EasyChair Preprint

№ 9818

Bidirectional Cross-Scale Feature Fusion for Long Video Micro-Expression 3D Spotting Network

Xiaosong He, Xiao Wu, Jun Peng, Qingxia Li, Xinkai Ma and Yuanmin He

March 5, 2023

Bidirectional Cross-scale Feature Fusion for Long Video Micro-Expression 3D Spotting Network

Xiaosong He*††§, Xiao Wu§, Jun Peng‡¶§, Qingxia Li§, Xinkai Ma§, Yuanmin He§
*Informatization Office
Chongqing University of Science and Technology, Chongqing 401331, China
E-mail: xiaosh@cqust.edu.cn
††Corresponding author
‡College of Mathematics, Physics and Data Science
Chongqing University of Science and Technology, Chongqing 401331, China
¶Chongqing Sino-German Future Factory Research Institute, Chongqing 401331, China
E-mail: jpeng@cqust.edu.cn
§College of Intelligent Technology and Engineering
Chongqing University of Science and Technology, Chongqing 401331, China
E-mail: 1291383707@qq.com, 690887086@qq.com,
irvingmxk@163.com, 523031150@qq.com

***Abstract*—Psycho-cognitive computing is an important part of intelligent human-computer interaction technology, which has received extensive attention in recent years. The research of micro-expressions can reflect the depth and breadth of mental cognitive computing. Micro-expression(ME) is a spontaneous, short-lived, and inadvertent facial expression. The research on ME is of great significance in sentiment analysis, criminal investigation, and psychology research. ME spotting refers to locating sequences of micro-expressions in a long video. ME detection is an extremely important step in the field of ME analysis. Based on the I3D backbone network of long video optical flow extraction and original video feature extraction, this paper extracts effective feature layers, performs downsampling, and uses the BiFPN module with a spatial attention mechanism to selectively fuse the extracted multi-scale feature layers. The final classification-regression network discriminates the detected expressions and locates the temporal boundaries where the expressions occur. The experimental results show that the proposed method effectively improves the F1 score index. Compared with other deep learning methods, this method performs better on both CAS(ME)^2 and SAMM. The proposed method provides a reliable foundation for the downstream tasks of ME research, such as ME recognition, and also provides an effective method for collecting natural macro- or micro-expression datasets from a large amount of video data in the future.**

***Index Terms*—Deep learning; Psycho-cognitive computing; 3D convolutional neural networks; BiFPN; Optical flow; Macro- and micro-expression spotting**

I. Introduction

Facial expressions usually contain a lot of emotional cognitive interaction information, representing a person's emotional expression at that time. In 1969, research by Ekman and Friesen [1] discovered a special kind of facial expression: Facial micro-expression (ME). ME is a subtle, involuntary expression that lasts no more than 0.5s, as opposed to macro-expression (MaE), which has a duration of 0.5s to 4s [2]. When ME occur, they are generally manifested as weak facial movements that are not controlled by humans, short-lived, and subconsciously contain real emotional information but are suppressed. ME are important cues for lie detection [3], [4], which are distinguished by local motion, short duration, and low intensity. ME analysis has many potential applications in psychology, criminal investigation, depression treatment and national security. ME spotting is the basis of ME analysis. The research on ME detection is also in full swing in the field of computer vision with the development of deep learning, especially object detection. ME detection refers to finding the start, climax and end time of ME in a picture sequence or video. ME recognition is to classify the detected micro-expression sequences. Compared with the detection task, ME recognition is much simpler and is essentially a classification task.

In the early related research on ME, some traditional methods were proved to be effective, such as comparing the feature difference (FD) within a fixed-length time window to determine the peak frame of the ME sequence. Among them, features such as LBP [5], HOG [6], Optical Flow [7] [8] [9] are common features of ME detection based on FD method. Although the FD algorithm considers the temporal characteristics, the captured movements are not necessarily ME movements, but only other facial movements with similar intensity or duration. Therefore, the FD algorithm is insufficient in distinguishing ME movements from other facial movements.

In recent years, ME spotting and recognition are increasingly combined with deep learning algorithms for feature extraction. Since deep learning algorithms often use a large number of feature extraction network layers to extract image features, a large number of data samples are needed to train these models to achieve the desired effect, but ME is a small-sample task, and existing datasets cannot support

large-scale depth Learning the network for training, based on this, deep learning methods often require a large number of preprocessing operations on the ME dataset, such as temporal normalization of the original data.

In 2018, Zhang [10] first proposed the use of convolutional neural networks to extract ME image features, and achieved good results in extracting ME peak frame positions. After the long video databases of CAS(ME)^2 [11] and SAMM [12] were made public, the academic community began to study ME detection based on long videos. In 2020, Zhang [13]et al. proposed to calculate the difference between the optical flow superposition of the local motion and the global motion vector by estimating the average optical flow of the nose area to obtain the local optical flow field, and use the spatiotemporal feature fusion matrix to describe the spatiotemporal information, Finally, the start frame, peak frame and end frame information of the expression occurrence are obtained according to the SP mode. It turns out that fusing optical flow features into the features extracted by deep learning models has a great positive effect on the task of detecting ME in long videos.

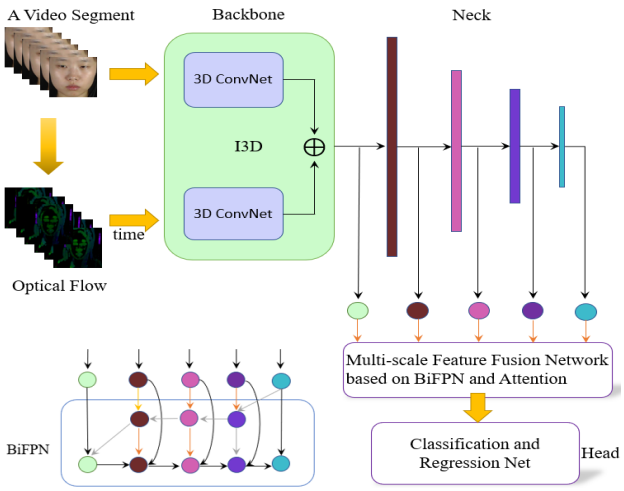

Fig. 1. ME spotting method based on I3D and BiFPN networks.

II. Related work

ME spotting based on long video can be summarized into the category of behavior recognition or video understanding in the field of computer vision.

In the field of video understanding, the fusion model of LSTM [14] and CNN became popular first. The main idea of this method is to use CNN to extract spatial features and use LSTM to obtain time series information. The performance of the above model is undistinguishe. Later, with the emergence of 3D convolutional neural networks and two-stream networks, this method was gradually abandoned. The 3D convolutional neural network(3D-CNN) [15] divides the long video into short video segments, each video segment contains from 1 to k pictures, and then the pictures are input into the network as a volume, the convolution kernel of the network must be 3D , in order to satisfy the processing of spatial features and information in the temporal dimension of two-dimensional images. Given enough datasets, 3D-CNN can show good performance, but the amount of parameters will be very large.

The Two-Stream [16] network not only avoids using LSTM for time series simulation, but also does not need to use 3D network to learn spatiotemporal information. It uses the extracted optical flow image to represent the time series information of the video as one time stream, while the other stream is the input One or more video frames are used as the spatial stream, and the 2D convolutional network extracts features from the input of the temporal stream and the spatial stream, and finally classifies the features. Therefore, the complexity of the model is greatly reduced, it is easier to train, and the effect is perfect. Subsequently, the researchers found that the effect of using 3D-CNN and optical flow information as auxiliary input in the field of video understanding is obvious when the dataset is sufficient. Therefore, Joao Carreira [17] and others proposed a two-stream I3D network. The essence of I3D is to inflate a 2D convolutional network into a 3D convolutional network. After comparing the experimental results of the previous LSTM+CNN and 3D+CNN on different datasets, the dual-stream 3D networks can achieve the best results, and it also shows from the side that extracting optical flow is always a beneficial operation for video understanding.

In summary, the light flow and spatial characteristics of using I3D convolution extraction are helpful for video understanding tasks. In response to the microscopicity of the ME spotting task, the method of downsampling and multi-scale fusion of the I3D extraction features can effectively locate the occurrence of ME.

III. Proposed method

ME spotting tasks belong to small sample tasks. The amplitude of the movement of micro-expressions is subtle, which leads to the improvement of detection difficulty. Thanks to the good performance of light flow characteristics in ME spotting tasks, Using I3D-based feature extraction network to extract optical flow features and motion features from videos. Feeding the effective feature layer to the BiFPN [18] network based on the fast restricted fusion method, thereby spotting the ME.

The main work is to extract the optical flow and spatial feature of the cropped image, then use BiFPN for multi-scale feature fusion, and finally design a classification and regression network to realize the classification and positioning of expressions.The overall structure of the model is shown in “Fig. 1”.

A. Two-Stream Inflated 3D ConvNet

Video can be understood as adding a time dimension to the image. Therefore, in the field of video, different network forms are usually distinguished by how to process the timing information that comes with the video. In the past video understanding models, 2D convolutional neural networks, 3D convolutional neural networks and dual-stream networks combined with LSTM are the main schools. After research, it is

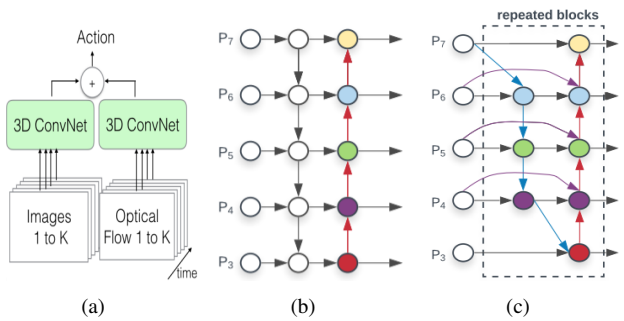

Fig. 2. (a) I3D Schematic; (b) PANET Structure diagram; (c) BiFPN Structure diagram.

found that when the network is pre-trained on a large scale and then fine-tuned on a small data set, the results of the above different networks are uneven, and even if there is an improvement, it is not very significant. Therefore, Joao Carreira et al. Based on the fusion improvement, a "Two-Stream Inflated 3D ConvNets" (I3D) based on Inception-v1 is proposed. Based on existing 2D image classification networks, such as AlexNet, VGG, etc., the success of I3D proves that roughly expanding 2D networks into 3D networks is a very effective operation. Subsequently, the Non-local neural networks of Wang [19] and others also implemented I3D on ResNet. The core idea is to expand 2D to 3D, still use multiple residual blocks to stack, and still go through four stages, each residual's layer structure (Conv, BN, ReLU) of the difference block is also not changed. This strategy is also reflected in the latest video swin transformer [20]. In view of the excellent results of I3D in the field of video understanding and the important position of video-based ME datasets in the ME spotting task. In this paper, I3D is used as the backbone network for the ME detection task to extract the optical flow (as timing information) and spatial features of the original video of the ME dataset.

B. Bidirectional Feature Network

FPN [21] is an important method for deal with multi-scale features to target detection. Its core idea is to combine multi-scale features in a top-down method. Later, PANET [22] also added an additional bottom-up aggregation network on the basis of FPN, while NAS-FPN [23] changed the neural structure search to automatically design the feature network topology, although it obtained better performance, but in the search process Requires thousands of GPU hours. So Mingxing Tan and colleagues [18] proposed BiFPN to optimize multi-scale feature fusion in a more intuitive and principled way.

As can be seen from "Fig. 2. c", BiFPN removes nodes with only one input in PANET, this can reduce calculation, in addition, adds some skip connections to fuse more features without adding too much cost, also have the ability to increase the enhanced characteristics of residual links. And finally, each bidirectional path as a feature network layer, higher-level feature fusion can be achieved.

The feature fusion method of BiFPN integrates both the bidirectional cross-scale connections and the fast normalized fusion. "Equation (1) is the feature fusion method of two nodes in the P6."

$$P_6^{td} = Conv(\frac{w_1 P_6^{in} + w_2 Resize(P_7^{in})}{w_1 + w_2 + \epsilon}) \quad (1)$$

$$P_6^{out} = Conv(\frac{w_1' P_6^{in} + w_2' P_6^{td} + w_3' Resize(P_5^{out})}{w_1' + w_2' + w_3' + \epsilon}) \quad (2)$$

where w is the weight, also known as the attention ratio, which is used to determine the degree of attention to input. In order to ensure that the weight(w) is greater than 0, the ReLU activation function is used before the weight. P_6^{td} is the intermediate node feature of $P6$ in the up-down path, and P_6^{out} is the output node feature of $P6$ in the bottom-up path. Features of other layers have been constructed in a similar manner. To improve efficiency, depthwise separable convolution is used for feature fusion, and BN and ReLU operations are added after each convolution.

C. Fusion Improved I3D and BiFPN

In order to adapt to the ME spotting task and improve the efficiency, this paper integrates and improves the I3D network and BiFPN. The model details are shown in "Fig. 3". The ME spotting task based on deep learning generally consists of three steps, namely datasets preprocessing, feature extraction and fusion, and classification and localization of the spotting expressions. Optical flow can describe the motion information between frames and capture the subtle changes of the sample frame sequence, so the Two-Stream network based on 3D-CNN combined with optical flow is used to detect the expressions in the video clips.

First, Shorter video clips are merged before extracting optical flow with the TV-L1 algorithm. During training, the data is randomly cropped and left-right flipped. Second, because the spatiotemporal resolution of the ME dataset meets the requirements of I3D, the backbone network uses the I3D model as a feature extraction network. Each convolutional layer is followed by a batch normalization layer and a ReLU activation function, and Adam is used as an optimizer for training videos. After the 3rd, 4th, and 5th pooling layers, the feature maps are extracted as the input $P3\_in$, $P4\_in$, and $P5\_in$ of the feature fusion network. P3, P4, and P5 can be obtained through the backbone feature extraction network, and P6 and P7 are obtained by downsampling twice on P5. P3, P4, and P5 are used as $P3\_in$, $P4\_in$, and $P5\_in$ after adjusting the number of channels by 1×1 convolution. After all inputs are obtained, $P7\_in$ is upsampled, and after upsampling, they are stacked with $P6\_in$ to obtain $P6\_td$; Then $P6\_td$ is upsampled and stacking with $P5\_in$ to obtain $P5\_td$; similarly, after upsampling $P5\_td$ and stacking with $P4\_in\_1$ to obtain $P4\_td$; And then, upsampling $P4\_td$ and stacking with $P3\_in$ to obtain $P3\_out$. After the upsampling is completed, $P3\_out$, $P4\_td$, $P4\_in\_2$, $P5\_td$, $P5\_in\_2$, $P6\_in$, $P6\_td$, and $P7\_in$ can be obtained, and then downsampling is performed. The downsampling process is similar

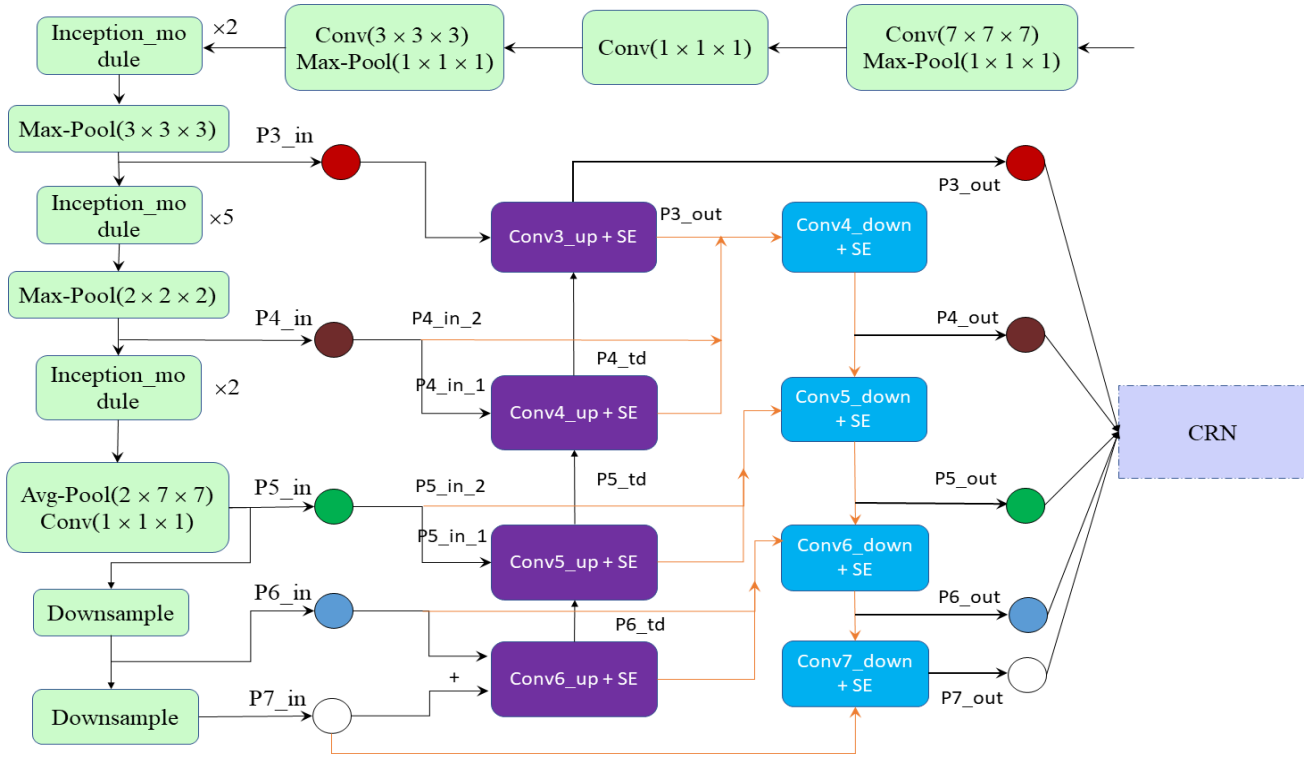

Fig. 3. Proposed network details: the backbone network extracts three effective feature layers, downsamples the last layer twice, and 5 effective feature layers are used as the input of the feature fusion network.

to the upsampling. Finally, the effective feature layers of $P3\_out$, $P4\_out$, $P5\_out$, $P6\_out$, and $P7\_out$ are obtained. Furthermore, in order to determine the attention to each input weight, a SE attention module is added.

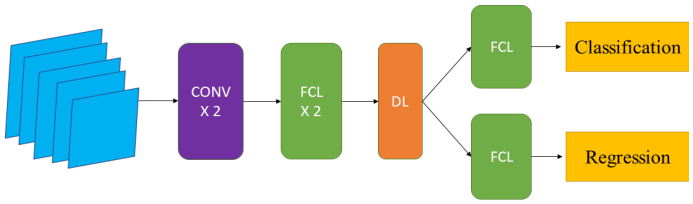

Fig. 4. Classification and Regression Network.

D. Classification and Regression Network

The effective feature layer is transmitted to the classification and regression network(CRN) to obtain the prediction result. The architecture of CRN is illustrated in "Fig. 4". Where the CONV refers to two 1D convolution layers with the same configuration of kernel size 3, no padding, stride 1, and ReLU activation function. FCL is two fully connected layers with the same configuration, and DL refers to the dropout layer with 0.5 ratio to reduce overfitting. Finally, two parallel fully-connected layers are appended. One has 2 neurons and a Softmax activation function to classify the proposed clip into "ME" or "non-ME". Another also has 2 neurons but without any activation function to regresses the temporal boundaries of the proposed clip. The classification loss of the CRN network is calculated by the cross-entropy loss function, and the regression loss is calculated by the Smooth L1 loss function.

After the network outputs the probability and regression value, a post-processing module is set to convert the output value into a detection segment, and finally NMS is used to remove the overlapping area. The classification network is mainly responsible for distinguishing the generated ME candidates (MaE or ME), and the regression network is mainly responsible for the classification results. Boundary regression of dimensions to locate the onset of occurrence of ME or MaE.

IV. Experiments

A. Datasets, Performance Metrics and Configuration

We validate the proposed method on the CAS(ME)^2 and SAMM datasets. CAS(ME)^2 contains 22 subjects, 57 MEs and 300 MaEs. It provides 98 long videos at 30FPS, each of which contains many spontaneous MaEs and MEs, with an average duration of 86 seconds. The SAMM dataset contains 343 MaEs and 159 MEs from 32 subjects, providing 147 long videos at 200 FPS with an average duration of 35 seconds. Action units and categories are annotated for both datasets.

In this paper, the model evaluation metrics proposed in MESNet [24] are borrowed. For each segment detected, use the

TABLE I
VARIOUS INDICATOR PARAMETERS UNDER DIFFERENT TEVAL ON CAS(ME)^2 DATASET.

T_eval	MaE			ME			Overall		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
0.50	**0.314**	**0.513**	**0.390**	**0.067**	**0.235**	**0.104**	**0.259**	**0.515**	**0.345**
0.40	**0.358**	**0.681**	**0.467**	**0.113**	**0.362**	**0.171**	**0.347**	**0.669**	**0.455**
0.35	**0.366**	**0.694**	**0.478**	**0.127**	**0.384**	**0.190**	**0.356**	**0.688**	**0.468**
0.30	**0.410**	**0.712**	**0.520**	**0.145**	**0.432**	**0.216**	**0.402**	**0.724**	**0.511**

intersection ratio (Equation (3)) to determine if the prediction was correct.

$$\frac{W_{spotted} \bigcap W_{groundTruth}}{W_{spotted} \bigcup W_{groundTruth}} \geq T\_eval \tag{3}$$

where T_{eval} is the evaluation threshold, which is 0.5 by default. When the intersection ratio is greater than or equal to T, it is regarded as TP, and the TP numbers of MaEs and MEs in the video are set as m and n, the total number of MaEs and MEs detected according to the experimental records are recorded as a and b, The total number of macro-micro expressions in the video is denoted as M and N, respectively. Finally, Then the Recall, Precision and F1-score are evaluated as follows:

$$Recall = \frac{m+n}{a+b} \tag{4}$$

$$Precision = \frac{m+n}{M+N} \tag{5}$$

$$F1-score = \frac{2 \times (Recall \times Precision)}{Recall + Precision} \tag{6}$$

B. Configuration and Result Analysis

TABLE II
F1-SCORE WITH DIFFERENT METRICS ON CAS(ME)^2 AND SAMM DATASET.

Top_threshold	CAS(ME)^2			SAMM		
	MaE	ME	All	MaE	ME	All
0.30	**0.252**	**0.069**	**0.241**	**0.268**	**0.054**	**0.253**
0.35	**0.263**	**0.061**	**0.255**	**0.272**	**0.044**	**0.264**
0.40	**0.273**	**0.059**	**0.262**	**0.280**	**0.050**	**0.271**
0.45	**0.291**	**0.047**	**0.287**	**0.331**	**0.043**	**0.317**
0.50	**0.327**	**0.036**	**0.306**	**0.344**	**0.038**	**0.324**
0.55	**0.340**	**0.028**	**0.319**	**0.359**	**0.036**	**0.341**
0.60	**0.356**	**0.039**	**0.334**	**0.367**	**0.040**	**0.352**

At the training stage of the fusion model, the batch size is 32 with 30 epochs, and the learning rate is 1×10^{-4} for CAS(ME)^2 and SAMM. We used Adam as an optimizer. Introduce focal loss to balance positive and negative samples. From the loss curve in Figure 5, it can be seen that the network model convergence. At the evaluation stage, top_threshold is used as the evaluation index. Table II shows the F1_score under different thresholds. The experimental thresholds are 0.30:0.05:0.6. Since Top_threshold is based on probability, the larger the threshold, the more expressions are output. The number of positive samples (micro-expressions) in the dataset only accounts for a small part, and with the increase of experimental samples, the number of negative samples also increases.

It can be seen from Table I that by reducing the requirement for the coincidence rate, that is, when the threshold is reduced, the F1-score score is greatly improved. Indicates that the detected expression may be associated with a label whose IOU is less than the default threshold.

TABLE III
COMPARING THE F1-SCORE OF DIFFERENT MODELS IN $CAS(ME)^2$ AND SAMM DATASET.

Models	CAS(ME)^2			SAMM		
	MaE	ME	All	MaE	ME	All
Zhang [25]	**0.213**	**0.055**	**0.140**	**0.133**	**0.073**	**0.100**
MESNet [24]	**-**	**-**	**0.036**	**-**	**-**	**0.088**
Yap [26]	**0.216**	**0.025**	**0.142**	**0.192**	**0.043**	**0.107**
He [27]	**0.417**	**0.120**	**0.353**	**0.411**	**0.235**	**0.370**
LSSNet [28]	**0.380**	**0.063**	**0.327**	**0.336**	**0.218**	**0.290**
Proposed	**0.390**	**0.104**	**0.345**	**0.382**	**0.240**	**0.367**

The comparison with other model data is listed in Table III, which shows that the proposed method has a certain improvement in the field of ME detection using deep learning.

V. CONCLUSION

This paper proposes a ME spotting method based on 3D network and long video understanding. This method mainly focuses on the weighted fusion of cross-scale features, and adds a self-attention mechanism when calculating the weights, so that it pays more attention to the input of the feature layer. Although this method has achieved excellent results, there is still a certain gap compared with methods based on optical flow and motion units. There are two reasons. Based on the optical flow and facial motion unit, the extracted optical flow area is greatly reduced, and only the optical flow of 14 regions of interest is extracted. In addition, limited by the characteristics of the deep learning model, Training with a large amount of data, the method still needs to be improved.

ACKNOWLEDGMENT

The presented work is partially funded by the Postgraduate Innovation Project of CQUST (No. CX2021003 and No.CX2021021). The authors would like to thank the anonymous reviewers for their valuable suggestions and comments.

REFERENCES

[1] P. Ekman, W. V. Friesen. Nonverbal leakage and clues to deception[J]. Psychiatry, 1969, 32(1): 88-106.

[2] W. J. Yan, Q. Wu, J. Liang, et al. How fast are the leaked facial expressions: The duration of micro-expressions[J]. Journal of Nonverbal Behavior, 2013, 37(4): 217-230.

[3] P. Ekman, Darwin, deception, and facial expression[J]. Annals of the new York Academy of sciences, 2003, 1000(1): 205-221.

[4] P. Ekman. Lie catching and microexpressions[J]. The philosophy of deception, 2009, 1(2): 5.

[5] X. B. Li, X. P. Hong, A. Moilanen, et al. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods[J]. IEEE transactions on affective computing, 2017, 9(4): 563-577.

[6] A. Davison, W. Merghani, C. Lansley, et al. Objective micro-facial movement detection using facs-based regions and baseline evaluation. IEEE international conference on automatic face & gesture recognition. 2018: 642-649.

[7] D. Patel, G. Zhao, M. Pietikäinen. Spatiotemporal integration of optical flow vectors for micro-expression detection. International conference on advanced concepts for intelligent vision systems. Springer, Cham, 2015: 369-380.

[8] S. J. Wang, S. Wu, X. Qian, et al. A main directional maximal difference analysis for spotting facial movements from long-term videos[J]. Neurocomputing, 2017, 230: 382-389.

[9] Y. Han, B. Li, Y.-K. Lai, et al. "CFD: A collaborative feature difference method for spontaneous micro-expression spotting," in Proc. 25th IEEE Int. Conf. Image Process. (ICIP), Oct. 2018, pp. 1942–1946.

[10] Z. Zhang, T. Chen, H. Meng, et al. SMEConvNet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos[J]. IEEE Access, 2018, 6: 71143-71151.

[11] F. Qu, S. J. Wang, W. J. Yan, et al. CAS $(ME)^2$: a database for spontaneous macro-expression and micro-expression spotting and recognition[J]. IEEE Transactions on Affective Computing, 2017, 9(4): 424-436

[12] C. H. Yap, C. Kendrick, M. H. Yap. Samm long videos: A spontaneous facial micro-and macro-expressions dataset. IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). 2020: 771-776.

[13] L. W. Zhang, J. Li, S. J. Wang, et al. Spatio-temporal fusion for macro-and micro-expression spotting in long video sequences. IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). 2020: 734-741.

[14] J. Donahue, L. Anne Hendricks, S. Guadarrama, et al. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2625-2634.

[15] S. W. Ji, W. Xu, M. Yang, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 221-231.

[16] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems, pages 568–576,2014.

[17] J. Carreira, A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[18] M. Tan, R. Pang, Q. V. Le. Efficientdet: Scalable and efficient object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[19] X. Wang, R. Girshick, A. Gupta, et al. Non-local neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

[20] Z. Liu, J. Ning, Y. Cao, et al. Video swin transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3202-3211.

[21] T. Y. Lin, P. Dollár, R. Girshick, et al. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[22] S. Liu, L. Qi, H. Qin, et al. Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[23] G. Ghiasi, T. Y. Lin, Q, V. Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7036-7045.

[24] S. J. Wang, Y. He, J. Li, et al. MESNet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos[J]. IEEE Transactions on Image Processing, 2021, 30: 3956-3969.

[25] L. W. Zhang, J. Li, S. J. Wang, et al. Spatio-temporal fusion for macro-and micro-expression spotting in long video sequences. IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). 2020: 734-741.

[26] C. H. Yap, M. H. Yap, A. K. Davison, et al. 3d-cnn for facial micro-and macro-expression spotting on long video sequences using temporal oriented reference frame[J]. arXiv preprint arXiv:2105.06340, 2021.

[27] H. Yuhong. Research on micro-expression spotting method based on optical flow features. Proceedings of the 29th ACM International Conference on Multimedia. 2021: 4803-4807.

[28] W. W. Yu, J. Jiang, Y. J. Li. 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. 2021: 4745-4749.