Computing and Informatics, Vol. 43, 2024, 458–481, doi: 10.31577/cai_2024_2_458

FESNET: SPOTTING FACIAL EXPRESSIONS USING LOCAL SPATIAL DISCREPANCY AND MULTI-SCALE TEMPORAL AGGREGATION

Bohao Zhang, Jiale Lu, Changbo Wang, Gaoqi He*

Chongqing Key Laboratory of Precision Optics
Chongqing Institute of East China Normal University
Chongqing 401120, China
&
School of Computer Science and Technology
East China Normal University
Shanghai 200333, China
e-mail: {52215901021, 51215901082@}stu.ecnu.edu.cn,
cbwang@dase.ecnu.edu.cn, gqhe@cs.ecnu.edu.cn

Abstract. Facial expressions (FEs) spotting aims to split long videos into intervals of neutral expression, macro-expression, or micro-expression. Recent works mainly focus on feature descriptor or optical flow methods, suffering from difficulty capturing subtle facial motion and efficient temporal aggregation. This paper proposes a novel end-to-end network, named FESNet (**F**acial **E**xpression **S**potting **Net**work), to solve the above challenges. The main idea is to model the subtle facial motion as local spatial discrepancy and incorporate temporal correlation by multi-scale temporal convolution. The FESNet comprises a local spatial discrepancy module (LSDM) and a multi-scale temporal aggregation module (MTAM). The LSDM first extracts the static spatial features from each frame by residual convolution and learns the inner spatial correlation by multi-head attention. Moreover, the subtle facial motion of facial expression is modeled as the discrepancy between the first frame and the current frame of the input interval, making frame-wise spatial proposals. Using the local spatial discrepancy features and proposals as input, the MTAM incorporates the temporal correlation by multi-scale temporal convolution and performs cascade refinement to make the final prediction. Furthermore, this paper proposes a smooth loss to ensure the temporal consistency of the cascade refined proposals from MTAM. Comprehensive experiments show that FESNet achieves competitive performance compared to state-of-the-art methods.

* Corresponding author

Keywords: Facial expression analysis, micro-expression spotting, video understanding, convolutional neural networks

1 INTRODUCTION

Facial expression recognition plays a vital role in human-computer interaction and security, and thus, there are many related works [1]. Most of these works have one assumption that FEs frames or intervals have been segmented from the original long videos. However, splitting a long video into segments of different facial features is a challenging task called FEs spotting. One reason is that FEs spotting takes unprocessed long videos as input. It is hard to capture the accurate boundary frame between different expression content. Moreover, FEs spotting is a multi-classification task. Frames in a specific interval are mapped to neutral, macro-expression (MaE), or micro-expression (ME).

As a proverb goes, 'The face is no index to the heart.' Gaining real insight into others' states of mind from facial expressions is unreliable, as macro-expressions can be disguised [2]. Therefore, involuntary micro-expressions become crucial cues to reveal the genuine states of mind that others try to conceal in high-risk tasks such as negotiation, criminal investigation, and national security. MaE and ME generally coexist with a lot of neutral frames in long videos, and it is necessary for facial expression analysis to first segment macro-expression and micro-expression intervals from long videos. However, as shown in Figure 1, with short duration and low intensity, ME is highly similar to neutral expressions. It is almost impossible to distinguish ME from neutral expressions with a single frame. Furthermore, spotting ME in long videos with the naked eye is still challenging and labor-intensive, even for experts after intense training. To segment MaE and ME intervals from long videos, researchers have explored kinds of feature extraction methods, including feature descriptor-based methods, optical flow-based methods, and deep learning-based methods [2].

Early facial expression spotting methods primarily focused on optical flow or other feature descriptors, relying heavily on the experience and observation of facial expression videos. Deep learning-based methods generally can automatically extract high-level implicit features invisible to the naked eye. Therefore, recent research utilizes deep learning methods to capture subtle facial motion. However, numerous researchers have found that optical flow features are more suitable than the original frames for deep learning models in achieving significant improvements. This is probably because the orientation and amplitude of optical flow can explicitly model the muscle movements [3, 4] associated with facial expressions. However, optical flow contains noise caused by head movements, losing the details of the original static distribution, as illustrated in Figure 1. Furthermore, nearly all current mainstream methods are region-based [2, 5], by dividing the original frames into regions or selecting specific regions based on facial landmarks to extract features.

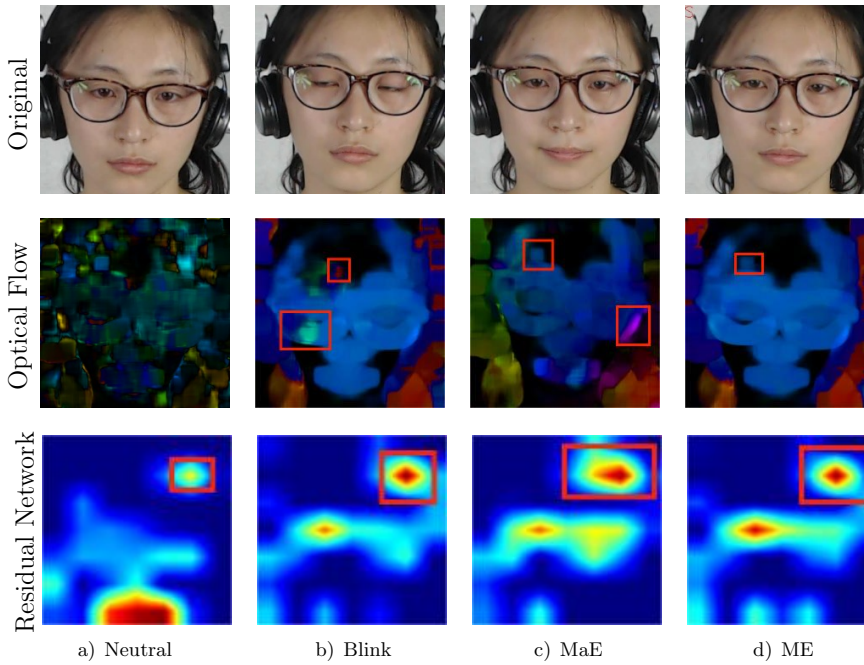


Figure 1. The original frames and their corresponding optical flow and latent features are analyzed. The optical flow contains noise, which can easily be confused with subtle facial movements. The latent features are focused on areas with significant differences, such as reflections on the forehead and blinking of the eyes.

The current state-of-the-art facial expression spotting methods precisely compute the optical flow of specific regions as input for the neural network [6]. Extracting features from carefully selected facial regions alleviates the interference from head movements and noise, achieving higher facial expression spotting accuracy. This strategy is effective because facial expressions, especially micro-expressions, are local facial motions; however, it does not account for the correlation between regions.

This paper addresses the facial expression (FE) spotting task by proposing an end-to-end network that comprises a **L**ocal **S**patial **D**iscrepancy **M**odule (LSDM) and a **M**ulti-Scale **T**emporal **A**ggregation **M**odule (MTAM). The LSDM employs residual convolution to extract static spatial features at the frame level and learns the inner spatial correlations using multi-head attention [7]. It then models the subtle facial motions as discrepancies between the first frame and the current frame of the input interval, generating spatial proposals based on these discrepancies. The MTAM utilizes local spatial discrepancy features and spatial proposals as input, incorporating temporal correlations to model facial expressions' temporal motion patterns and make the final prediction. As illustrated in Figure 2, ME and MaE

exhibit fixed temporal motion patterns. Furthermore, it has been observed that the first and last frames of facial expression (FE) intervals are invariably similar to neutral frames, as discussed in [8]. Assuming the sliding window is positioned near the first frame of an FE interval, the aggregated temporal features from the MTAM will align with the corresponding temporal motion pattern of either MaE or ME, leading to an accurate prediction. Conversely, suppose the sliding window is positioned far from the first frame. In that case, a mismatch between the temporal feature and its corresponding temporal motion pattern will result in the accurate rejection of negative samples.

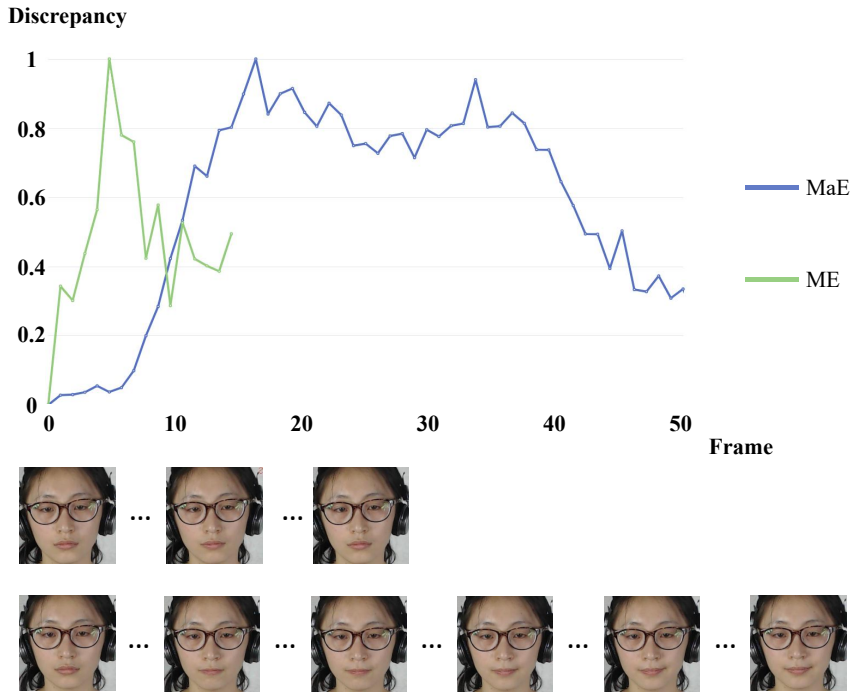


Figure 2. The variation of spatial discrepancy during MaE and ME from the same individuals in CAS(ME)^2. The estimation of spatial discrepancy is based on the similarity between the spatial static features of the first frame and the current frame. It can be seen that there is a significant difference between the temporal patterns of MaE and ME.

The LSDM employs residual convolution to compact spatial features towards the intersecting facial region of ME and MaE. FEs comprise multiple facial motions; however, their inner spatial correlations are complex and defy explicit modeling. Consequently, multi-head attention is utilized to learn the inner spatial correlations from single frame-level features, yielding attention-weighted static spatial features. In order to model the spatial discrepancy alongside obtaining static spatial features,

the difference between the static features of the first frame and the current frame of the input interval is calculated. The resultant discrepancy feature, being sparse, is upsampled and integrated with the original static spatial feature to derive a compact local spatial discrepancy feature, which serves as the basis for making spatial proposals.

The MTAM employs multi-scale temporal convolution to aggregate temporal correlations and model the motion patterns of FEs. By integrating spatial and temporal proposals as input, MTAM models the temporal motion patterns and makes the final prediction. Based on a pyramid structure, a temporal search strategy is adopted to learn the local temporal patterns from local spatial discrepancy features. MTAM aggregates local spatial discrepancy features across multi-scale temporal ranges by varying the temporal receptive field. The aggregated features represent the temporal motion patterns within the input interval and are utilized to generate temporal proposals. Since neutral frames may suppress FEs features during aggregation, MTAM incorporates spatial proposals to mitigate this effect and make the final prediction.

Overall, the main contributions of this paper are summarized as follows:

- A novel end-to-end network for FEs spotting from long videos. The proposed FESNet efficiently extracts spatial discrepancy features and performs multi-scale temporal aggregation for FEs spotting. Moreover, FESNet surpasses state-of-the-art methods, offering a novel solution for FEs spotting.
- An efficient LSDM is designed to extract FEs-related spatial discrepancy features using residual convolution and multi-head attention. Besides, the LSDM is constrained by a temporal local diversity loss to model the subtle facial movement.
- A multi-scale MTAM is proposed to incorporate multi-scale temporal correlations and model the temporal motion patterns of FEs. MTAM employs multi-scale temporal convolution and cascade refinement processes to model the temporal motion patterns of FEs accurately and make the final prediction.

2 RELATED WORK

2.1 Feature Descriptor-Based Methods

Early FEs spotting methods mainly focus on designing effective feature descriptors to model the subtle facial motions of FEs. Polikovsky et al. [9] first divided the face into specific regions and then utilized a 3D-Gradient orientation histogram descriptor to recognize the motion in the corresponding areas. They also considered the K-Means algorithm and estimated specific ME characteristics in psychological analysis [10]. Li et al. [11] first proposed to utilize the feature difference contrast to spotting ME from long videos. They divided the face into 6×6 blocks based

on three tracked facial feature points and then extracted LBP [12] and HOOF [13] features to analyze the feature difference. All the above three works aim to detect the onset, apex, and offset frames of ME from long videos. Yan et al. [14] used the constraint local model to track feature points of the face and employed the LBP feature from ROIs to measure the difference between frames and locate the apex frame of ME. Similarly, Liong et al. [15] also utilized LBP from RoIs but combined LBP with a binary search method to spot apex from long videos. Esmaeili and Shahdi [16] proposed a new LBP-based feature descriptor named Cubic-LBP, which computes LBP on fifteen introduced planes to grab the most vital information and detect apex frame. Davison et al. [17] calculated the chi-square distance of a 3D histogram of oriented gradients and detected the apex frame automatically.

Diverging from the above methods focused on spotting key frames of ME, Davison et al. [18] proposed a method based on the histogram of oriented gradients to spotting ME intervals from long videos directly. Li et al. [19] introduced LTP-ML, applying temporal PCA to windows of specific facial regions for detecting facial movements associated with ME. Molianen et al. [20] initially divided the face into block structures and calculated the dissimilarity between blocks of different frames using chi-squared distance. Threshold-based peak detection was employed for spotting ME intervals in videos. Xia et al. [21] utilized the geometric deformation of facial regions as a feature descriptor and proposed a probabilistic framework for FEs spotting. Considering the computational time for ME spotting, Soh et al. [22] proposed a many-core parallel LBP-TOP [23] algorithm to leveraging compute unified device architecture. Zhao et al. [24] leveraged improved face alignment methods, more robust optical flow techniques, and superior facial landmark detectors. They employed a Bayesian optimization hybrid approach for optimizing parameters typically set manually.

2.2 Optical Flow-Based Methods

Optical flow, a well-established motion estimation method, is widely studied for its capability to model the subtle facial motions characteristic of ME. Shreve et al. [25] introduced optical strain, derived from robust optical flow, to analyze FEs. They further leveraged optical flow to calculate skin strain from non-rigid facial motions, marking an initial attempt to detect both MaE and ME [26]. Patel et al. [27] extracted the optical flow from local spatial regions and utilized the direction continuity to spot the onset and offset frames. To address the problem that optical flow may be affected by noise, such as head motion, Liong et al. [28] extracted pixel-wise optical strain magnitudes, generating a feature histogram for ME recognition. Considering the magnitude and angle of optical flow, Guo et al. [29] proposed a novel decision criterion focusing on the four most discriminative facial regions for ME spotting. Similarly, Yan et al. [3] concentrate on the direction of optical flow, applying a robust method on ROIs to extract the Main Directional Mean Optical-flow (MDMO) feature. Xu et al. [4] introduced the Facial Dynam-

ics Map (FDM) to accurately model subtle facial motions for estimating optical flow between frames. Additionally, Shreve et al. [30] utilized optical strains to represent non-rigid facial motions and visualize ME progression over time. Zhang et al. [31], accounting for head movement, proposed a method to separate FE-related local movements from the global optical flow field, constructing optical flow sequences as spatial-temporal features for identifying FE intervals from extracted SP-patterns.

2.3 Deep Learning Methods

Compared with traditional methods, deep learning-based FEs spotting methods face significant challenges due to the limited available public datasets. Kim et al. [32] pioneered using CNN to extract spatial features of MEs, leveraging expression states through objective functions. To elucidate the temporal relations of ME frames, the extracted spatial features were integrated into Long Short-Term Memory (LSTM) networks to derive temporal features. ELRCN [33], adopting a similar strategy to Kim et al., innovated by training CNN and LSTM jointly to ensure ME features' internal consistency and to decrease computational time. Nag et al. [34] introduced a joint network to extract discriminative temporal features, distinguishing MEs from rapid muscle movements. Wang et al. [35] introduced MESNet, which employs 2D convolution for spatial feature extraction and 1D convolution for modeling temporal relations. Yap et al. [36] introduced 3D Convolutional Neural Networks (3D-CNNs) for the simultaneous extraction of spatial features and analysis of temporal relations. ABPN [6] initially calculates optical flow using a video encoding module to mitigate noise impact, extracting temporal features through 1D convolution. The 1D convolution within ABPN's PEM module infers frame-level auxiliary probabilities, contributing to the final prediction. Xie et al. [37] proposed AEM-Net, which extracts features at various depths and identifies discriminative ME intervals through an attention module.

STCAN [38] accounted for the inconsistency in duration and estimated the weight of frame-level spatial features in the temporal domain for spotting MaE and ME intervals within video sequences. Yang et al. [39] proposed using facial action units (AUs) for MaE and ME spotting. They introduced the Concat-CNN model to discern the relationships between AUs across distinct frames. LSSNet [40] leveraged the I3D [41] model to extract optical flow-related spatial features of a fixed length, making the final prediction while suppressing location. Liong et al. [42] approached ME spotting as a regression challenge, employing pseudo-labeling to enhance learning. The proposed SOFTNet utilized optical flow from specific facial regions, aggregating scores to detect the apex frame. DynGeoNet [43] aimed to enhance ME spotting performance through hybrid feature engineering, extracting robust features from geometric and appearance aspects.

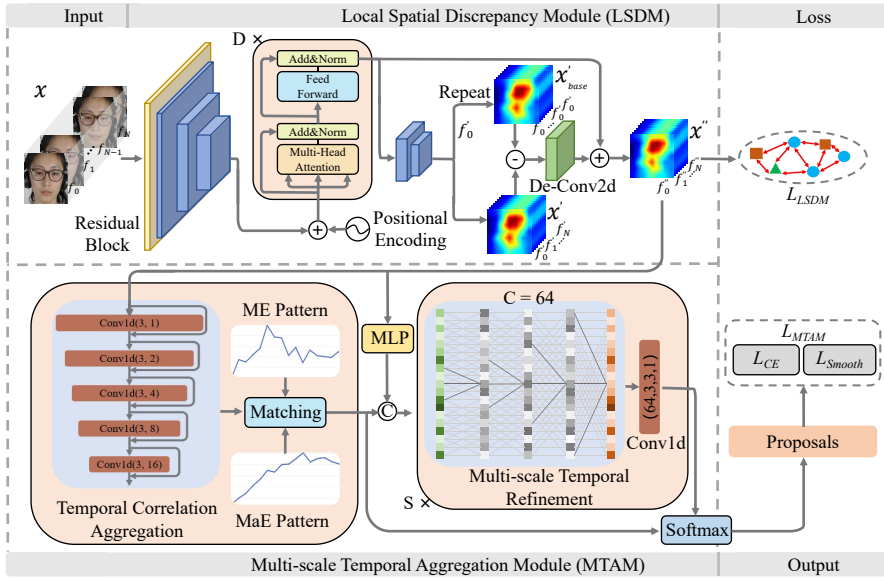


Figure 3. Illustration of our pipeline. The LSDM extracts the static spatial independent features x' and calculates the difference between f'_0 and f'_i to obtain x''. The loss function $\mathcal{L}_{LSDM}$ is employed to guide the LSDM towards extracting distinctive features. Subsequently, the MTAM integrates the temporal discrepancy, refining proposals from previous layers and the LSDM through multi-scale temporal convolution. The dilation parameters of the multi-scale refinement process are gradually increased to maintain local consistency and prevent overfitting in global prediction.

3 METHODOLOGY

3.1 Overview

This paper focuses on the task of FEs spotting, and the overall pipeline of the proposed method is shown in Figure 3. Specifically, the FESNet is composed of LSDM and MTAM. The LSDM employs residual convolution to circumvent feature vanishing, modeling the subtle facial motions characteristic of ME as discrepancies in spatial static features and generating spatial proposals. The MTAM aggregates temporal correlations to align with ME patterns for temporal proposal generation. Significantly, MTAM processes spatial proposals and merges them with temporal proposals, executing final predictions through multi-scale temporal refinement.

3.2 Preprocessing

FEs videos usually contain background information, and the frame rate differs. Therefore, we sample all original videos in SAMM [44] into 60 FPS (frames per

second) while maintaining the original frame rate of CAS(ME)^2 [45]. Specifically, for the raw video, we first sample it to 60 FPS, then divide it into intervals of 32 frames according to the duration constraints in the definition of micro-expressions. We then use these segments as inputs to FESNet. To avoid the interference of background information and unrelated motion, we crop the face from the original frames and make face alignment based on the face detection result of MTCNN [46]. We calculate a square box instead of a rectangle to crop the faces. The purpose of using a square box is to avoid unnecessary interference, which may be caused by the deformation of facial details during the resizing process before the cropped face is fed to the network. The preprocessed frames are resized to 112×112 pixels as shown in Figure 5 a).

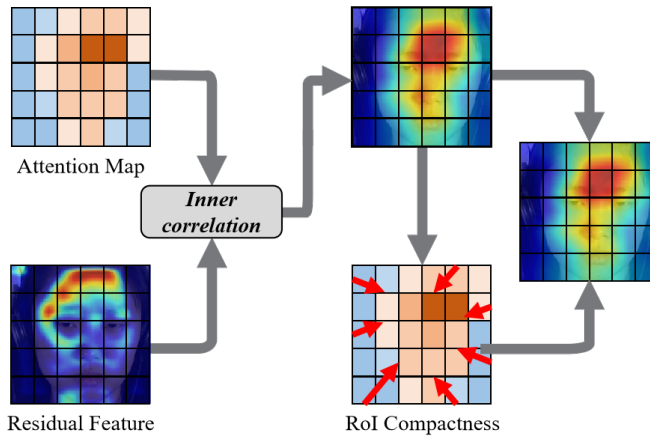


Figure 4. Illustration of the proposed LSDM. Residual blocks obtain the residual feature and will be fed into a multi-head attention module to model the inner correlation. With a loss constraint, the facial motion modeling process will compact the residual map towards the intersection of ME and MaE.

3.3 Local Spatial Discrepancy Module

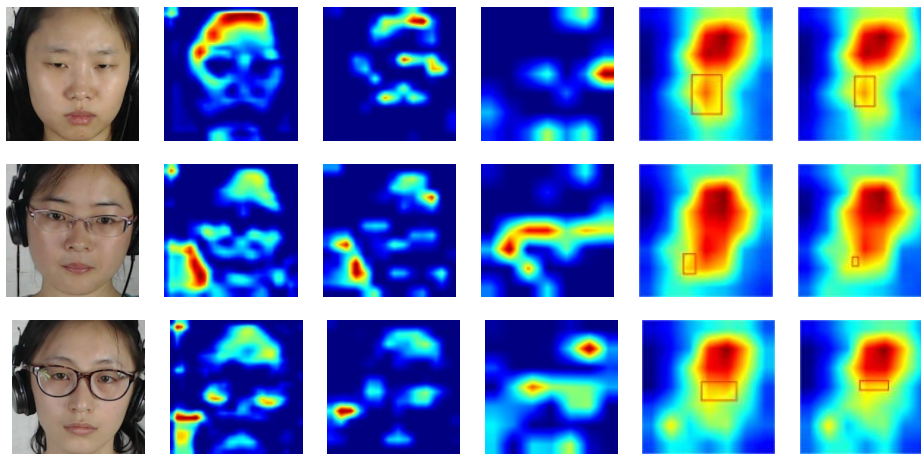
As shown in Figure 5, although residual blocks can extract features from the eyes and mouth, these layers have a low activation response to these regions. To this end, we utilize a multi-head mechanism to extract the static spatial independent features f_i' from the i^{th} frame f_i of the input interval. It is worth noticing that f_i' is not yet spatial motion features since f_i' is extracted from a single frame. The spatial motion features f_i'' of f_i is modeled by the discrepancy between f_i' and f_0'. After rearranging all f_i'' along the temporal dimension, we obtain the local spatial motion features x'' of the original x. The main idea of the proposed LSDM is shown in Figure 4. An attention map models the inner correlation of the input RGB frames and obtains a refined feature map that can describe FEs. However, MaE-related features will

dominate the latent space, easily understood by facial intensity. Therefore, we make the features compact by modeling FEs as the spatial discrepancy and adding loss constraints, making the model focus on the intersection of the MaE and ME regions. Such an operation aims to ensure that FEs (mainly ME) can be distinguished from neutral FEs, and MTAM will distinguish MaE and ME by the temporal motion pattern.

As shown in Figure 3, for each RGB image from the given interval x, we first perform a single convolutional layer and three residual blocks to obtain the unweighted residual spatial features X_{init}. To exclude redundant features in X_{init} and enhance FEs-related features, we construct a multi-head attention module of depth D with the number of heads H. D and H are set to 4 and 8, respectively. With the attention-weighted spatial features, X_{Atten}, we first perform two residual blocks to obtain the deeper features x', and then repeat f_0' for N times to obtain x'_{base}:

$$x'' = ReLU\left(BN\left(DeConv\left(x' - x'_{base}\right)\right) + X_{Atten}\right), \tag{1}$$

where $DeConv(\cdot)$ denotes deconvolution, $ReLU(\cdot)$ and $BN(\cdot)$ denote activation and normalization, respectively.



a) Input b) Res-Block 1 c) Res-Block 2 d) Res-Block 3 e) x' f) x''

Figure 5. The original input frames of the proposed FESNet and feature maps from different layers of LSDM. It can be seen from b), c), and d) that feature maps of residual blocks cannot focus on crucial facial regions such as eyebrows and mouth. As for e) and f), these feature maps are feature maps before and after calculating spatial discrepancy. It can be seen that the salient regions (regions in red box) in x'' are much smaller than that in x'. We explain that the spatial discrepancy calculation excludes redundant facial features unrelated to FEs, making the network focus on specific areas and effectively improving the accuracy of FEs spotting.

3.4 Multi-Scale Temporal Aggregation Module

Although the LSDM can capture subtle facial motion, the process is temporal independent and thus cannot be directly used for FEs spotting. To efficiently spot ME and MaE intervals based on the differences in temporal motion pattern, we construct the multi-scale temporal aggregation module, extracting multi-scale temporal correlation from x'' and making final prediction based on spatial and temporal proposals and multi-scale refinement.

As shown in Figure 3, the MTAM first aggregates the temporal correlation by five 1D convolutions with the same kernel size but different temporal receptive fields (the numbers in parentheses). The aggregated temporal correlation can be seen as the temporal pattern of the input interval. It will match the ME pattern in latent space, resulting in temporal proposals. To avoid the repression of FEs frames caused by a large number of neutral frames in temporal proposals, the MTAM will also consider spatial proposals and make the final prediction by multi-scale temporal refinement. The refinement includes S basic blocks composed of 1D convolutional layers with different dilation operations, where S is set to 2 in our experiment, respectively. These basic blocks can be described as:

$$X_l^s = \begin{cases} W_l^s * softmax\left(X_4^{s-1}\right) + b_l^s, & l = 0, \\ ReLU\left(BN\left(W_l^s * X_{l-1}^s + b_l^s\right)\right), & \text{else}, \end{cases} \tag{2}$$

where W_l^s and b_l^s are the weights and bias vectors of 1D convolution, $l \in [0, 4]$, $s \in [0, S)$, the dilation parameter of l^{th} layers in basic blocks is 2^l. With different dilation operations, layers in basic blocks have different temporal local receptive fields, and the max dilation parameters match the duration of ME with a frame rate of 32. The gradual increase of the dilation parameters is to refine the proposals of previous layers, and the smaller dilation parameters ensure local consistency, alleviating the over-segmentation. Meanwhile, the larger dilation parameters ensure that the receptive fields are wide enough to obtain the global pattern of the input interval. Besides, each basic block refines the previous basic block's output, given input $x = \{f_0, f_1, \dots f_n\}$, the l^{th} 1D convolutional layer in S^{th} basic block of MTAM will make frame-level proposal $p^{S|l}$ restored in $P_{MTAM} = \{p^{0|0}, \dots p^{0|l}, \dots, p^{S|l}\}$.

3.5 Loss Function

To make sure the LSDM can extract more discriminative spatial features and the MTAM can make a more accurate and smoother prediction, we introduce $\mathcal{L}_{LSDM}$ and $\mathcal{L}_{MTAM}$ in our method. For $\mathcal{L}_{LSDM}$, we calculate:

$$\mathcal{L}_{LSDM} = -\frac{1}{\log\sqrt{\sum\left(x'' - \mathcal{M}(x'')\right)}}, \tag{3}$$

where $\mathcal{M}(\cdot)$ denotes the mean operation. The main purpose of $\mathcal{L}_{LSDM}$ is to discard redundant features to capture the subtle facial motions and reduce the deviation within spatial static features, which is a trade-off. Irrelevant factors are more significant than the subtle motion of ME, which will increase the deviation within the spatial static features. This will cause a mix of ME and neutral FEs in the latent space, leading to the failure of FEs spotting. We tried other solutions to make samples away from each other, but the scarcity and specificity of ME samples make deep learning-based methods not applicable to solve this problem. Besides, a more complex form of the $\mathcal{L}_{LSDM}$ will lead to an unstable training process on some extreme FEs samples.

For $\mathcal{L}_{MTAM}$, we use a weighted cross entropy loss $\mathcal{L}_{CE}$ and a mean square error based smooth loss $\mathcal{L}_{smooth}$. Given labels $Y = \{y_0, y_1, \dots y_n\}$ and final output $O = \{o_0, o_1, \dots o_n\}$, $\mathcal{L}_{MTAM}$ can be expressed as:

$$\begin{aligned} \mathcal{L}_{MTAM} &= \mathcal{L}_{CE} + \gamma \mathcal{L}_{smooth}, \\ \mathcal{L}_{CE} &= -\sum_{n} \sum_{c=1}^{3} w_c y_n \log\left(o_n\right), \\ \mathcal{L}_{smooth} &= \frac{1}{S} \sum_{i}^{S} \sum_{j}^{l} \left(p_{[1:n]}^{i|j} - p_{[0:n-1]}^{i|j}\right)^2, \end{aligned} \tag{4}$$

where w_c is the inverse of the proportion of neutral, ME, and MaE in the corresponding dataset, γ is the weight of L_{smooth} and is set to 0.15 empirically according to our experiments, as shown in Table 4. The final loss function is:

$$\mathcal{L} = \mathcal{L}_{LSDM} + \mathcal{L}_{MTAM}. \tag{5}$$

4 EXPERIMENTS

4.1 Datasets

To evaluate the proposed method, we conducted a five-fold cross-validation on CAS(ME)^2 and SAMM. CAS(ME)^2 contains 98 long videos with a frame rate of 30 FPS, including 57 ME and 300 MaE samples. SAMM contains 224 long videos with a frame rate of 200 FPS, including 159 ME samples and 343 MaE samples. In addition, we also sampled the videos in SAMM to alleviate the problem of significant frame rate difference, and the frame rate of the sampled videos was 60 FPS.

4.2 Evaluation Metrics

In this paper, we report three evaluation metrics to prove the performance of the proposed FESNet on the task of FEs spotting, including frame-wise accuracy *Acc*, F1-score with IoU of 0.5 [47] and edit score *Edit*. Both *Acc* and *Edit* are calculated

similarly as [48]. Given the input interval $x = \{f_0, f_1, \cdots, f_n\}$ of n frames and corresponding labels $Y = \{y_0, y_1, \cdots, y_n\}$, FESNet will make the final output $O = \{o_0, o_1, \cdots, o_n\}$, the frame-wise accuracy Acc can be formulated as:

$$Acc = \frac{O \cap Y}{O \cup Y - O \cap Y}. \tag{6}$$

The F1-score of FEs spotting is different from that of traditional classify tasks. The true positive (TP) is determined based on the Intersection over Union (IoU) of Y and O:

$$TP_{O,Y} = \begin{cases} 1, & \frac{O \cap Y}{O \cup Y} \geq T_{\text{threshold}}, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $T_{\text{threshold}}$ is set to 0.5 in our experiment. Given input of R_n FEs intervals and prediction of P_n FEs intervals, the F1-score of FEs spotting is:

$$\begin{aligned} \text{Recall} &= \frac{TP}{R_n}, \\ \text{Precision} &= \frac{TP}{P_n}, \\ \text{F1-score} &= \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} = \frac{2T}{R_N + P_N}. \end{aligned} \tag{8}$$

In addition to Acc and F1-score, we also report edit score $Edit$ to compare the difference between FEs labels and the prediction of FESNet. $Edit$ is a quantitative measure of the difference between two strings and can be formulated as:

$$Edit = 100 \times \left(1 - \frac{\text{lev}_{GT,Pred}(i,j)}{n}\right), \tag{9}$$

$$\text{lev}_{Y,O}(i,j) = \begin{cases} \max(i,j), & \text{if } \min(i,j) = 0, \\ \min_{Y,O}(i,j), & \text{otherwise}, \end{cases} \tag{10}$$

$$\min_{Y,O}(i,j) = \min \begin{cases} \text{lev}_{a,b}(i-1,j) + 1, \\ \text{lev}_{a,b}(i,j-1) + 1, \\ \text{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)}. \end{cases} \tag{11}$$

4.3 Results and Analysis

To the best of our knowledge, we are the first to report not only the F1-score but also the edit score and frame-wise accuracy to further investigate the proposed method's

Methods	CAS(ME)^2			SAMM		
	$F1_{MaE}$	$F1_{ME}$	$F1_{Overall}$	$F1_{MaE}$	$F1_{ME}$	$F1_{Overall}$
Li et al. [19] (2019)	–	1.79	–	–	3.16	–
He et al. [49] (2020)	11.96	0.82	3.76	6.29	3.64	4.45
Zhang et al. [31] (2020)	21.31	5.47	14.03	7.25	13.31	9.99
MESNet [35] (2021)	–	3.60	–	–	8.80	–
DynGeoNet [43] (2021)	–	5.01	–	–	9.74	–
STCAN [38] (2021)	12.50	2.50	11.68	14.69	1.25	12.57
Concat-CNN [39] (2021)	25.05	1.53	20.19	35.53	11.55	27.36
SOFTNet [42] (2021)	24.10	11.73	20.22	21.69	15.20	18.81
LSSNet [40] (2021)	38.0	6.3	32.7	33.6	21.8	29.0
3D CNN [36] (2022)	21.45	7.14	16.75	15.95	4.66	10.84
ABPN [6] (2022)	33.57	15.90	31.17	33.49	16.89	29.08
Our Method	**41.34**	**17.90**	**34.29**	**44.24**	**22.46**	**33.20**

1. The best results are marked in bold font.
2. '–' indicates that the corresponding metric is not reported in the original literature.

Table 1. Comparison with the state-of-the-art methods on CAS(ME)^2 and SAMM

performance in FEs spotting. The proposed FESNet and state-of-the-art methods' quantitative comparison is shown in Figure 1.

It can be seen that the F1-score of early FEs spotting methods needs to be more satisfactory, and even some deep learning-based methods fail on ME spotting. The failure of early FEs spotting methods is that these methods mainly focus on feature descriptors to explicitly extract the motion patterns of ME. Besides, most early FEs spotting methods rely heavily on empirically selected thresholds. Some deep learning-based methods fail to spot FEs because the distribution of samples in ME datasets is unbalanced. A few deep learning-based methods converge to neutral FEs, failing MaE and ME spotting. Most deep learning-based methods can spot MaE intervals but fail to distinguish ME from noise and neutral FEs.

The quantitative experiment shows that our method significantly outperforms the existing methods, especially in the F1-score of MaE spotting, which is improved by 8.7 % to 23 % on CAS(ME)^2 and 31.7 % to 32 % on SAMM. Moreover, the spotting performance of ME is improved by 12.6 % on CAS(ME)^2 and 33 % on SAMM. It can be observed that both *Acc* and *Edit* drop significantly on SAMM compared to that on CAS(ME)^2, while the F1-score of ME and MaE is relatively high. This is because SAMM's high-speed camera captures lots of MaE with low intensity. Although we have sampled the raw sequences, many low-intensity MaE are still easily confusing with ME. Besides, the raw videos in SAMM are much longer than that in CAS(ME)^2, which means there are more negative samples in SAMM, resulting in the low score of *Edit* and *Acc*.

In order to evaluate the performance of the proposed FESNet more comprehensively and avoid data coincidences, a five-fold cross-validation experiment was conducted on CAS(ME)^2 and SAMM. We found that dataset partitioning significantly impacts FEs spotting and may even directly determine the overall performance of

	CAS(ME)^2					SAMM				
	$F1_{MaE}$	$F1_{ME}$	$F1_{Overall}$	*Edit*	*Acc*	$F1_{MaE}$	$F1_{ME}$	$F1_{Overall}$	*Edit*	*Acc*
Fold 1	40.04	15.87	34.08	**88.50**	90.53	**55.49**	18.26	**42.34**	52.72	59.80
Fold 2	45.30	13.81	**39.89**	87.33	**91.12**	40.82	21.19	26.52	42.91	50.51
Fold 3	34.99	22.85	25.58	74.72	82.31	45.03	26.31	35.32	44.16	49.57
Fold 4	**45.82**	13.83	38.05	78.47	84.72	47.35	**29.60**	35.75	**65.98**	**71.08**
Fold 5	40.54	**23.11**	33.85	70.71	76.66	32.48	16.96	26.05	57.75	63.14
Overall	41.34	17.90	34.29	79.95	85.07	44.24	22.46	33.20	52.71	58.82

1. The best result of each metric on each dataset is marked in bold font.

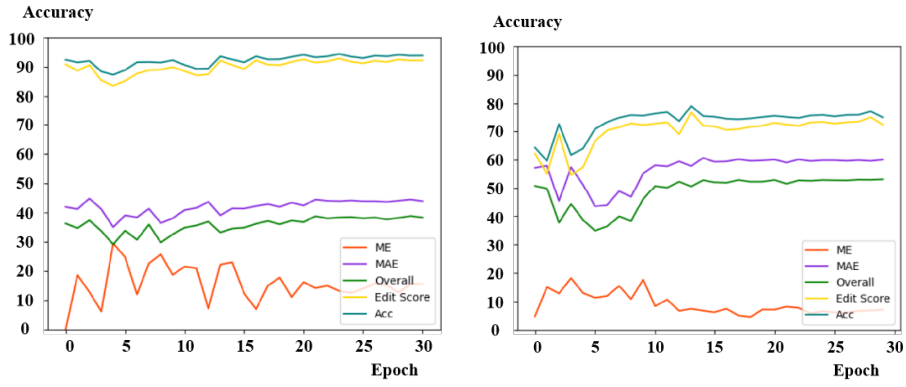
Table 2. Five-fold cross validation of proposed methods on CAS(ME)^2 and SAMM

the corresponding method from the numerical level. The quantitative comparison of the five reported metrics during the five-fold cross-validation experiment is shown in Figure 2. It can be observed that all five reported metrics vary significantly with different partitions of the dataset. We believe this is mainly caused by the distribution of FEs intervals in different long videos. Since multiple FEs intervals exist in a single original video, the dataset is divided according to the original video to avoid sample duplication.

Furthermore, this may lead to differences in the proportion of actual samples in the training and validation sets among different partitions. Moreover, even from the same individuals, there is a significant difference between hard and easy samples, and thus, we cannot ensure that these two kinds of samples are always balanced. This also proves the necessity to conduct cross-validation in FEs spotting tasks. Besides, the *Edit* and *Acc* on SAMM are relatively low. We believe that the higher frame rate of SAMM causes this. Although the high-speed camera SAMM uses can capture the fine-grained facial motions of FEs, the recorded frames will contain more FEs frames, which are much more similar to neutral frames. Besides, more negative samples are in SAMM, resulting in degraded FEs spotting performance.

In addition, we found that there seems to be a game relationship among the five indicators, as shown in Figure 6. We believe that the similarity between ME and neutral frames causes this. When network weights are too biased towards MaE, low-intensity ME is more likely to be classified as a neutral frame. On the contrary, when the network weights are too biased toward ME, MaE with relatively low intensity and neutral frames with local facial motion will be classified as ME. Similarly, the classification weight of $\mathcal{L}_{\mathcal{CE}}$ also significantly impacts the performance of FEs spotting. This is mainly due to the unbalanced distribution of neutral and FEs frames in long videos. Moreover, ME with high intensity and MaE with low intensity are also easily confused. Hence, the performance of FEs spotting is sensitive to the weight of the loss function $\mathcal{L}_{\mathcal{CE}}$.

Figure 8 shows the results of FESNet on the FEs spotting task, where it can be seen that the ME and MaE intervals in ground truth are segmented into many short intervals separated by neutral frames in the proposal of LSDM. The multi-



a) CAS(ME)^2 b) SAMM

Figure 6. Experimental records of the five metrics reported in this paper on CAS(ME)^2 and SAMM

scale refinement of MTAM will aggregate the discrete short intervals into relatively more complete FEs intervals. However, the refinement will also make short false positive intervals more obvious. The neutral frames near false positive intervals in the LSDM proposal are predicted as FEs in the final prediction. In addition, since FESNet is mainly designed for ME spotting, MaE spotting is likely to fail when there is a MaE with significant head movement in the original input, which is also one of the limitations of FESNet.

	F1-score			Edit	Acc
	MaE	*ME*	*Overall*		
w/o $\mathcal{L}_{LTDM}$	49.83	15.88	38.24	46.37	53.79
w/o MTAM	47.24	9.35	36.62	**54.78**	58.11
FESNet	**55.50**	**16.26**	**42.34**	52.72	**59.80**

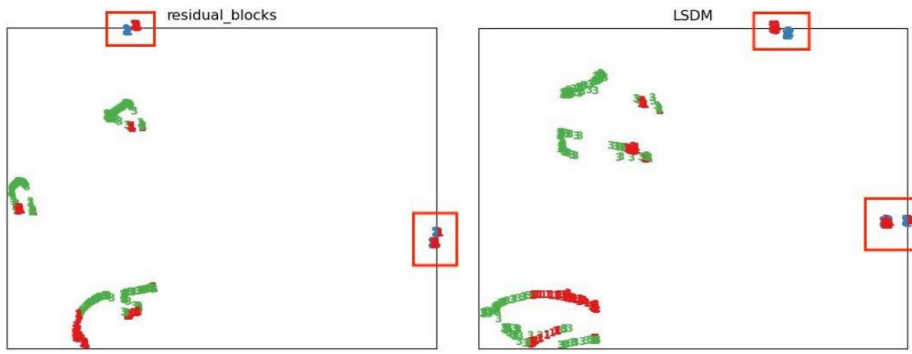
Table 3. Ablation studies on SAMM

4.4 Ablation Studies

We carried out ablation studies on SAMM to explore the effectiveness of $\mathcal{L}_{LSDM}$ and MTAM. The quantitative results are shown in Figure 3. It can be observed that the FEs spotting performance drops obviously without the constraint of $\mathcal{L}_{LSDM}$. This is because the constraint of $\mathcal{L}_{LSDM}$ is to make the LSDM extract facial motion-related features. With the absence of $\mathcal{L}_{LSDM}$, the LSDM will extract motion-irrelevant spatial static features which interfere with FEs spotting, especially when there is FEs-irrelevant motion in neutral frames.

γ	F1-score			Edit	Acc
	MaE	*ME*	*Overall*		
0.00	45.22	1.63	40.25	86.89	89.33
0.05	39.92	6.89	33.23	85.29	87.91
0.10	40.91	12.50	34.66	86.49	90.07
0.15	45.30	**13.81**	39.89	**87.33**	**91.12**
0.20	47.18	4.51	41.54	86.99	90.49
0.25	48.16	3.10	40.19	85.70	90.39
0.30	**49.63**	6.25	**43.93**	85.42	89.31

Table 4. Analysis of the impact of γ. Experiments were conducted on CAS(ME)^2 to analyze the impact of the weight of L_{smooth} on the performance of FESNet.



a) residual blocks b) LSDM

Figure 7. TSNE visualization results of feature distribution obtained by residual blocks and LSDM. Colors indicate different FEs: red for neutral FEs, green for MaE, and blue for ME. As can be seen from the red box area, the class centers of LSDM features are further apart, leading to a better result, especially for the classification of neutral FEs and ME.

As shown in Figure 7, it can be found easily from the areas marked by red boxes that the distribution of neutral FEs and ME from residual blocks are mixed, which the proposed LSDM alleviates. The visualization results show that the proposed LSDM can extract more discriminative spatial static features, especially for ME and neutral FEs. In addition, the FEs spotting performance drops more obviously in the absence of MTAM. As shown in Figure 8, the proposals from LSDM for FEs spotting are unstable within the input interval, resulting in the low-performance in Figure 3. Without the MTAM incorporating the multi-scale temporal correlation, the final prediction depends on the independent spatial features. Thus, the frame-wise classification results do not contain the correlation in the temporal dimension. Therefore, the FEs spotting results will suffer from severe over-segmentation, one of the main challenges in FEs spotting. Besides, we

found a fascinating phenomenon: individuals' identifying information will specifically impact FEs spotting. As shown in Figure 7, multiple cluster centers exist for the same FEs. Each cluster center represents an individual in the latent space.

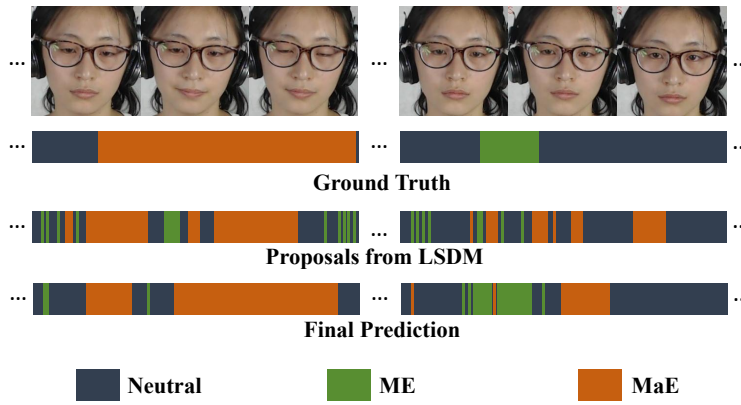


Figure 8. Visualization of the FEs spotting results of our method. Although there are still failure cases in the final prediction, the over-segmentation has been effectively alleviated compared with the proposals of LSDM.

5 CONCLUSION

This paper proposes an end-to-end FESNet network to extract efficient spatial and multi-scale temporal features. Comprehensive experiments show that our method can extract more discriminative spatial features and effectively model the temporal motion pattern of FEs. Besides, the proposed FESNet significantly outperforms existing state-of-the-art methods and supplies a new solution for FEs spotting. In our future work, we will combine FESNet with 3D CNN to extract the spatial features and simultaneously model the temporal motion pattern. In addition, we also plan to introduce biological models and consistency loss constraints to achieve better performance in ME recognition and generation.

Acknowledgement

This work was sponsored by the Natural Science Foundation of Chongqing, China (No. CSTB2022NSCQ-MSX0552), the National Natural Science Foundation of China (No. 62002121, 62072183), the Shanghai Science and Technology Commission (No. 21511100700, 22511104600), the Open Project Program of the State Key Lab of CAD & CG (No. A2203), Zhejiang University.

REFERENCES

[1] Jampour, M.—Javidi, M.: Multiview Facial Expression Recognition, A Survey. IEEE Transactions on Affective Computing, Vol. 13, 2022, No. 4, pp. 2086–2105, doi: 10.1109/TAFFC.2022.3184995.

[2] Ben, X.—Ren, Y.—Zhang, J.—Wang, S. J.—Kpalma, K.—Meng, W.—Liu, Y. J.: Video-Based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, 2022, No. 9, pp. 5826–5846, doi: 10.1109/TPAMI.2021.3067464.

[3] Liu, Y. J.—Zhang, J. K.—Yan, W. J.—Wang, S. J.—Zhao, G.—Fu, X.: A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition. IEEE Transactions on Affective Computing, Vol. 7, 2016, No. 4, pp. 299–310, doi: 10.1109/TAFFC.2015.2485205.

[4] Xu, F.—Zhang, J.—Wang, J. Z.: Microexpression Identification and Categorization Using a Facial Dynamics Map. IEEE Transactions on Affective Computing.

[5] Li, J.—Dong, Z.—Lu, S.—Wang, S. J.—Yan, W. J.—Ma, Y.—Liu, Y.—Huang, C.—Fu, X.: CAS(ME)3: A Third Generation Facial Spontaneous Micro-Expression Database with Depth Information and High Ecological Validity. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, 2023, No. 3, pp. 2782–2800, doi: 10.1109/TPAMI.2022.3174895.

[6] Leng, W.—Zhao, S.—Zhang, Y.—Liu, S.—Mao, X.—Wang, H.—Xu, T.—Chen, E.: ABPN: Apex and Boundary Perception Network for Micro- and Macro-Expression Spotting. Proceedings of the 30th ACM International Conference on Multimedia (MM '22), 2022, pp. 7160–7164, doi: 10.1145/3503161.3551599.

[7] Dosovitskiy, A.—Beyer, L.—Kolesnikov, A.—Weissenborn, D.—Zhai, X.—Unterthiner, T.—Dehghani, M.—Minderer, M.—Heigold, G.—Gelly, S.—Uszkoreit, J.—Houlsby, N.: An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR 2021), 2021, doi: 10.48550/arXiv.2010.11929.

[8] Li, J.—Soladié, C.—Séguier, R.: Local Temporal Pattern and Data Augmentation for Micro-Expressions Spotting. IEEE Transactions on Affective Computing, Vol. 14, 2023, No. 1, pp. 811–822, doi: 10.1109/TAFFC.2020.3023821.

[9] Polikovsky, S.—Kameda, Y.—Ohta, Y.: Facial Micro-Expressions Recognition Using High Speed Camera and 3D-Gradient Descriptor. 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009), IET, 2009, pp. 1–6, doi: 10.1049/ic.2009.0244.

[10] Polikovsky, S.—Kameda, Y.—Ohta, Y.: Detection and Measurement of Facial Micro-Expression Characteristics for Psychological Analysis. IEICE Technical Report, Vol. 110, 2010, No. 97, Art. No. PRMU2010-48.

[11] Li, X.—Hong, X.—Moilanen, A.—Huang, X.—Pfister, T.—Zhao, G.—Pietikäinen, M.: Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-Expression Spotting and Recognition Methods. IEEE Transactions on Affective Computing, Vol. 9, 2018, No. 4, pp. 563–577, doi: 10.1109/TAFFC.2017.2667642.

[12] Ojala, T.—Pietikäinen, M.—Mäenpää, T.: Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, 2002, No. 7, pp. 971–987, doi: 10.1109/TPAMI.2002.1017623.

[13] Liu, C.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. Ph.D. Thesis. Massachusetts Institute of Technology, Cambridge, MA, USA, 2009, https://hdl.handle.net/1721.1/53293.

[14] Yan, W. J.—Wang, S. J.—Chen, Y. H.—Zhao, G.—Fu, X.: Quantifying Micro-Expressions with Constraint Local Model and Local Binary Pattern. In: Agapito, L., Bronstein, M. M., Rother, C. (Eds.): Computer Vision - ECCV 2014 Workshops. Springer, Cham, Lecture Notes in Computer Science, Vol. 8925, 2014, pp. 296–305, doi: 10.1007/978-3-319-16178-5_20.

[15] Liong, S. T.—See, J.—Wong, K.—Le Ngo, A. C.—Oh, Y. H.—Phan, R.: Automatic Apex Frame Spotting in Micro-Expression Database. 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2015, pp. 665–669, doi: 10.1109/ACPR.2015.7486586.

[16] Esmaeili, V.—Shahdi, S. O.: Automatic Micro-Expression Apex Spotting Using Cubic-LBP. Multimedia Tools and Applications, Vol. 79, 2020, No. 27-28, pp. 20221–20239, doi: 10.1007/S11042-020-08737-5.

[17] Davison, A.—Merghani, W.—Lansley, C.—Ng, C. C.—Yap, M. H.: Objective Micro-Facial Movement Detection Using FACS-Based Regions and Baseline Evaluation. 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), 2018, pp. 642–649, doi: 10.1109/FG.2018.00101.

[18] Davison, A. K.—Yap, M. H.—Lansley, C.: Micro-Facial Movement Detection Using Individualised Baselines and Histogram-Based Descriptors. 2015 IEEE International Conference on Systems, Man, and Cybernetics, 2015, pp. 1864–1869, doi: 10.1109/SMC.2015.326.

[19] Li, J.—Soladié, C.—Séguier, R.—Wang, S. J.—Yap, M. H.: Spotting Micro-Expressions on Long Videos Sequences. 2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019), 2019, pp. 1–5, doi: 10.1109/FG.2019.8756626.

[20] Moilanen, A.—Zhao, G.—Pietikäinen, M.: Spotting Rapid Facial Movements from Videos Using Appearance-Based Feature Difference Analysis. 2014 22nd International Conference on Pattern Recognition, IEEE, 2014, pp. 1722–1727, doi: 10.1109/ICPR.2014.303.

[21] Xia, Z.—Feng, X.—Peng, J.—Peng, X.—Zhao, G.: Spontaneous Micro-Expression Spotting via Geometric Deformation Modeling. Computer Vision and Image Understanding, Vol. 147, 2016, pp. 87–94, doi: 10.1016/j.cviu.2015.12.006.

[22] Soh, X. R.—Baskaran, V. M.—Buhari, A. M.—Phan, R. C. W.: A Real Time Micro-Expression Detection System with LBP-TOP on a Many-Core Processor. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2017, pp. 309–315, doi: 10.1109/APSIPA.2017.8282041.

[23] Zhao, G.—Pietikäinen, M.: Dynamic Texture Recognition Using Local Binary

Patterns with an Application to Facial Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, 2007, No. 6, pp. 915–928, doi: 10.1109/TPAMI.2007.1110.

[24] Zhao, Y.—Tong, X.—Zhu, Z.—Sheng, J.—Dai, L.—Xu, L.—Xia, X.—Jiang, Y.—Li, J.: Rethinking Optical Flow Methods for Micro-Expression Spotting. Proceedings of the 30th ACM International Conference on Multimedia (MM '22), 2022, pp. 7175–7179, doi: 10.1145/3503161.3551602.

[25] Shreve, M.—Godavarthy, S.—Manohar, V.—Goldgof, D.—Sarkar, S.: Towards Macro- and Micro-Expression Spotting in Video Using Strain Patterns. 2009 Workshop on Applications of Computer Vision (WACV), IEEE, 2009, pp. 1–6, doi: 10.1109/WACV.2009.5403044.

[26] Shreve, M.—Godavarthy, S.—Goldgof, D.—Sarkar, S.: Macro- and Micro-Expression Spotting in Long Videos Using Spatio-Temporal Strain. 2011 IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2011, pp. 51–56, doi: 10.1109/FG.2011.5771451.

[27] Patel, D.—Zhao, G.—Pietikäinen, M.: Spatiotemporal Integration of Optical Flow Vectors for Micro-Expression Detection. In: Battiato, S., Blanc-Talon, J., Gallo, G., Philips, W., Popescu, D., Scheunders, P. (Eds.): Advanced Concepts for Intelligent Vision Systems (ACIVS 2015). Springer, Cham, Lecture Notes in Computer Science, Vol. 9386, 2015, pp. 369–380, doi: 10.1007/978-3-319-25903-1_32.

[28] Liong, S. T.—See, J.—Phan, R. C. W.—Le Ngo, A. C.—Oh, Y. H.—Wong, K.: Subtle Expression Recognition Using Optical Strain Weighted Features. In: Jawahar, C. V., Shan, S. (Eds.): Computer Vision - ACCV 2014 Workshops. Springer, Cham, Lecture Notes in Computer Science, Vol. 9009, 2014, pp. 644–657, doi: 10.1007/978-3-319-16631-5_47.

[29] Guo, Y.—Li, B.—Ben, X.—Ren, Y.—Zhang, J.—Yan, R.—Li, Y.: A Magnitude and Angle Combined Optical Flow Feature for Microexpression Spotting. IEEE MultiMedia, Vol. 28, 2021, No. 2, pp. 29–39, doi: 10.1109/MMUL.2021.3058017.

[30] Shreve, M.—Brizzi, J.—Fefilatyev, S.—Luguev, T.—Goldgof, D.—Sarkar, S.: Automatic Expression Spotting in Videos. Image and Vision Computing, Vol. 32, 2014, No. 8, pp. 476–486, doi: 10.1016/j.imavis.2014.04.010.

[31] Zhang, L. W.—Li, J.—Wang, S. J.—Duan, X. H.—Yan, W. J.—Xie, H. Y.—Huang, S. C.: Spatio-Temporal Fusion for Macro- and Micro-Expression Spotting in Long Video Sequences. 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 734–741, doi: 10.1109/FG47880.2020.00037.

[32] Kim, D. H.—Baddar, W. J.—Ro, Y. M.: Micro-Expression Recognition with Expression-State Constrained Spatio-Temporal Feature Representations. Proceedings of the 24th ACM Conference on Multimedia (MM '16), 2016, pp. 382–386, doi: 10.1145/2964284.2967247.

[33] Khor, H. Q.—See, J.—Phan, R. C. W.—Lin, W.: Enriched Long-Term Recurrent Convolutional Network for Facial Micro-Expression Recognition. 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), 2018, pp. 667–674, doi: 10.1109/FG.2018.00105.

[34] Nag, S.—Bhunia, A. K.—Konwer, A.—Roy, P. P.: Facial Micro-Expression Spotting and Recognition Using Time Contrasted Feature with Visual Memory. ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 2022–2026, doi: 10.1109/ICASSP.2019.8683737.

[35] Wang, S. J.—He, Y.—Li, J.—Fu, X.: MESNet: A Convolutional Neural Network for Spotting Multi-Scale Micro-Expression Intervals in Long Videos. IEEE Transactions on Image Processing, Vol. 30, 2021, pp. 3956–3969, doi: 10.1109/TIP.2021.3064258.

[36] Yap, C. H.—Yap, M. H.—Davison, A.—Kendrick, C.—Li, J.—Wang, S. J.—Cunningham, R.: 3D-CNN for Facial Micro- and Macro-Expression Spotting on Long Video Sequences Using Temporal Oriented Reference Frame. Proceedings of the 30th ACM International Conference on Multimedia (MM '22), 2022, pp. 7016–7020, doi: 10.1145/3503161.3551570.

[37] Xie, Z.—Cheng, S.—Liu, X.—Fan, J.: Adaptive Enhanced Micro-Expression Spotting Network Based on Multi-Stage Features Extraction. In: Deng, W., Feng, J., Huang, D., Kan, M., Sun, Z., Zheng, F., Wang, W., He, Z. (Eds.): Biometric Recognition (CCBR 2022). Springer, Cham, Lecture Notes in Computer Science, Vol. 13628, 2022, pp. 287–296, doi: 10.1007/978-3-031-20233-9_29.

[38] Pan, H.—Xie, L.—Wang, Z.: Spatio-Temporal Convolutional Attention Network for Spotting Macro- and Micro-Expression Intervals. In: Cheng, W. H., Li, J., Yap, M. H. (Eds.): Proceedings of the 1st Workshop on Facial Micro-Expression: Advanced Techniques for Facial Expressions Generation and Spotting (FME '21). ACM, 2021, pp. 25–30, doi: 10.1145/3476100.3484463.

[39] Yang, B.—Wu, J.—Zhou, Z.—Komiya, M.—Kishimoto, K.—Xu, J.—Nonaka, K.—Horiuchi, T.—Komorita, S.—Hattori, G.—Naito, S.—Takishima, Y.: Facial Action Unit-Based Deep Learning Framework for Spotting Macro- and Micro-Expressions in Long Video Sequences. Proceedings of the 29th ACM International Conference on Multimedia (MM '21), 2021, pp. 4794–4798, doi: 10.1145/3474085.3479209.

[40] Yu, W. W.—Jiang, J.—Li, Y. J.: LSSNet: A Two-Stream Convolutional Neural Network for Spotting Macro- and Micro-Expression in Long Videos. Proceedings of the 29th ACM International Conference on Multimedia (MM '21), 2021, pp. 4745–4749, doi: 10.1145/3474085.3479215.

[41] Carreira, J.—Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4724–4733, doi: 10.1109/CVPR.2017.502.

[42] Liong, G. B.—See, J.—Wong, L. K.: Shallow Optical Flow Three-Stream CNN for Macro- and Micro-Expression Spotting from Long Videos. 2021 IEEE International Conference on Image Processing (ICIP), 2021, pp. 2643–2647, doi: 10.1109/ICIP42928.2021.9506349.

[43] Tran, T. K.—Vo, Q. N.—Zhao, G.: DynGeoNet: Fusion Network for Micro-Expression Spotting. Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), ACM, 2021, pp. 745–749, doi: 10.1145/3462244.3479958.

[44] Yap, C. H.—Kendrick, C.—Yap, M. H.: SAMM Long Videos: A Spontaneous Facial Micro- and Macro-Expressions Dataset. 2020 15th IEEE International Confer-

ence on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 771–776, doi: 10.1109/FG47880.2020.00029.

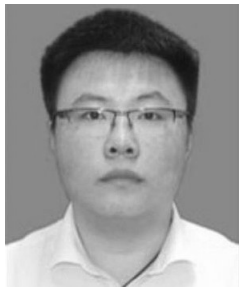
[45] Qu, F.—Wang, S. J.—Yan, W. J.—Li, H.—Wu, S.—Fu, X.: CAS(ME)^2: A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. IEEE Transactions on Affective Computing, Vol. 9, 2018, No. 4, pp. 424–436, doi: 10.1109/TAFFC.2017.2654440.

[46] Zhang, K.—Zhang, Z.—Li, Z.—Qiao, Y.: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters, Vol. 23, 2016, No. 10, pp. 1499–1503, doi: 10.1109/LSP.2016.2603342.

[47] Li, J.—Yap, M. H.—Cheng, W. H.—See, J.—Hong, X.—Li, X.—Wang, S. J.: FME '21: 1st Workshop on Facial Micro-Expression: Advanced Techniques for Facial Expressions Generation and Spotting. Proceedings of the 29th ACM International Conference on Multimedia (MM '21), 2021, pp. 5700–5701, doi: 10.1145/3474085.3478579.

[48] Farha, Y. A.—Gall, J.: MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3575–3584, doi: 10.1109/CVPR.2019.00369.

[49] He, Y.—Wang, S. J.—Li, J.—Yap, M. H.: Spotting Macro-and Micro-Expression Intervals in Long Video Sequences. 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 742–748, doi: 10.1109/FG47880.2020.00036.



Bohao Zhang received his M.Sc. degree in software engineering from the School of Software Engineering Institute, East China Normal University, Shanghai, China, in 2021. He is currently pursuing his Ph.D. degree at the School of Computer Science and Technology, East China Normal University, Shanghai, China. His research interests include computer vision, pattern recognition, and intelligent education.

Jiale Lu received the bachelor degree in the School of Computer Science and Technology at the East China Normal University, Shanghai, China, in 2021. He is working toward a Master's degree in the School of Computer Science and Technology, East China Normal University, Shanghai, China. His main research interests include computer vision and pattern recognition.

Changbo Wang is Professor in the School of Computer Science and Technology, East China Normal University, China. He received his Ph.D. degree at the State Key Laboratory of CAD and CG, Zhejiang University in 2006, and received his B.Eng. degree in 1998, and his M.Eng. degree in civil engineering in 2002, respectively, both from the Wuhan University of Technology. His research interests include physically based modeling and rendering, computer animation and realistic image synthesis, information visualization, and others.

Gaoqi He is currently Professor in the School of Computer Science and Technology at the East China Normal University. He received his Ph.D. degree from the State Key Laboratory of CAD and CG, Zhejiang University, China, in 2007. His research interests include computer graphics, computer vision and machine learning. His work aims to develop efficient and practical scene understanding and visual analysis algorithms, with applications including video processing, crowd simulation, virtual reality and augmented reality.