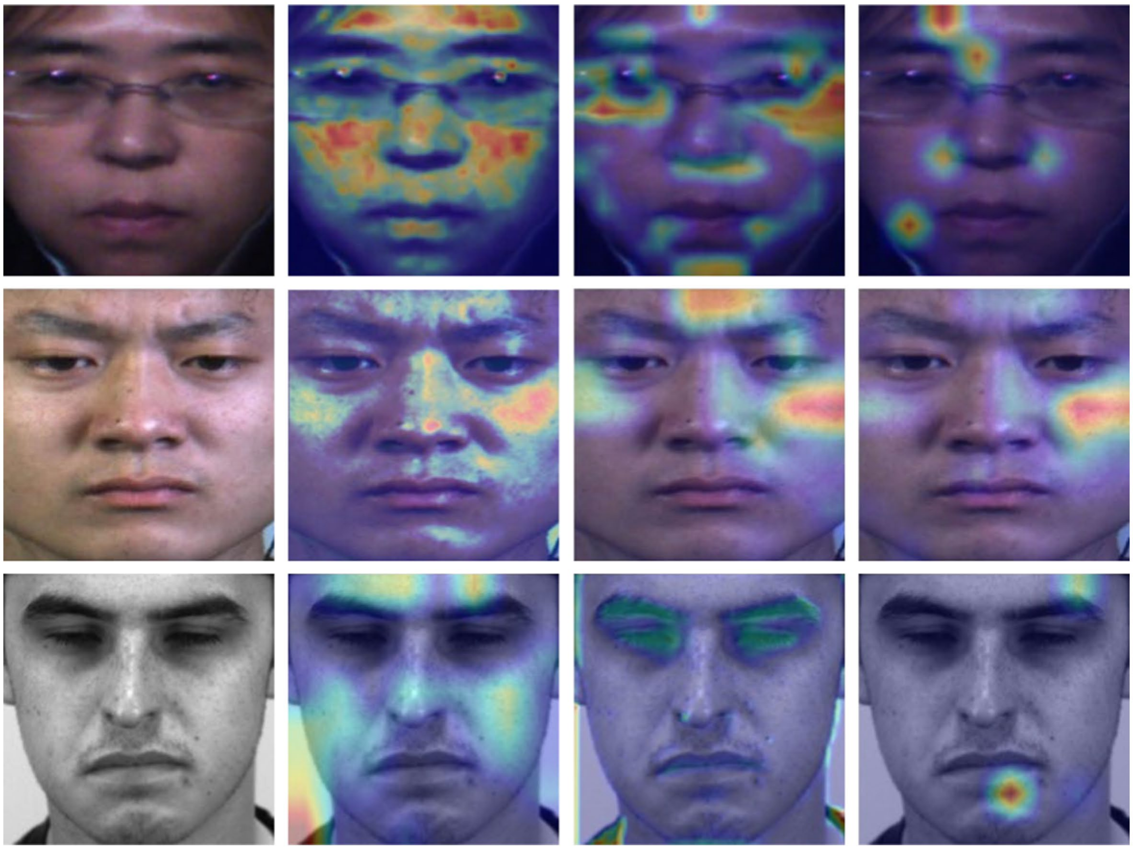

FIGURE 6
The Grad-weighted Class Activation Mapping of the multi-scale features on the SMIC, CASME II, and SAMM.

## 5. Conclusion

In this paper, we propose a multi-scale fusion visual attention network model that fuses the local attention weights of the multiple-scale feature maps of the removing identity attributes network for micro-expression recognition. For the problem of the small micro-expression dataset, a combination of unsupervised and transfer learning is used to reduce the influence of identity attributes by learning the mapping relationship from apex to onset frame in micro-expression video sequences. Then, the local detail features are extracted by focusing on multi-scale local attention weights. Finally, micro-expressions are classified by fusing global features with local features with high weights. In general, we reveal the impact of individual attributes on the localization of local ROIs. The experimental results show that a multi-scale fusion visual attention network contributes to micro-expression recognition.

The research work related to micro-expression analysis in this paper mainly discusses the micro-expression recognition problem, but often there is still how to locate the occurrence of micro-expressions in the real environment. In a real environment, the occurrence of micro-expressions is often to conceal true emotions, so micro-expressions are often accompanied by the occurrence of macro-expressions. How to locate the location of micro-expressions in a complex environment and emotional changes is also important to research in future work.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

HY: data curation, software, validation, investigation, writing—original draft preparation, and visualization. HP: methodology, software, validation, and visualization. LX: conceptualization, formal analysis, investigation, resources, and funding acquisition. ZW: resources and visualization. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported in part by the National Key R&D Program of China under Grant 2018YFC2001700 and in part by the Beijing Natural Science Foundation under Grant L192005.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Ahmad, M. I., Mubin, O., Shahid, S., and Orlando, J. (2019). Robot's adaptive emotional feedback sustains children's social engagement and promotes their vocabulary learning: a long-term child–robot interaction study. *Adapt. Behav.* 27, 243–266. doi: 10.1177/1059712319844182

Bai, M. (2020). "Detection of micro-expression recognition based on spatio-temporal modelling and spatial attention", in ACM international conference on multimodal interaction, New York: ACM. 703–707

Ben, X., Ren, Y., Zhang, J., Wang, S.-J., Kpalma, K., Meng, W., et al. (2021). Video-based facial micro-expression analysis: a survey of datasets, features and algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1–5846. doi: 10.1109/TPAMI.2021.3067464

Bisogni, C., Castiglione, A., Hossain, S., Narducci, F., and Umer, S. (2022). Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Trans. Ind. Inf.* 18, 5619–5627. doi: 10.1109/TII.2022.3141400

Chen, B., Zhang, Z., Liu, N., Tan, Y., Liu, X., and Chen, T. (2020). Spatiotemporal convolutional neural network with convolutional block attention module for micro-expression recognition. *Information* 11:380. doi: 10.3390/info11080380

Chowdary, M. K., Nguyen, T. N., and Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Comput. Appl.*, 33, 1–18. doi: 10.1007/s00521-021-06012-8

Davison, A. K., Lansley, C., Costen, N., Tan, K., and Yap, M. H. (2016). SAMM: A spontaneous micro-facial movement dataset. *IEEE Trans. Affect. Comput.* 9, 116–129. doi: 10.1109/TAFFC.2016.2573832

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: transformers for image recognition at scale Available at: http://arxiv.org/abs/2010.11929

Ekman, P. (2009). "Lie catching and microexpressions" in *The philosophy of deception*. ed. C. Martin (New York, NY: Oxford University Press), 118–133.

Ekman, P., and Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry* 32, 88–106. doi: 10.1080/00332747.1969.11023575

Ekman, P., and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* 17, 124–129. doi: 10.1037/h0030377

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

Happy, S., and Routray, A. (2017). Fuzzy histogram of optical flow orientations for micro-expression recognition. *IEEE Trans. Affect. Comput.* 10, 394–406. doi: 10.1109/TAFFC.2017.2723386

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition", in IEEE conference on computer vision and pattern recognition (CVPR), Piscataway: IEEE 770–778.

Holler, J., and Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends Cognit. Sci.* 23, 639–652. doi: 10.1016/j.tics.2019.05.006

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks", in IEEE conference on computer vision and pattern recognition (CVPR), 7132–7141.

Huang, X., Wang, S.-J., Liu, X., Zhao, G., Feng, X., and Pietikäinen, M. (2019). Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. *IEEE Trans. Affect. Comput.* 10, 32–47. doi: 10.1109/TAFFC.2017.2713359

Huang, J., Zhao, X., and Zheng, L. (2020). "SHCFNet on Micro-expression recognition system", in International congress on image and signal processing, BioMedical Engineering and Informatics (IEEE), 163–168.

Khor, H.-Q., See, J., Liong, S.-T., Phan, R.C., and Lin, W. (2019). "Dual-stream shallow networks for facial micro-expression recognition", in IEEE international conference on image processing, Piscataway: IEEE 36–40.

Kim, D.H., Baddar, W.J., and Ro, Y.M. (2016). "Micro-expression recognition with expression-state constrained spatio-temporal feature representations", in ACM international conference on multimedia (MM), 382–386.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). "Imagenet classification with deep convolutional neural networks", in Advances conference on neural information processing systems (NeurIPS), 1097–1105.

Kumar, A.J.R., and Bhanu, B. (2021). "Micro-expression classification based on landmark relations with graph attention convolutional network", in IEEE/CVF conference on computer vision and pattern recognition, Piscataway: IEEE 1511–1520.

Lei, L., Chen, T., Li, S., and Li, J. (2021). "Micro-expression recognition based on facial graph representation learning and facial action unit fusion", in IEEE/CVF conference on computer vision and pattern recognition, Piscataway: IEEE 1571–1580.

Li, S., and Deng, W. (2020). Deep facial expression recognition: a survey. *IEEE Trans. Affect. Comput.* 13, 1195–1215. doi: 10.1109/TAFFC.2020.2981446

Li, J., Dong, Z., Lu, S., Wang, S.-J., Yan, W.-J., Ma, Y., et al. (2022). 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, 1–2800. doi: 10.1109/TPAMI.2022.3174895

Li, Y., Huang, X., and Zhao, G. (2021). Joint local and global information learning with single apex frame detection for Micro-expression recognition. *IEEE Trans. Image Process.* 30, 249–263. doi: 10.1109/TIP.2020.3035042

Li, Y.-K., Meng, Q.-H., Yang, T.-H., Wang, Y.-X., and Hou, H.-R. (2022). Touch gesture and emotion recognition using decomposed spatiotemporal convolutions. *IEEE Trans. Instrum. Meas.* 71, 1–9. doi: 10.1109/TIM.2022.3147338

Li, X., Pfister, T., Huang, X., Zhao, G., and Pietikäinen, M. (2013). "A spontaneous micro-expression database: inducement, collection and baseline", in IEEE international conference on automatic face & gesture recognition, Piscataway: IEEE 1–6.

Li, J., Soladie, C., and Seguier, R. (2020). Local temporal pattern and data augmentation for micro-expression spotting. *IEEE Trans. Affect. Comput.* 14, 811–822. doi: 10.1109/TAFFC.2020.3023821

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

Liu, Y.-J., Li, B.-J., and Lai, Y.-K. (2018). Sparse MDMO: learning a discriminative feature for micro-expression recognition. *IEEE Trans. Affect. Comput.* 12, 1–261. doi: 10.1109/TAFFC.2018.2854166

Merghani, W., and Yap, M.H. (2020). "Adaptive mask for region-based facial Micro-expression recognition", in IEEE international conference on automatic face & gesture recognition, Piscataway: IEEE 765–770.

Nie, X., Takalkar, M. A., Duan, M., Zhang, H., and Xu, M. (2021). GEME: dual-stream multi-task GEnder-based micro-expression recognition. *Neurocomputing* 427, 13–28. doi: 10.1016/j.neucom.2020.10.082

Niu, M., Tao, J., Li, Y., Huang, J., and Lian, Z. (2019). "Discriminative video representation with temporal order for micro-expression recognition", in IEEE international conference on acoustics, speech and signal processing (ICASSP), Piscataway: IEEE 2112–2116.

O'sullivan, M., Frank, M. G., Hurley, C. M., and Tiwana, J. (2009). Police lie detection accuracy: the effect of lie scenario. *Law Human. Behav.* 33, 530–538. doi: 10.1007/s10979-008-9166-4

Pfister, T., Li, X., Zhao, G., and Pietikäinen, M. (2011). "Recognising spontaneous facial micro-expressions", in IEEE international conference on computer vision, Piscataway: IEEE 1449–1456.

See, J., Yap, M.H., Li, J., Hong, X., and Wang, S.-J. (2019). "MEGC 2019–the second facial micro-expressions grand challenge", in IEEE international conference on automatic face & gesture recognition Piscataway: IEEE 1–5.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: visual explanations from deep networks via gradient-based localization", in IEEE international conference on computer vision, Piscataway: IEEE 618–626.

Shao, J., and Qian, Y. (2019). Three convolutional neural network models for facial expression recognition in the wild. *Neurocomputing* 355, 82–92. doi: 10.1016/j.neucom.2019.05.005

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., et al. (2015). "Going deeper with convolutions", in IEEE conference on computer vision and pattern recognition (CVPR), Piscataway: IEEE 1–9.

Tran, T.-K., Vo, Q.-N., Hong, X., Li, X., and Zhao, G. (2021). Micro-expression spotting: a new benchmark. *Neurocomputing* 443, 356–368. doi: 10.1016/j.neucom.2021.02.022

Valstar, M.F., and Pantic, M. (2011). Fully automatic recognition of the temporal phases of facial actions. IEEE transactions on systems, man, and cybernetics, Part B (Cybernetics). Piscataway: IEEE 42, 28–43.

Wang, X., Kou, L., Sugumaran, V., Luo, X., and Zhang, H. (2020b). Emotion correlation mining through deep learning models on natural language text. *IEEE Trans. Cybern.* 51, 4400–4413. doi: 10.1109/TCYB.2020.2987064

Wang, C., Peng, M., Bi, T., and Chen, T. (2020a). Micro-attention for micro-expression recognition. *Neurocomputing* 410, 354–362. doi: 10.1016/j.neucom.2020.06.005

Wang, Y., See, J., Phan, R.C.-W., and Oh, Y.-H. (2014). "LBP with six intersection points: reducing redundant information in lbp-top for micro-expression recognition", in Asian conference on computer vision Berlin Springer 525–537.

Wang, Y., See, J., Phan, R. C.-W., and Oh, Y.-H. (2015). Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. *PLoS One* 10:e0124674. doi: 10.1371/journal.pone.0124674

Wang, Y., Zheng, S., Sun, X., Guo, D., and Lang, J. (2022). Micro-expression recognition with attention mechanism and region enhancement. *Multimedia Syst.*, 28, 1–9. doi: 10.1007/s00530-022-00934-6

Wei, J., Lu, G., and Yan, J. (2021). A comparative study on movement feature in different directions for micro-expression recognition. *Neurocomputing* 449, 159–171. doi: 10.1016/j.neucom.2021.03.063

Wei, J., Lu, G., Yan, J., and Liu, H. (2022a). Micro-expression recognition using local binary pattern from five intersecting planes. *Multimedia Tools Appl.* 81, 20643–20668. doi: 10.1007/s11042-022-12360-x

Wei, J., Lu, G., Yan, J., and Zong, Y. (2022b). Learning two groups of discriminative features for micro-expression recognition. *Neurocomputing* 479, 22–36. doi: 10.1016/j.neucom.2021.12.088

Wu, M., Teng, W., Fan, C., Pei, S., Li, P., and Lv, Z. (2023). An investigation of olfactory-enhanced video on eeg-based emotion recognition. *IEEE Trans. Neural Syst. Rehabil. Eng.* 31, 1602–1613. doi: 10.1109/TNSRE.2023.3253866

Xia, Z., Peng, W., Khor, H. Q., Feng, X., and Zhao, G. (2020). Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Trans. Image Process.* 29, 8590–8605. doi: 10.1109/TIP.2020.3018222

Xie, H.-X., Lo, L., Shuai, H.-H., and Cheng, W.-H. (2020). "Au-assisted graph attention convolutional network for micro-expression recognition", in ACM international conference on multimedia New York: ACM 2871–2880.

Xu, F., Zhang, J., and Wang, J. Z. (2017). Microexpression identification and categorization using a facial dynamics map. *IEEE Trans. Affect. Comput.* 8, 254–267. doi: 10.1109/TAFFC.2016.2518162

Yan, W.-J., Li, X., Wang, S.-J., Zhao, G., Liu, Y.-J., Chen, Y.-H., et al. (2014). CASME II: an improved spontaneous micro-expression database and the baseline evaluation. *PLoS One* 9:e86041. doi: 10.1371/journal.pone.0086041

Yu, M., Guo, Z., Yu, Y., Wang, Y., and Cen, S. (2019). Spatiotemporal feature descriptor for micro-expression recognition using local cube binary pattern. *IEEE Access.* 7, 159214–159225. doi: 10.1109/ACCESS.2019.2950339

Zhao, Z., Li, Q., Zhang, Z., Cummins, N., Wang, H., Tao, J., et al. (2021). Combining a parallel 2D CNN with a self-attention dilated residual network for CTC-based discrete speech emotion recognition. *Neural Netw.* 141, 52–60. doi: 10.1016/j.neunet.2021.03.013

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

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

Zhou, L., Shao, X., and Mao, Q. (2021). A survey of micro-expression recognition. *Image Vis. Comput.* 105:104043. doi: 10.1016/j.imavis.2020.104043

Zhu, J., Zong, Y., Chang, H., Xiao, Y., and Zhao, L. (2022). A sparse-based transformer network with associated spatiotemporal feature for micro-expression recognition. *IEEE Signal Process Lett.* 29, 2073–2077. doi: 10.1109/LSP.2022.3211200