

Received 13 April 2024, accepted 26 April 2024, date of publication 29 April 2024, date of current version 10 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3395116

RESEARCH ARTICLE

AST+SVMNet: A Novel Decomposition Method for Micro-Expression Recognition Based on Fusion Attention and Improved Spatio-Temporal Convolution by Feature Transfer

PEIYUN XUE^{1,2}, XIAOLONG GUO¹, JING BAI¹, AND BO YUAN³

¹College of Electronic Information and Optical Engineering, Taiyuan University of Technology, Taiyuan 030024, China

²Shanxi Academy of Advanced Research and Innovation, Taiyuan 030032, China

³North Automatic Control Technology Institute, Taiyuan 030006, China

Corresponding author: Peiyun Xue (xuepeiyun@tyut.edu.cn)


This work was supported in part by the Natural Science Foundation of Shanxi Province under Grant 20210302124544, and in part by the Applied Basic Research Project of Shanxi Province under Grant 201901D111094.

ABSTRACT Micro-expression (ME) is spontaneous, rapid, and subtle facial mechanism that can reveal the concealed emotions. However, the short duration, low motion intensity, and small dataset of MEs make the extraction and learning of features from ME samples more challenging for existing micro-expression recognition (MER) methods. To address this issue, we propose a novel decomposition MER method, called AST+SVMNet, and the primary architecture of our method integrates improved fusion attention and spatio-temporal convolutional neural network, achieving efficient MER through feature transfer to SVM. This method consists of four main components: feature extraction, fusion attention, spatio-temporal feature extraction, and feature transfer modules. In the feature extraction part, we designed a novel ME texture feature called the image sequence difference feature (ISDF). It mitigates the negative impact of optical flow calculation noise on MER task when applying optical flow features simultaneously. In the fusion attention part, we designed a fusion attention module (FAM) that reduces the extraction of redundant information for ME samples, optimizing the extraction of finer-grained spatio-temporal information. In the third part, we reduced the parameter count through 2D and 3D Inception Modules without compromising the performance of spatio-temporal feature extraction. In the feature transfer part, we achieved rapid and efficient MER by training the SVM classifier through feature transfer on high-dimensional spatio-temporal features. Finally, the performance of our proposed method on four publicly available spontaneous ME datasets surpasses that of existing baseline methods in MER. In addition, through effectiveness experiments and ablation studies, we demonstrated the effectiveness of the proposed texture feature ISDF and the MER method AST+SVMNet.

INDEX TERMS Micro-expression recognition, fusion attention, spatio-temporal feature, feature transfer, differential feature.

I. INTRODUCTION

Facial expression (FE) is a universal, generalized, natural pattern of human emotion transmission, which is closely related

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang .

to the mental state and emotional state of human beings in a specific environment [1]. FEs usually refer to macro-expressions, but when macro-expressions are subjectively suppressed, inner emotions can also be conveyed through micro-expressions (MEs) [2]. Unlike the macro-expressions, which are one of the components of FEs, and are directly

produced by people in their daily life and has obviously emotional orientation characteristics, MEs are **spontaneous** (the producer does not have the sense to control it) [3], **rapid** (the duration of MEs is usually only 1/25 second to 1/5 second) [4], and **subtle** (the facial area involved is small, and the amplitude of facial muscle-groups movement is small) [5]. These are the direct reasons why MEs do not reflect the emotional characteristics they point to. However, because of its own unique characteristics, MEs can more accurately reveal the real emotions that people try to hide, and reflect the actual feelings of people in the specific environment [6]. Therefore, micro-expression recognition (MER) has been applied and developed in many fields, including lie detection [7], clinical diagnosis [8], commercial negotiation [9], national security [10] and others.

The phenomenon of MEs was proposed by Haggard and Isaacs in 1966 when they explored the influence of FEs in psychotherapy [11]. Specifically for MER, because of its unique properties, ME features have led to inaccurate identification even by trained psychology professionals to ME [12]. Thus, requiring experts to accomplish MER well is inefficient, common and costly. In recent years, due to the advance of computer science, researchers have also proved that the effect and efficiency of using algorithms to realize MER is better than that of expert recognition [13], [14]. Therefore, how to improve the accuracy and effect of MER has become a crucial issue that requires resolution.

At the initial phase of the MER research, academics mainly based on hand-crafted feature-based descriptors to realize automatic recognition of ME [15]. Serving as an illustration of the classic approach of image texture features, the local binary pattern (LBP) from Three-Orthogonal Planes (LBP-TOP) [16] algorithm achieve good recognition results; As a representative of the geometric features of images, Optical flow describes the motion information of image sequences, and the representative work is TV-L1 [17] optical flow. In order to promote the healthy and orderly development of the field of MER, researchers have created a certain number of spontaneous ME datasets, such as CASME series datasets [18], [19], [20], [21], SMIC [22], SAMM [23], etc. In addition, the organization of Facial Micro-Expression Recognition Grand Challenges (MEGC) [24] has significantly propelled the development and innovation of MER technology.

As the successful expansion of deep learning applications in the field of computer vision continues, researchers are also attempting to apply deep learning to the domain of MER [25], [26], [27], [28]. Although the existing ME datasets have limited sample sizes, which pose certain challenges to network architecture design, professors have nonetheless proposed effective solutions tailored to these characteristics.

One of the representative examples is the proposal lightweight Dual-Stream Shallow Network (DSSN) [29] based AlexNet backbone which effectively mitigates the issues of low intensity in ME samples and the overfitting problems. In addition to that, there has been significant

development in MER based on Generative Adversarial Networks (GANs) [30]. On the level of features, unlike single optical flow and image texture features, multi-input features represent a promising research direction [31], [32]. Kim et al. [33] employed CNN to encode and LSTM to classify, achieving effective MER. Therefore, it can be seen that multi-input features can effectively explore spatio-temporal information and highlight the unique advantages of each sub-feature to some extent [34]. In the current research landscape, deep learning-based MER is the preferred solution for MER, offering state-of-the-art performance compared to other implementations [25], [29].

In the current research landscape, due to the short duration and limited muscle amplitude of ME samples, existing studies tend to extract a significant amount of irrelevant noise when capturing spatio-temporal correlations among frames in ME sample sequences [34]. To address these issues and improve the current state of research, this paper proposes a novel texture feature called Image Difference Sequence Feature (IDSF), which mitigates noise introduced by uneven lighting conditions during optical flow feature computation. Furthermore, existing feature extraction approaches exhibit limited capabilities in extracting spatio-temporal features of MEs and often involve a high number of parameters and computational complexity [25]. To enhance the feature extraction capabilities, reduce parameter count, and prevent overfitting due to the small size of ME datasets, we also propose an improved Fusion Attention and Spatio-temporal Convolutional Neural Network (AST+SVMNet) based on 3D and 2D Inception Module, which also includes independently designed Fusion Attention Module (FAM) that combines the spatial attention and channel attention mechanisms.

In summary, our main contributions are as follows:

- 1) We propose a novel decomposition method for the MER task: AST+SVMNet. The proposed method has a simplified network structure that utilizes the specially designed 2D or 3D Inception Module in combination with an SVM classifier to improve classification speed and efficiency.
- 2) We propose a novel texture feature for ME, called Image Difference Sequence Feature (IDSF). This feature reduces the negative impact of optical flow computational noise on the classification results through a predefined difference operation, which ultimately improves the classification accuracy.
- 3) We design a Fusion Attention Module (FAM), embedded in our method. This module combines the channel and spatial attention mechanisms and realizes the coupling of them through a unique sampling mechanism to reduce redundant information extraction and enhance model performance.

The remaining organization of this paper is as follows. **Section II** provides a brief overview of the current hand-crafted features for MEs and classic deep-learning based models applied to MER. **Section III** elaborates on the technical details of our proposed Image Difference Sequence

Feature (IDSF) and the AST+SVMNet model including its module. **Section IV** presents a detailed description of the used dataset, the experimental setup and experimental results. Finally, in **Section V**, we summarize the experimental findings and draw conclusions.

II. RELATED WORK

This section reviews existing MER methods. Based on the differences in feature extraction methods and classifiers, the MER approach can be primarily categorized into two types: handcrafted-machine learning methods and deep learning methods.

A. HANDCRAFTED-MACHINE LEARNING METHODS

In this type, the extracted features are primarily divided into two categories: texture features and geometric features.

The most representative texture feature is LBP-TOP and its variants. The LBP-TOP algorithm [16] involves three directions, where X and Y represent spatial coordinates, and T represents the time sequence. Based on these three directions, the algorithm extracts pixels from XY, XT, and YT planes using the LBP operator, generates histograms for each plane, and concatenates them in the order of XY, XT, YT to form the LBP-TOP feature. In addition to using the LBP-TOP operator, its variants have also been proposed to address various problems in MER. Based on Dual-Cross Patterns from Three Orthogonal Planes (DCP-TOP) [35], it is possible to enhance the directional information of features, further improving the accuracy of MER. Wang et al. [36] proposed LBP with six intersection points (LBP-SIP), it can remove duplicated encoding of the six intersection points, reducing redundancy and histogram length, which in turn improves computational speed. STLBP-IP [37] combines spatio-temporal LBP operator with integral projection (IP) to simultaneously capture texture information and temporal information, thereby enhancing recognition performance. Huang et al. [38] proposed Completed Local Quantized Patterns (CLQP), segmented quantization makes the features more robust to variations in lighting intensity and noise. Additionally, they extended this approach to three-dimensional space, calling it STCLQP.

The most representative geometric feature is Optical Flow and its variants. Optical flow descriptors capture relative motion information for MER by calculating changes in pixel intensity between the sequence of image frames [39]. Verburg and Menkovski [40] used the Histogram of Oriented Optical Flow (HOOF) to encode subtle changes in ME frame sequences. However, due to the susceptibility of HOOF to lighting variations, Happy and Routray [41] proposed Fuzzy HOOF (FHOOF) to overcome its shortcomings, making it less sensitive to computational noise and uneven lighting. Liang et al. [42] introduced another optical flow descriptor, Bi-Weighted Oriented Optical Flow, called Bi-WOOF, which calculates horizontal and vertical optical flow vectors between two frames and then constructs

histograms based on direction, magnitude, and optical flow strain. Bi-WOOF achieves better MER performance compared to HOOF. Liu et al. [43] proposed the Main Directional Mean Optical-flow (MDMO) feature, which computes the principal direction and average optical flow magnitude for each region of interest within the facial region. This feature vector extraction method takes into account both local motion information and spatial location, resulting in improved MER capabilities.

In this approach, various classifiers are applied, including SVM, RF, K-NN, SVD, and others. Among them, SVM is the most widely used [15]. Due to its robust classification performance, generalization ability, SVM is extensively applied in early MER research.

B. DEEP LEARNING METHODS

Deep learning methods can be primarily categorized into two types: handcrafted-deep learning methods and other deep learning methods.

In handcrafted-deep learning methods, the combination of traditional handcrafted feature extraction techniques with deep learning has achieved superior performance. Liang et al. [44] proposed the Off-ApexNet, which calculates optical flow information from the starting frame to the apex frame of each ME frame sequence, and this optical flow information is fed into a CNN model for feature enhancement and classification. Jin et al. [45] incorporated genetic algorithms (GA) into the Apex Frame Network to eliminate irrelevant information that does not contribute to expression prediction, enhancing features and improving recognition performance. Khor et al. [46] proposed ELRCN, which takes images, optical flow, and optical strain as input, feeds them into a CNN to extract spatial and spectral features, and then passes them into an LSTM for ME prediction. Choi and Song [47] transformed the landmarks of ME sequences into 2-D image information (LFM) and fed them into a cascaded network of CNN and LSTM to achieve MER. Liang et al. [48] introduced STST-Net, which extracts optical flow features and optical strain from Onset to Apex in each ME segment, and then vertical optical flow, horizontal optical flow, and optical strain are input into a shallow 3DCNN for classification. In addition to CNN, Capsule Neural Networks have also been applied to MER [49]. The CapsuleNet proposed by Quang et al. [50] achieves good MER performance through routing mechanisms and improved hierarchical relationships. Song et al. [51] implemented an end-to-end ME amplification model MEMM through the encoder-decoder network to simplify the MER task. Lei et al. [52] designed AU-GCN combining AU and graph convolutional network to realize end-to-end MER. In this method, the most widely used classifiers are MLP and softmax function.

Other deep-learning methods include attention mechanism-based methods and transfer learning-based methods. In the MER approach, the use of attention modules can enhance the encoding of spatial information for the Regions of

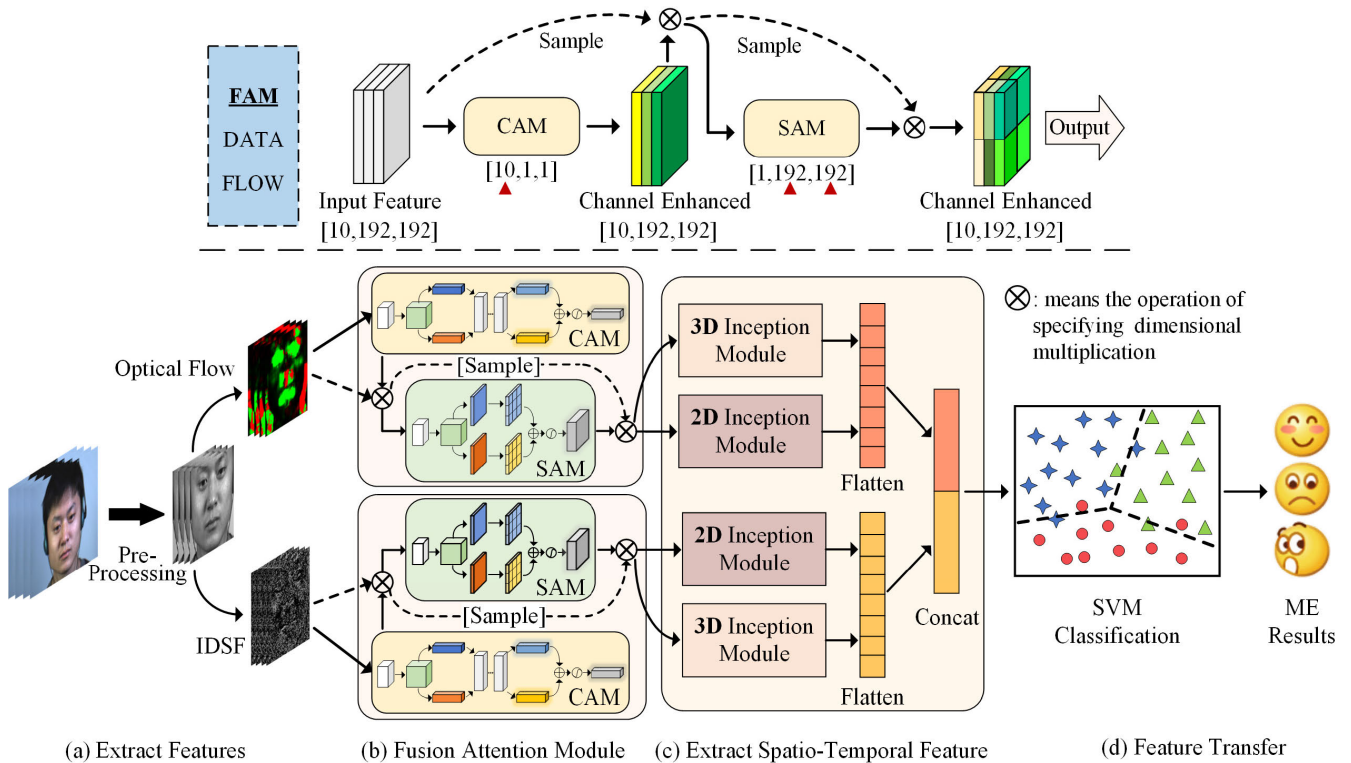


FIGURE 1. Basic framework for AST+SVMNet.

Interest (RoIs) of ME samples and improve the extraction of temporal information [53]. Wang et al. [54] proposed the Global Spatial-Temporal Attention Module (GAM), which simultaneously encodes both spatial and temporal information, enabling the extraction of more advanced features. Chen et al. [55] introduced CBAMNet, which utilizes a Convolutional Block Attention Module (CBAM) by cascading spatial attention modules and channel attention modules, achieving more accurate MER. Zhou et al. [56] proposed Feature Refinement (FR) and combined the attention mechanism to achieve ME feature refinement with MER. Ruisheng et al. [57] enhanced the performance of MER using crossVit as a backbone network combined with an improved cross-attention mechanism. In transfer learning-based methods, transfer learning from the FE dataset to the ME dataset can be achieved through fine-tuning [58], [59], knowledge distillation [60], and domain adaptation [61]. Such as Li et al. [62] proposed the domain adaptation-based DS-3DCNN, which bridges the gap between ME and macro-expressions through domain adaptation and achieves excellent results.

In summary, combining the advantages of handcrafted machine learning methods and deep-learning methods, we propose a novel decomposition method AST+SVMNet for MER task, resulting in promising experimental performance. The AST+SVMNet possesses the ability to extract multi-scale spatio-temporal features, enhanced local information attention capability, and feature fusion capability. And

through the feature transfer operation, AST+SVMNet has a better classification ability for ME. We also propose a new ME texture feature called image sequence difference feature (ISDF). ISDF employs a constrained difference algorithm to reduce the introduction of computational noise caused by uneven lighting in optical flow, which is a common issue in handcrafted features and can lead to poor recognition performance. In addition, we introduce the attention mechanism and design a fusion attention module (FAM) embedded in AST+SVMNet to enhance its feature extraction capability. We describe them in detail in the next section.

III. AST+SVMNET

This section provides a detailed description of the architecture of AST+SVMNet for MER, and Fig. 1 illustrates the general framework of the proposed method.

As shown in Fig. 1, our method is divided into four main parts: feature extraction module, fusion attention module, multiplexed spatio-temporal feature extraction module and feature transfer module. Among them, within the first part, we extracted dual-path features after preprocessing the input ME sequence: optical flow features and ISDF. Then, the dual-path features are respectively fed into the fusion attention module to realize the enhancement of the face motion information to enhance the ME representation. In the third part, we feed the enhanced dual-path features into the spatio-temporal feature extraction module respectively to realize spatio-temporal feature extraction and feature

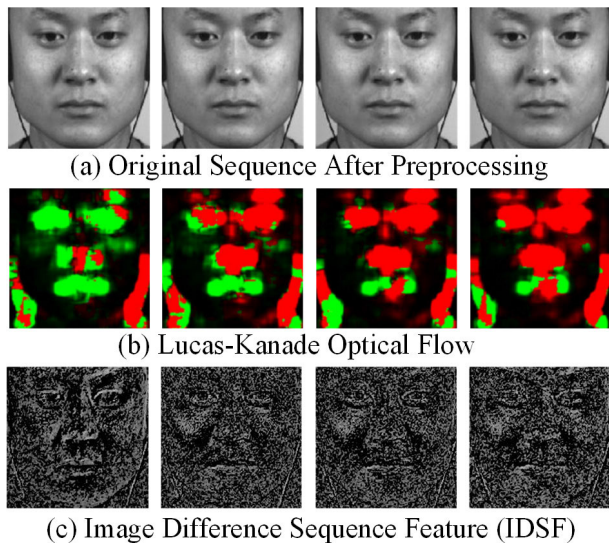


FIGURE 2. Example of ME features. Group (a) is the original frame after preprocessing; group (b) is an example of results for optical flow feature of group (a); group (c) is an example of ISDF of group (a).

fusion of the ME information. Finally, we feed the obtained high-dimensional features into SVM classifier for efficient classification.

A. EXTRACT FEATURES

1) DATA PRE-PROCESSING

In the preprocessing stage, the process for the ME samples mainly includes cropping, alignment and motion amplification, and finally the temporal interpolation model is used to normalize the number of video frames of the ME samples to a uniform value. We crop the ME samples from different ME datasets from the onset frame to the offset frame using the face detection method of the dlib library [63] based on their corresponding face sizes, retaining only the face information, with an image size of 192×192 . We use Active Shape Models (AAM) [15] to achieve face registration and eliminate the influence of head posture on MER. We use Eulerian Video Magnification method (EVM) for motion amplification between consecutive frames. Finally, we interpolated all ME sequences to 10 frames using Temporal Interpolation Model (TIM) [25] for temporal normalization. As a result, the size of the preprocessed single ME sample is $10 \times 192 \times 192$. Fig.2 group (a) shows a ME sample example of Happiness from the CASME II dataset, which is the result after preprocessing.

2) OPTICAL FLOW

In this part, we used the classical Lucas-Kanade algorithm [34] to implement the estimation of optical flow between sequences of ME sample frames. The algorithm reflects the features of ME by calculating the pixel position shift between two adjacent frames of ME samples. Specifically, the core step is to use the gradient information between pixels and

neighboring pixels within a local window to estimate the displacement vector using the least squares method. As a result, the subject's small facial movement changes are mapped to a pixel-level matching problem.

Fig.2 group (b) shows a simple example of LK optical flow extraction after preprocessing of a ME sample of Happiness from the CASME II dataset. Red area represents the horizontal component of the optical flow and green color represents the vertical component. We can find that the facial motion in this ME sample is mainly concentrated between the eyebrows and near the nose.

However, the calculation of LK optical flow is based on three assumptions: (1) constant brightness: the gray value of pixels between adjacent frames does not change; (2) small motion: the pixel positions between adjacent frames do not produce sharp changes; (3) spatial consistency: neighboring pixel points in the previous frame are also neighboring in the following frame. In practice, affected by the position of the light source, light intensity and other factors, it is difficult to ensure that the brightness of the surface of the target remains constant when it is in motion, so assumption (1) is difficult to be satisfied, and when the target moves faster, assumption (2) and assumption (3) are also difficult to be satisfied.

In order to extract the motion features of ME sample sequences more efficiently, and to supplement the feature information lost due to computational noise caused by unfounded assumptions in extracting optical flow features, we propose the image difference sequence feature (IDSF).

3) IMAGE DIFFERENCE SEQUENCE FEATURE (IDSF)

The expression of the image difference algorithm used in this paper is as follows when the effect of changing lighting conditions is not considered:

$$D_i(x, y) = I_i(x, y) - I_{onset}(x, y) \quad (1)$$

where, $I_i(x, y)$ is the gray value of a pixel of the i -th frame of the ME sequence at the point (x, y) , $I_{onset}(x, y)$ means the gray value of a pixel of the onset frame of the ME sequence at the point (x, y) . $D_i(x, y)$ is the feature matrix obtained of the ME sequence by the image difference operation. $D_i(x, y)$ means the difference value at point (x, y) after the image difference operation.

When using image difference algorithms to extract differential image sequence features for MEs, the feature values are affected by changes in lighting conditions. To solve this problem, an image difference operation is performed for each frame in the ME sequence with the onset frame, so that a differential image sequence with the same number of frames as the ME sequence can be obtained. Since the difference operation between the corresponding frame and the onset frame of the ME sequence, each frame in the differential image sequence takes into account the effects produced by changes in illumination. When the time interval between the i -th frame and the previous frame in the ME sequence is sufficiently small, the variation of lighting conditions is negligible for the extraction of ME motion features.

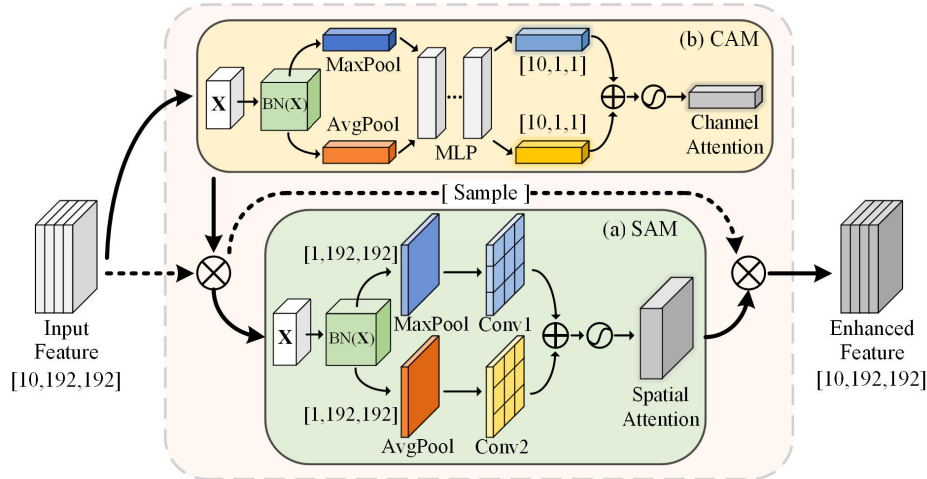


FIGURE 3. The actual architecture of the fusion attention module (FAM). Block (a) is the spatial attention module (SAM). Block (b) is the channel attention module (CAM).

In order to represent the noise introduced between the i -th frame and the onset frame due to changes in lighting conditions, we introduce a time dimension variable with the following expression:

$$\begin{aligned}
 F_i(x, y, t_i) &= F_i(x, y, t_0) + \delta(t_i) \\
 F &= \{F_1, F_2, \dots, F_n\} \\
 &= \{D_1 + \delta(t_1), \\
 &= D_2 + \delta(t_2), \\
 &\dots \\
 &= D_n + \delta(t_n)\}
 \end{aligned} \tag{2}$$

where, $\delta(t_i)$ means the noise introduced due to the change of illumination at the moment t_0 to t_i during the computation of the i -th frame of the differential image. F_i means the i -th differential image taking into account the noise. $F = \{F_1, F_2, \dots, F_n\}$ means the obtained differential image sequence features.

In this paper, when normalizing the ME frames, we use the equal interval sampling method, and 10 frames are taken at equal time intervals for each ME sequence. Thus, there is $n = 10$, and the expression for the interval time Δt is shown as:

$$\Delta t = t_1 - t_0 = t_2 - t_1 = \dots = t_n - t_{n-1} \tag{4}$$

Therefore, equation (3) can be written as:

$$\begin{aligned}
 F &= \{F_1, F_2, \dots, F_n\} \\
 &= \{D_1 + \delta(t_1), \\
 &= D_2 + \delta(t_1 + \Delta t), \\
 &\dots, \\
 &= D_n + \delta(t_1 + (n - 1)\Delta t)\}
 \end{aligned} \tag{5}$$

Since the change in luminance with time is smooth and continuous and n is a finite value, the noise due to illumination

during the time interval Δt is negligible when the Δt is small enough, there is:

$$\delta(t + \Delta t) = \delta(t), \Delta t \rightarrow 0 \tag{6}$$

So, equation (5) can be written as:

$$\begin{aligned}
 F' &= \{F'_1, F'_2, \dots, F'_n\} \\
 &= \{D_1 + \delta(t_1), \\
 &= D_2 + \delta(t_1), \\
 &\dots, \\
 &= D_n + \delta(t_1)\}
 \end{aligned} \tag{7}$$

Thus, we obtain the image difference sequence feature (IDSF) $F' = \{F'_1, F'_2, \dots, F'_n\}$. Each frame in the IDSF takes into account the noise with respect to the onset frame, and the noise of each frame is $\delta(t_1)$, which means the noise between the first frame of the ME sequence and the onset frame. When IDSF is used as the input feature of the neural network, since the noise introduced in each frame is $\delta(t_1)$, which is equivalent to adding a constant to the original feature, it does not affect the training results during training, which avoids the effect of changes in lighting conditions on the classification results. Fig.2 group (c) shows a simple example of IDSF extraction after preprocessing of a ME sample labeled Happiness in the CASME II dataset, which can visualize the motion information of ME samples.

B. FUSION ATTENTION MODULE

Cause both the KL optical flow features and the IDSF used in our method have multiple dimensions, we not only have to localize spatially significant regions, but also learn the significance of different channel. We introduced both the spatial attention mechanism and the channel attention mechanism [55] to design a fusion attention module, which ensure that all the important regions of the input features with high contribution to the final classification and recognition results

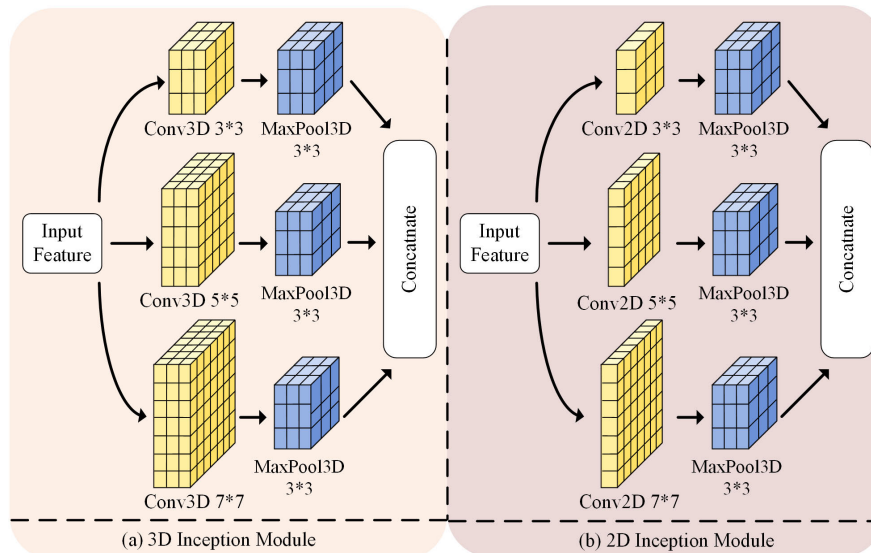


FIGURE 4. The specific architecture of the 3D and 2D inception module.

are localized in order to extract higher dimensional features with finer granularity through the spatio-temporal feature extraction module. Fig. 3 shows the design idea and details of this fusion attention module.

For the classical spatial attention model, after extracting the coarse model and fine model using average pooling and max pooling respectively, we concatenate these two models and process them with the same convolutional kernel. In order to more precisely locate regions and identify feature information that contributes significantly to the final result, we separated the average pooling and max pooling. We processed the coarse model and fine model separately using two channels, allowing distinct treatment of the coarse and fine information. Fig. 3 (a) illustrates the details of the separable SAM, where X represents the input features of the model with dimensions $10 \times 192 \times 192$. Simultaneously, we perform batch normalization on the input features to prevent gradient vanishing during training. After the one path of max pooling and one path of average pooling operations, the output dimensions become $1 \times 192 \times 192$. Finally, we perform a convolutional operation on the features from the two paths and add them together, using activation function to obtain the output of the SAM.

The processing flow of the CAM also consists of two branches. Similarly, after performing batch normalization on the input features, we apply average pooling and max pooling operations, resulting in an output with dimensions $10 \times 1 \times 1$. However, what differs is that the output needs to go through a multi-layer perceptron (MLP) for transformation to obtain two-channel excitation signals. Subsequent operations are similar to the SAM. Fig. 3 (b) illustrates the details of the CAM.

In the design of FAM, we use the Softmax activation function. For the internal connections of the two sub-blocks, inspired by the residual connection [64], we multiply the

channel values of the feature map with the corresponding channel values of the CAM output. The result is then fed into the SAM. We refer to the former operation as sampling. For the subsequent processing, similarly, we perform multiplication between the channel values of the output from the SAM and the corresponding channel values of the sampling input features of that sub-block. This forms the design framework of FAM. Therefore, we achieve enhanced feature representation for ME features through FAM to facilitate subsequent modules in extracting more fine-grained spatio-temporal features.

C. EXTRACTION OF SPATIO-TEMPORAL FEATURES

Considering the challenge that rapid and subtle movements in ME can make it difficult to extract effective features in ME samples. Inspired by the parallel processing of InceptionNet [53] and TSNN [65], we designed a neural network architecture for MER called the 3D/2D Inception Module, which utilizes both 3D or 2D convolutional kernels to simultaneously extract spatio-temporal information. This module is capable of capturing information about ME features along the temporal axis and the motion information in spatial axis. Fig. 1 (c) illustrates the simplified process of spatio-temporal feature extraction.

Fig. 4 shows the specific architecture of the 3D/2D inception module. The input to each module is the ME features with enhanced feature representation output from the previous module, and its dimensions are down-sampled to $10 \times 32 \times 32$. We use the 2D inception module to extract motion information in spatio axis for ME, and the 3D inception module is employed to extract the temporal sequence information of the input features. After extracting spatio-temporal information from both branches of ME features, we use the concatenate operation to achieve the fusion of the two streams. In terms of activation function selection, we employed the ReLU

TABLE 1. The detailed parameters and architectures of the 3D and 2D inception module.

Element	Layer Name	Kernel Size	Output	Parameter	Connect To
Input	<i>input_layer</i>	N/A	10×32×32	0	N/A
2D Inception Module	<i>conv2d_1</i>	16(1×1×1)	10×32×32×16	32	<i>input_layer</i>
	<i>conv2d_2</i>	16(1×3×3)	10×32×32×16	160	<i>input_layer</i>
	<i>conv2d_3</i>	16(1×5×5)	10×32×32×16	416	<i>input_layer</i>
	<i>max_poll3d_1</i>	3×3×3	4×11×11×16	0	<i>conv2d_1</i>
	<i>max_poll3d_2</i>	3×3×3	4×11×11×16	0	<i>conv2d_2</i>
	<i>max_poll3d_3</i>	3×3×3	4×11×11×16	0	<i>conv2d_3</i>
3D Inception Module	<i>conv3d_4</i>	16(3×3×3)	10×32×32×16	448	<i>input_layer</i>
	<i>conv3d_5</i>	16(3×5×5)	10×32×32×16	1216	<i>input_layer</i>
	<i>conv3d_6</i>	16(3×7×7)	10×32×32×16	2368	<i>input_layer</i>
	<i>max_poll3d_4</i>	3×3×3	4×11×11×16	0	<i>conv3d_4</i>
	<i>max_poll3d_5</i>	3×3×3	4×11×11×16	0	<i>conv3d_5</i>
	<i>max_poll3d_6</i>	3×3×3	4×11×11×16	0	<i>conv3d_6</i>
Concatenate	<i>Concatenate</i>	N/A	4×11×11×96	0	<i>max_poll3d_1</i> <i>max_poll3d_2</i> <i>max_poll3d_3</i> <i>max_poll3d_4</i> <i>max_poll3d_5</i> <i>max_poll3d_6</i>
Flatten	<i>Flatten</i>	N/A	46464	0	<i>concatenate</i>

function in all convolutional layers and only used the Softmax function in the final concatenate layer.

With the increase in the number of network layers, effective features in ME samples are prone to loss. The 3D/2D inception module increases network width while reducing network depth. Without sacrificing performance, it reduces the original trainable parameters, thereby improving the efficiency of model training. Table 1 shows the detailed parameters for this section.

D. FEATURE TRANSFER

In this part, we use SVM as the classifier to achieve lightweight and high-quality MER. The high-dimensional features extracted by the 3D/2D inception module are transferred to the SVM classifier for training to generate the decision boundary. In our method, we employ the radial basis function kernel for the SVM and use the One Versus Rest (OVR) decision function. This configuration requires training three binary classifiers to achieve three-class classification. When determining the category, the test data is input into the three classifiers separately, and the one with the highest evaluation score is chosen as the final determined category.

IV. EXPERIMENT

This section provides a detailed description of the dataset used, experimental configuration details, and specifies the evaluation metrics and strategies for the experimental results. Finally, based on the above information, we evaluate and analyze the experimental results.

A. DATASET

In order to evaluate the performance of the proposed method, we conducted experiments on the following four existing public spontaneous ME datasets.

1) CASME II

The CASME II dataset [19] was introduced and has been made publicly available by the Institute of Psychology of the Chinese Academy of Sciences in 2014. In comparison to its predecessor, CASME, this dataset includes more ME samples, totaling 247 samples collected from 26 subjects with an average age of 22.59 years. The ME samples were collected in a completely controlled laboratory environment. Additionally, high-speed cameras were used during ME sample collection, achieving a frame rate of 200Hz and a resolution of 640*480. The face size in the video samples is approximately 250*340. The dataset labels include five emotional types: Happiness (33), Surprise (60), Disgust (25), Repression (27), Others (102).

2) SMIC

The SMIC dataset [22] was introduced by the University of Oulu in 2013. It consists of three subsets: HS, VIS, and NIR, which differ in the type of camera used during sample collection. For this experiment, we primarily utilized the SMIC (HS) subset. This dataset comprises 164 ME samples collected from 16 subjects, with an average age of 26.7 years. The resolution of the samples is 1280*720, and due to the use of a high frame rate camera, the frame rate is 100Hz, higher than the frame rates of the other two subsets. The face size

TABLE 2. The specific distribution of the MEGC2019 dataset.

Class	Number	Original Label	Original Dataset
Negative	250	Anger	SAMM
		Sadness	SAMM
		Fear	SAMM
		Contempt	CASME II
		Negative	SMIC(HS)
		Repression	CASME II
		Disgust	CASME II, SAMM
Positive	109	Happiness	CASME II, SAMM
		Positive	SMIC(HS)
Surprise	83	Surprise	CASME II, SAMM SMIC(HS)

is approximately 190*230. The dataset labels include three emotional types: Positive (51), Negative (70), Surprise (43).

3) SAMM

The SAMM dataset [23] was introduced and has been made publicly available by Manchester Metropolitan University in 2016, comprises a total of 159 ME samples collected from 32 subjects with an average age of 33.34 years. The subjects come from 13 different ethnicities. The ME samples were collected in a completely controlled laboratory environment. The samples have a resolution of 2040*1088, a frame rate of 200Hz, and the face size in the video samples is approximately 400*400. The dataset labels include eight emotional types: Happiness (24), Anger (20), Surprise (13), Disgust (8), Fear (7), Sadness (3), Others (84).

4) MEGC2019

The challenge integrated the SMIC(HS), CASME II, and SAMM datasets, standardized the ME emotion label criteria, and mapped the sample label types to a set of three common emotion categories: Negative, Positive, Surprise. The MEGC2019 dataset [24] includes a total of 442 samples, with 145, 164, and 133 samples extracted from SMIC(HS), CASME II, and SAMM, respectively. Table 2 illustrates the specific distribution of the MEGC2019 dataset.

In order to evaluate the performance of the model on different datasets, we designed the experiments based on the three-category ME emotion label standard of the MEGC2019 dataset, including the three-category experiments, effectiveness experiments and ablation experiments.

B. EXPERIMENTAL SETUP

To accurately and independently evaluate each ME sample in the dataset, we adopted the Leave-One-Subject-Out (LOSO) cross-validation protocol to assess the performance of the proposed method. In each round of training, the ME samples of an individual subject were used as the validation set, while the samples of the remaining subjects constituted the training set for model training. This process was repeated for each subject's ME samples until all of them had been used as the

TABLE 3. Environment configurations for experiments.

Category	Specification
System	Windows 10
CPU	Intel i9 12900H
GPU	Nvidia RTX 3060 8G
RAM	64G
Python	3.6.13
Tensorflow	2.2.0
Keras	2.4.3
CUDA	10.2

training set. The results obtained using this method represent the overall model performance.

Due to the uneven sample sizes in the ME datasets, we employed Unweighted Average Recall (UAR) and Unweighted F1-Score (UF1) as experimental evaluation metrics. The definitions of UAR and UF1 are as follows:

- UAR: An unweighted average of all recalls. After calculating the Recall for each class, the average is taken, without considering the class imbalance. The formula is as follows:

$$UAR = \frac{\sum_{i=1}^{N_c} Recall_i}{N_c} \quad (8)$$

where, $Recall_i$ represents the recall of a category, N_c means the total number of labels.

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \quad (9)$$

In this formula, TP_c represents the final number of the samples correctly predicted as positive in the class c , FN_c represents the final number of the samples incorrectly predicted as negative in the class c .

- UF1-Score: An unweighted average of precision and recalls. The UF1-Score measures the model's performance by balancing precision and recall. The formula is as follows:

$$UF1 = \frac{\sum_{i=1}^{N_c} UF1_i}{N_c} \quad (10)$$

$$UF1_c = \frac{2 \times TP_c}{2 \times TP_c + FP_c + FN_c} \quad (11)$$

where, $UF1_c$ represents the UF1-score of a category, N_c , TP_c and FN_c is same as above. FP_c represents the final number of the samples incorrectly predicted as positive in class c .

In the training cycle during experiments, the optimizer is Adam (Adaptive Moment Estimation) with a learning rate of $1e-3$. The batch size is set to 32. Table 3 shows additional configurations of the experimental platform in hardware and software environment.

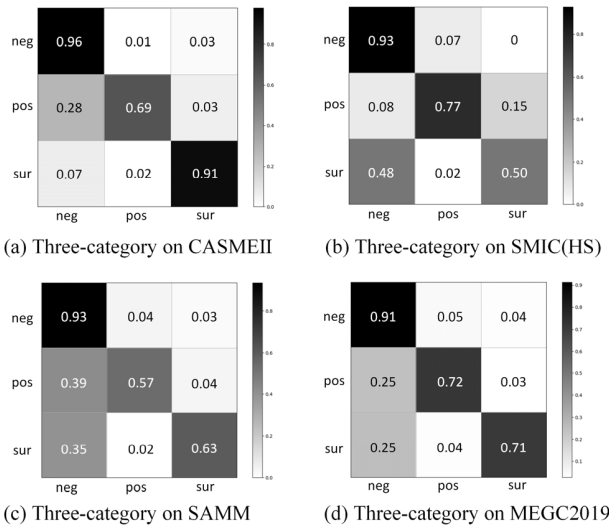


FIGURE 5. Confusion matrices on four spontaneous ME datasets.

C. RESULT AND DISCUSSION

The experimental results consist of three components. Firstly, we horizontally compared the performance of the proposed AST+SVMNet with other state-of-the-art models in the ME three-classification task across four different datasets. Fig. 5 shows the primary confusion matrices on four spontaneous ME datasets. Secondly, we evaluated the performance of the proposed IDSF in comparison with other ME features, validating the efficacy of the IDSF feature. Finally, to prove the effectiveness of the proposed method, we conducted a comprehensive ablation analysis on AST+SVMNet, encompassing input features, structural aspects, network layers and classification simultaneously.

1) THREE-CATEGORY EXPERIMENT ON CASME II

We conduct a three-category experiment on the CASME II dataset. We compare our proposed method with representative approaches in recent years in the field of MER. In this three-category experiment, we employ UAR and UF1-score as evaluation metrics to assess the effectiveness of our proposed method compared to advanced methods in recent years, evaluating its performance. We compared the performance of our model with traditional handcrafted features such as LBP-TOP [16] and deep learning-based methods like STST-Net [48], AU-GCN [52], and DACrossViT [57], etc. Table 4 shows the specific results of this experiment.

Our proposed AST+SVMNet achieved a UAR of 0.892 and an F1-score of 0.913 on the CASME II dataset, demonstrating a slight performance improvement compared to the MER-baseline method and recent advanced methods. In comparison to AU-GCN [52], our method exhibited a UAR improvement of 0.021, while compared to FR [56], our proposed method showed a UF1-score improvement of 0.029. The experimental results confirm the strong competitiveness

TABLE 4. CASME II dataset three-category experiments.

Year	Methods Approach	UAR	UF1
2011	LBP-TOP [16]	0.743	0.703
2018	Bi-WOOF [42]	0.803	0.781
2019	OFF-ApexNet [44]	0.868	0.876
2020	CapsuleNet [50]	0.702	0.707
2020	RCN [46]	0.856	0.809
2020	LFM [47]	0.840	0.870
2020	STSTNet+GA [45]	0.887	0.887
2021	AU-GCN [52]	0.871	0.879
2022	FR [56]	0.891	0.884
2023	DACrossViT [57]	0.864	0.866
2023	Ours	0.892	0.913

TABLE 5. SMIC(HS) dataset three-category experiments.

Year	Methods Approach	UAR	UF1
2011	LBP-TOP [16]	0.528	0.200
2018	Bi-WOOF [42]	0.583	0.573
2019	OFF-ApexNet [44]	0.670	0.682
2020	CapsuleNet [50]	0.588	0.582
2020	RCN [46]	0.599	0.598
2020	STSTNet+GA [45]	0.644	0.717
2021	TSNN-LF [65]	0.683	0.692
2022	SLSTT-LSTM [28]	0.720	0.740
2023	TFT [32]	0.718	0.741
2023	RES-CapsNet [49]	0.685	0.690
2023	Ours	0.758	0.712

of our proposed AST+SVMNet in the ME three-category experiment on the CASME II dataset.

2) THREE-CATEGORY EXPERIMENT ON SMIC(HS)

We conduct a three-category experiment on the SMIC(HS) dataset. We compare our proposed method with representative approaches in recent years in the field of MER. In this three-category experiment, we employ UAR and UF1-score as evaluation metrics to assess the effectiveness of our proposed method compared to advanced methods in recent years, evaluating its performance. We compared the performance of our model with traditional handcrafted features such as LBP-TOP [16] and deep learning-based methods like SLSTT-LSTM [28], RES-CapsNet [49] and TSNN-LF [65], etc. Table 5 shows the specific results of this experiment.

In the SMIC(HS) dataset, our proposed AST+SVMNet achieved the UAR of 0.758 and an F1-score of 0.712. Compared to TSNN-LF [65], our method demonstrated an improvement of 0.075 in UAR and an increase of 0.020 in UF1-score. In comparison to SLSTT-LSTM [28], our method exhibit a UAR improvement of 0.018, although it slightly lagged in UF1-score. The experimental results affirm that our proposed AST+SVMNet maintains a leading position in the three-category performance on the SMIC(HS) dataset.

TABLE 6. SAMM dataset three-category experiments.

Year	Methods Approach	UAR	UF1
2011	LBP-TOP [16]	0.410	0.395
2018	Bi-WOOF [42]	0.514	0.521
2019	OFF-ApexNet [44]	0.539	0.541
2020	CapsuleNet [50]	0.599	0.621
2020	RCN [46]	0.698	0.677
2020	STSTNet+GA [45]	0.670	0.668
2021	GEME [26]	0.545	0.584
2021	MERSiamC3D [27]	0.728	0.744
2022	MobileViT [59]	0.678	0.743
2022	SLSTT-LSTM[28]	0.643	0.715
2023	TFT [32]	0.656	0.709
2023	Ours	0.778	0.751

3) THREE-CATEGORY EXPERIMENT ON SAMM

Similarly, we conduct a three-category experiment on the SAMM dataset, utilizing UAR and UF1-score as evaluation metrics and LOSO protocol.

In this three-category experiment, we compared the performance of our model with traditional handcrafted features such as LBP-TOP [16], Bi-WOOF [42], and deep learning-based methods like GEME [26], MERSiamC3D [27], and TFT [32], etc. Table 6 shows the specific results of this experiment.

As shown in Table 6, our proposed AST+SVMNet achieved the UAR of 0.778 and an F1-score of 0.741 on the SAMM dataset. Compared to GEME [26] and TFT [32], our method demonstrated improvements of 0.233 and 0.122 in UAR, respectively. In terms of UF1-score, our method outperformed GEME by 0.167 and outperformed TFT by 0.042. The experimental results confirm that our proposed AST+SVMNet excels in UAR in the three-category experiment on the SAMM dataset compared to current advanced MER methods.

4) THREE-CATEGORY EXPERIMENT ON MEGC2019

As shown in Table 7, our proposed AST+SVMNet achieved the UAR of 0.832 and a UF1-score of 0.807 on the MEGC2019 fusion ME dataset. Compared to the existing MER-baseline method LBP-TOP [27], Bi-WOOF [42], and state-of-the-art methods such as MEMM4+Resnet50 [51] and Inceptr [53], our UAR in this experiment improved by 0.253, 0.209, 0.008, and 0.079. Our UF1-score also increased by 0.219, 0.177, 0.001, and 0.061, respectively. The experimental results affirm that our proposed AST+SVMNet exhibits strong competitiveness in the three-category experiment on the MEGC2019 mixed ME dataset.

5) IDSF VALIDITY EXPERIMENT

In this section, we validated the effectiveness of our proposed ME feature IDSF, across four datasets. The network framework utilized our proposed AST+SVMNet, with variations limited to the input features, specifically focusing on

TABLE 7. MEGC2019 dataset three-category experiments.

Year	Methods Approach	UAR	UF1
2011	LBP-TOP [16]	0.579	0.588
2018	Bi-WOOF [42]	0.623	0.630
2019	OFF-ApexNet [44]	0.710	0.720
2020	CapsuleNet [50]	0.651	0.652
2020	RCN [46]	0.716	0.705
2020	LFM [47]	0.750	0.770
2020	STSTNet+GA [45]	0.789	0.766
2022	MEMM [51]	0.824	0.806
2023	Inceptr [53]	0.753	0.746
2023	Ours	0.832	0.807

single-route handcrafted features. We compared the performance of existing classical handcrafted features, and the specific results are shown in Table 8. Compared to the best-performing method, our proposed ISDF demonstrated significant improvements. In the CASME II dataset, UAR improved by 0.098 and UF1-score increased by 0.067. In the SMIC(HS) dataset, UAR increased by 0.148 and UF1-score by 0.126. For the SAMM dataset, UAR showed a notable improvement of 0.248, with UF1-score increase of 0.228. In the MEGC2019 mixed ME dataset, UAR improved by 0.181, and UF1-score increased by 0.157. Substantial improvements in both UAR and UF1-score were observed across all four datasets, confirming the effectiveness of our proposed ISDF.

6) ABLATION EXPERIMENTS

To validate the effectiveness of our proposed AST+SVMNet, we conduct ablation experiments on four datasets, examining four aspects: input features, structural aspects, network layers and classification. We compare the proposed method with its corresponding variants.

At the part of input features, we compare the original framework using only single optical flow features and single IDSF with the dual-path feature framework on four datasets. The specific results are shown in Table 9. In addition, we also compare the performance of a single feature input under the influence of different classifiers at four ME datasets. On the CASME II, SAMM, and MEGC2019 datasets, the complete framework shows significant improvements compared to using only single-path features, with the UAR improving by a maximum of 0.159, 0.065, and 0.109, and the UF1-score improving by a maximum of 0.192, 0.049, and 0.073, respectively. Apart from this, the experimental results also show that the SVM classifier outperforms the FC+softmax classifier for classification with a single input feature. Because the small-parameter shallow parallel network designed by us to handle the ME samples does not have a high degree of dimensionality in the ME samples, it leads to a much smaller computational effort in generating the decision boundaries for the SVM, enabling it to generate more robust boundaries and improve the classification results. However, on the

TABLE 8. Experimental validation of the effectiveness of IDSF.

Input Feature	CASME II		SMIC (HS)		SAMM		MEGC2019	
	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1
LBP-TOP (2011) [16]	0.742	0.703	0.528	0.200	0.410	0.395	0.579	0.588
STLBP-IP (2015) [37]	0.701	0.714	0.543	0.547	0.502	0.509	0.594	0.617
BI-WOOF (2018) [42]	0.743	0.767	0.595	0.621	0.514	0.521	0.623	0.630
IDSF (Ours)	0.841	0.834	0.743	0.747	0.762	0.749	0.804	0.787

TABLE 9. Ablation experiment in different input features on four datasets.

Structure (classifier)	CASME II		SMIC (HS)		SAMM		MEGC2019	
	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1
Optical Flow-only (SVM)	0.743	0.726	0.698	0.701	0.716	0.703	0.729	0.738
Optical Flow-only (Softmax)	0.733	0.721	0.697	0.698	0.713	0.702	0.723	0.734
IDSF-only (SVM)	0.841	0.834	0.743	0.747	0.762	0.749	0.804	0.787
IDSF-only (Softmax)	0.824	0.817	0.736	0.704	0.753	0.744	0.789	0.783
AST+SVMNET	0.892	0.913	0.758	0.712	0.778	0.751	0.832	0.807

TABLE 10. Ablation experiment in structural aspects on four datasets.

Structure (classifier)	CASME II		SMIC (HS)		SAMM		MEGC2019	
	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1
Without FAM (SVM)	0.746	0.727	0.664	0.635	0.678	0.643	0.693	0.672
Without FAM (Softmax)	0.723	0.714	0.623	0.594	0.637	0.605	0.667	0.646
CAM-only (SVM)	0.865	0.844	0.747	0.709	0.754	0.743	0.801	0.784
CAM-only (Softmax)	0.846	0.831	0.734	0.698	0.736	0.731	0.789	0.765
SAM-only (SVM)	0.868	0.858	0.732	0.703	0.755	0.726	0.803	0.791
SAM-only (Softmax)	0.853	0.845	0.729	0.687	0.738	0.715	0.798	0.776
AST+SVMNET	0.892	0.913	0.758	0.712	0.778	0.751	0.832	0.807

TABLE 11. Ablation experiment in network layers on four datasets.

Network (classifier)	CASME II		SMIC (HS)		SAMM		MEGC2019	
	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1
2D IM-only (SVM)	0.836	0.847	0.667	0.634	0.748	0.703	0.753	0.792
2D IM-only (Softmax)	0.818	0.835	0.659	0.626	0.728	0.687	0.738	0.772
3D IM-only (SVM)	0.825	0.838	0.667	0.659	0.704	0.693	0.731	0.764
3D IM-only (Softmax)	0.814	0.821	0.647	0.634	0.689	0.678	0.720	0.738
AST+SVMNET	0.892	0.913	0.758	0.712	0.778	0.751	0.832	0.807

SMIC(HS) dataset, the complete framework only improved in UAR compared to single-path features, but in terms of UF1-score, the framework with single IDSF performed better. We consider that this difference is due to the lower frame rate in the SMIC(HS) dataset compared to the other datasets. In the limited duration of ME, the changes in ME per unit time are more significant in SMIC(HS), leading to better feature extraction performance for IDSF and 3D-Inception Module. Overall, based on the experimental results, the dual-path feature input for MER is superior to single-path input in our proposed method and SVM classifier has optimal results in the framework where only the classifiers are different.

At the structural aspects, we compare the complete framework without using the FAM, using only the CAM, using only the SAM, and the complete framework on four datasets. On this basis, it is further divided into two categories accord-

ing to the different classifiers. The specific experimental results are shown in Table 10. On the CASME II, SMIC(HS), SAMM, and MEGC2019 datasets, the complete framework demonstrates the best MER performance. The UAR improved by a maximum of 0.169, 0.135, 0.141, and 0.165, and the UF1-score improved by a maximum of 0.199, 0.118, 0.146, and 0.161, respectively. Based on the experimental results, we observe that using either the CAM or the SAM significantly improved MER performance. That is because the channel attention mechanism in CAM weights and recalibrates the different channel's to enhance the representation of temporal information in ME samples, and the spatial attention mechanism in SAM focuses on the changing part of ME sample single frames to enhance the representation of spatial features. The introduced FAM further improved the MER performance on top of the already effective results achieved by individual

TABLE 12. Ablation experiment in classification on four datasets.

Classification	CASME II		SMIC (HS)		SAMM		MEGC2019	
	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1
FC+Softmax	0.867	0.868	0.728	0.693	0.751	0.716	0.763	0.801
AST+SVMNET	0.892	0.913	0.758	0.712	0.778	0.751	0.832	0.807

attention modules, because this module has the advantages of both and completes the missing information in both.

As shown in Table 11, at the network layers part, we compare the complete framework using only 2D Inception Module to extract spatial information, using only the 3D Inception Module to extract temporal information, and the complete framework using both of them. On this basis, it is further divided into two categories according to the different classifiers. On the CASME II, SMIC(HS), SAMM, and MEGC2019 datasets, the complete framework demonstrates the best MER performance. The UAR improved by a maximum of 0.078, 0.111, 0.089, and 0.112, and the UF1-score improved by a maximum of 0.092, 0.086, 0.073, and 0.069, respectively. Comparing the feature extraction methods, spatial information contributes more to MER performance than temporal information. Based on the experimental result, the complete framework simultaneously extracting both spatial and temporal information outperform the framework focusing on a single type of information. That is because 3D convolution captures categorical information about continuous temporal correlations in the semantics of ME samples for filling in facial movement information at the spatial level. We also compare the MER performance of the original framework using FC+Softmax classifier and SVM classifier in different feature extraction methods. The complete framework using SVM classifier achieves the best classification results on all four datasets.

At the classification aspects, we compare the classification performance of the complete framework with different classifiers on four datasets. On the CASME II, SMIC(HS), SAMM, and MEGC2019 datasets, the complete framework with SVM demonstrates the best MER performance. The UAR improved by 0.025, 0.03, 0.027, and 0.069, and the UF1-score improved by 0.045, 0.019, 0.035, and 0.006, respectively. Therefore, it is proved that the validity of our proposed method is established.

V. CONCLUSION

In this paper, we propose a novel decomposition method for MER task: AST+SVMNet. This method utilizes low-parameter parallel network modules to achieve efficient MER by extracting fine-grained spatial-temporal information from two streams of ME features and transferring the features for SVM retraining. Additionally, we introduce a new texture feature ISDF for MER, which mitigates the negative impact of optical flow calculation noise in MER methods that simultaneously introduce ME optical flow features. We apply the ISDF to AST+SVMNet, achieving excellent MER performance. Furthermore, we designed the

FAM for AST+SVMNet, which reduces redundant ME information through nested channel and spatial attention, optimizing the extraction of temporal and spatial representations. Experimental results on four publicly available spontaneous ME datasets demonstrate that the performance of AST+SVMNet is particularly outstanding in the task of fine-grained, lightweight MER when compared with existing representative methods of ME.

It is noteworthy that due to the different sample frame rates of the spontaneous ME datasets, the MER performance of our proposed method varies conspicuously. Therefore, we plan to improve our proposed method by devising an adaptive video frame normalization module with multi-scale low information loss, enabling it to accommodate more diverse and challenging ME datasets. In addition, due to the good generalization ability of SVM in small samples, we employed the SVM classifier to simplify the classification boundary. However, we have only made preliminary applications on this basis and have not identified a kernel function that adapts to the data from different ME datasets. In the future, we plan to discuss in detail the classification effects of different kernel functions in our proposed method on various datasets, to make targeted improvements. Building on the existing method, we aim to determine or design a robust kernel function that adapts to different ME datasets.

REFERENCES

- [1] L. Zhang and O. Arandjelović, "Review of automatic microexpression recognition in the past decade," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 2, pp. 414–434, May 2021, doi: [10.3390/make3020021](https://doi.org/10.3390/make3020021).
- [2] P. Ekman, "Lie catching and microexpressions," in *The Philosophy of Deception*. Oxford, U.K.: Oxford Univ. Press, 2009, pp. 118–136, doi: [10.1093/acprof:oso/9780195327939.003.0008](https://doi.org/10.1093/acprof:oso/9780195327939.003.0008).
- [3] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971, doi: [10.1037/h0030377](https://doi.org/10.1037/h0030377).
- [4] M. Zhang, Q. Fu, Y.-H. Chen, and X. Fu, "Emotional context influences micro-expression recognition," *PLoS ONE*, vol. 9, no. 4, Apr. 2014, Art. no. e95018, doi: [10.1371/journal.pone.0095018](https://doi.org/10.1371/journal.pone.0095018).
- [5] B. Bhushan, "Study of facial micro-expressions in psychology," in *Understanding Facial Expressions in Communication*. New Delhi, India: Springer, 2015, pp. 265–286, doi: [10.1007/978-81-322-1934-7_13](https://doi.org/10.1007/978-81-322-1934-7_13).
- [6] S. Porter and L. ten Brinke, "Reading between the lies," *Psychol. Sci.*, vol. 19, no. 5, pp. 508–514, May 2008, doi: [10.1111/j.1467-9280.2008.02116.x](https://doi.org/10.1111/j.1467-9280.2008.02116.x).
- [7] A. Awad, "Collective framework for fraud detection using behavioral biometrics," in *Information Security Practices*, Cham: Springer International Publishing, 2017, pp. 29–37, doi: [10.1007/978-3-319-48947-6_3](https://doi.org/10.1007/978-3-319-48947-6_3).
- [8] W. Huang, "Elderly depression recognition based on facial micro-expression extraction," *Traitement du Signal*, vol. 38, no. 4, pp. 1123–1130, Aug. 2021, doi: [10.18280/ts.380423](https://doi.org/10.18280/ts.380423).
- [9] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016, doi: [10.1109/TPAMI.2016.2515606](https://doi.org/10.1109/TPAMI.2016.2515606).

- [10] Y. Yang, B. Guo, Z. Wang, M. Li, Z. Yu, and X. Zhou, "BehaveSense: Continuous authentication for security-sensitive mobile apps using behavioral biometrics," *Ad Hoc Netw.*, vol. 84, pp. 9–18, Mar. 2019, doi: 10.1016/j.adhoc.2018.09.015.
- [11] E. A. Haggard and K. S. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of Research in Psychotherapy*. Boston, MA, USA: Springer, 1966, pp. 154–165, doi: 10.1007/978-1-4684-6045-2_14.
- [12] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel: Training laypeople and professionals to recognize fleeting emotions," in *Proc. Annu. Meeting Int. Commun. Assoc. Sheraton New York*, New York, NY, USA, 2009, pp. 1–35.
- [13] M. M. F. Donia, A. A. A. Youssif, and A. Hashad, "Spontaneous facial expression recognition based on histogram of oriented gradients descriptor," *Comput. Inf. Sci.*, vol. 7, no. 3, pp. 31–37, Jul. 2014, doi: 10.5539/cis.v7n3p31.
- [14] S. A. Khan, A. Hussain, and M. Usman, "Reliable facial expression recognition for multi-scale images using Weber local binary image based cosine transform features," *Multimedia Tools Appl.*, vol. 77, no. 1, pp. 1133–1165, Jan. 2018, doi: 10.1007/s11042-016-4324-z.
- [15] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, "A survey: Facial micro-expression recognition," *Multimedia Tools Appl.*, vol. 77, no. 15, pp. 19301–19325, Aug. 2018, doi: 10.1007/s11042-017-5317-2.
- [16] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.
- [17] D. Patel, G. Zhao, and M. Pietikainen, "Spatiotemporal integration of optical flow vectors for micro-expression detection," in *Advanced Concepts for Intelligent Vision Systems*, vol. 9386. Cham, Switzerland: Springer, 2015, pp. 369–380, doi: 10.1007/978-3-319-25903-1_32.
- [18] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7, doi: 10.1109/FG.2013.6553799.
- [19] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e86041, doi: 10.1371/journal.pone.0086041.
- [20] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS(ME): A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 424–436, Oct. 2018, doi: 10.1109/TAFFC.2017.2654440.
- [21] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, and X. Fu, "CAS(ME)3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2782–2800, Mar. 2023, doi: 10.1109/TPAMI.2022.3174895.
- [22] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6, doi: 10.1109/FG.2013.6553717.
- [23] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, Jan. 2018, doi: 10.1109/TAFFC.2016.2573832.
- [24] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "MEGC 2019—The second facial micro-expressions grand challenge," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5, doi: 10.1109/FG.2019.8756611.
- [25] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep learning for micro-expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2028–2046, Oct. 2022, doi: 10.1109/TAFFC.2022.3205170.
- [26] X. Nie, M. A. Takalkar, M. Duan, H. Zhang, and M. Xu, "GEME: Dual-stream multi-task Gender-based micro-expression recognition," *Neurocomputing*, vol. 427, pp. 13–28, Feb. 2021, doi: 10.1016/j.neucom.2020.10.082.
- [27] S. Zhao, H. Tao, Y. Zhang, T. Xu, K. Zhang, Z. Hao, and E. Chen, "A two-stage 3D CNN based learning method for spontaneous micro-expression recognition," *Neurocomputing*, vol. 448, pp. 276–289, Aug. 2021, doi: 10.1016/j.neucom.2021.03.058.
- [28] L. Zhang, X. Hong, O. Arandjelovic, and G. Zhao, "Short and long range relation based spatio-temporal transformer for micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1973–1985, Oct. 2022, doi: 10.1109/TAFFC.2022.3213509.
- [29] H.-Q. Khor, J. See, S.-T. Liong, R. C. W. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 36–40, doi: 10.1109/ICIP.2019.8802965.
- [30] J. Yu, C. Zhang, Y. Song, and W. Cai, "ICE-GAN: Identity-aware and capsule-enhanced GAN with graph-based reasoning for micro-expression recognition and synthesis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8, doi: 10.1109/IJCNN52387.2021.9533988.
- [31] B. Sun, S. Cao, J. He, and L. Yu, "Two-stream attention-aware network for spontaneous micro-expression movement spotting," in *Proc. IEEE 10th Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Oct. 2019, pp. 702–705, doi: 10.1109/ICSESS47205.2019.9040685.
- [32] Z. Wang, M. Yang, Q. Jiao, L. Xu, B. Han, Y. Li, and X. Tan, "Two-level spatio-temporal feature fused two-stream network for micro-expression recognition," *Sensors*, vol. 24, no. 5, p. 1574, Feb. 2024, doi: 10.3390/s24051574.
- [33] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1–7, doi: 10.1145/2964284.2967247.
- [34] X. Zeng, X. Zhao, X. Zhong, and G. Liu, "A survey of micro-expression recognition methods based on LBP, optical flow and deep learning," *Neural Process. Lett.*, vol. 55, no. 5, pp. 5995–6026, Oct. 2023, doi: 10.1007/s11063-022-11123-x.
- [35] X. Ben, X. Jia, R. Yan, X. Zhang, and W. Meng, "Learning effective binary descriptors for micro-expression recognition transferred by macro-information," *Pattern Recognit. Lett.*, vol. 107, pp. 50–58, May 2018, doi: 10.1016/j.patrec.2017.07.010.
- [36] Y. Wang, J. See, R. Raphael, and Y. H. Oh, "LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition," in *Computer Vision—ACCV 2014*. Cham, Switzerland: Springer, 2015, pp. 525–537, doi: 10.1007/978-3-319-16865-4_34.
- [37] X. Huang, S.-J. Wang, G. Zhao, and M. Pietikainen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 1–9, doi: 10.1109/ICCVW.2015.10.
- [38] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikainen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, Jan. 2016, doi: 10.1016/j.neucom.2015.10.096.
- [39] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 115–137, Jan. 2014, doi: 10.1007/s11263-013-0644-x.
- [40] M. Verburg and V. Menkovski, "Micro-expression detection in long videos using optical flow and recurrent neural networks," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–6, doi: 10.1109/FG.2019.8756588.
- [41] S. L. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 394–406, Jul. 2019, doi: 10.1109/TAFFC.2017.2723386.
- [42] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Process., Image Commun.*, vol. 62, pp. 82–92, Mar. 2018, doi: 10.1016/j.image.2017.11.006.
- [43] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 299–310, Oct. 2016, doi: 10.1109/TAFFC.2015.2485205.
- [44] Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "OFF-ApexNet on micro-expression recognition system," *Signal Process., Image Commun.*, vol. 74, pp. 129–139, May 2019, doi: 10.1016/j.image.2019.02.005.
- [45] Q.-S. Jin, H.-C. Xu, K.-H. Liu, S.-T. Liong, Y. S. Gan, and S.-W. Su, "GA-APEXNET: Genetic algorithm in apex frame network for micro-expression recognition system," *J. Phys., Conf. Ser.*, vol. 1544, no. 1, May 2020, Art. no. 012149, doi: 10.1088/1742-6596/1544/1/012149.

- [46] H.-Q. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 667–674.
- [47] D. Y. Choi and B. C. Song, "Facial micro-expression recognition using two-dimensional landmark feature maps," *IEEE Access*, vol. 8, pp. 121549–121563, 2020, doi: [10.1109/ACCESS.2020.3006958](https://doi.org/10.1109/ACCESS.2020.3006958).
- [48] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5, doi: [10.1109/FG.2019.8756567](https://doi.org/10.1109/FG.2019.8756567).
- [49] X. Shu, J. Li, L. Shi, and S. Huang, "RES-CapsNet: An improved capsule network for micro-expression recognition," *Multimedia Syst.*, vol. 29, no. 3, pp. 1593–1601, Jun. 2023, doi: [10.1007/s00530-023-01068-z](https://doi.org/10.1007/s00530-023-01068-z).
- [50] N. V. Quang, J. Chun, and T. Tokuyama, "CapsuleNet for micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–7.
- [51] Y. Song, W. Zhao, T. Chen, S. Li, and J. Li, "Recognizing microexpression as macroexpression by the teacher–student framework network," in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, Oct. 2022, pp. 548–553, doi: [10.1109/ISMAR-Adjunct57072.2022.00115](https://doi.org/10.1109/ISMAR-Adjunct57072.2022.00115).
- [52] L. Lei, T. Chen, S. Li, and J. Li, "Micro-expression recognition based on facial graph representation learning and facial action unit fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1571–1580, doi: [10.1109/CVPRW53098.2021.00173](https://doi.org/10.1109/CVPRW53098.2021.00173).
- [53] H. Zhou, S. Huang, and Y. Xu, "Incept: Micro-expression recognition integrating inception-CBAM and vision transformer," *Multimedia Syst.*, vol. 29, pp. 3863–3876, Aug. 2023, doi: [10.1007/s00530-023-01164-0](https://doi.org/10.1007/s00530-023-01164-0).
- [54] Y. Wang, H. Ma, X. Xing, and Z. Pan, "Eulerian motion based 3DCNN architecture for facial micro-expression recognition," in *MultiMedia Modeling*. Cham, Switzerland: Springer, 2020, pp. 266–277, doi: [10.1007/978-3-030-37731-1_22](https://doi.org/10.1007/978-3-030-37731-1_22).
- [55] B. Chen, Z. Zhang, N. Liu, Y. Tan, X. Liu, and T. Chen, "Spatiotemporal convolutional neural network with convolutional block attention module for micro-expression recognition," *Information*, vol. 11, no. 8, p. 380, Jul. 2020, doi: [10.3390/info11080380](https://doi.org/10.3390/info11080380).
- [56] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108275, doi: [10.1016/j.patcog.2021.108275](https://doi.org/10.1016/j.patcog.2021.108275).
- [57] R. Ruisheng, S. Kai, J. Xiaopeng, and W. Ning, "Micro-expression recognition based on dual attention CrossViT," *Nanjing Xixi Gongcheng Daxue Xuebao*, vol. 15, no. 5, pp. 541–550, 2023.
- [58] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2258–2263, doi: [10.1109/ICPR.2016.7899972](https://doi.org/10.1109/ICPR.2016.7899972).
- [59] Y. Liu, Y. Li, X. Yi, Z. Hu, H. Zhang, and Y. Liu, "Lightweight ViT model for micro-expression recognition enhanced by transfer learning," *Frontiers Neurobotics*, vol. 16, Jun. 2022, Art. no. 922761, doi: [10.3389/fnbot.2022.922761](https://doi.org/10.3389/fnbot.2022.922761).
- [60] B. Sun, S. Cao, D. Li, J. He, and L. Yu, "Dynamic micro-expression recognition using knowledge distillation," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 1037–1043, Apr. 2022, doi: [10.1109/TAFFC.2020.2986962](https://doi.org/10.1109/TAFFC.2020.2986962).
- [61] B. Xia, W. Wang, S. Wang, and E. Chen, "Learning from macro-expression: A micro-expression recognition framework," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2936–2944.
- [62] Z. Li, Y. Zhang, H. Xing, and K.-L. Chan, "Facial micro-expression recognition using double-stream 3D convolutional neural network with domain adaptation," *Sensors*, vol. 23, no. 7, p. 3577, Mar. 2023, doi: [10.3390/s23073577](https://doi.org/10.3390/s23073577).
- [63] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874, doi: [10.1109/CVPR.2014.241](https://doi.org/10.1109/CVPR.2014.241).
- [64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [65] C. Wu and F. Guo, "TSNN: Three-stream combining 2D and 3D convolutional neural network for micro-expression recognition," *IEEE Trans. Electr. Electron. Eng.*, vol. 16, no. 1, pp. 98–107, Jan. 2021, doi: [10.1002/tee.23272](https://doi.org/10.1002/tee.23272).



PEIYUN XUE received the B.S. and Ph.D. degrees in information and communication engineering, and electronic science and technology from Taiyuan University of Technology, Taiyuan, Shanxi, China, in 2013 and 2019, respectively. She is currently a Postdoctoral Researcher with Shanxi Institute of Advanced Innovation. In recent years, she took on and participated in many national and provincial projects. She has participated in many projects, such as the National Natural Fund and Science and Technology Research in Shanxi and also chairs a provincial youth fund. Her research interests include speech signal processing, pathological phonetics, gesture recognition, and artificial intelligence.



XIAOLONG GUO received the B.S. degree from the College of Information and Computer, Taiyuan University of Technology, Taiyuan, Shanxi, China, in 2022. He is currently pursuing the M.S. degree with the College of Electronic Information and Optical Engineering, Taiyuan University of Technology. His research interests include MER, sign language recognition, and computer vision.



JING BAI was born in Taiyuan, Shanxi, China. She received the bachelor's degree in electronic information engineering, the master's degree of information and signal processing, and the Ph.D. degree in circuits and systems from the College of Information Engineering, Taiyuan University of Technology, Shanxi, in 1985, 2004, and 2010, respectively. She has been a Teacher, a Professor, the Master Supervisor, and the Director of the Experiment Technology Center, College of Electronic Information and Optical Engineering, Taiyuan University of Technology. She edited a textbook named *Digital Signal and Logical Design* in 2009. Her research interests include digital signal processing and data mining. In 2011, she received the Teaching Achievements Second Prize of Shanxi Province, and the Science and Technology Progress Second Prize of Shanxi Province, in 2012.



BO YUAN received the B.S. degree from the College of Information Engineering, Zhengzhou University, in 2020, and the M.S. degree from the College of Information and Computer Science, Taiyuan University of Technology, in 2023. He is currently with Northern Institute of Automatic Control Technology. His research interests include computer vision and MER.