Integrating VideoMAE based model and Optical Flow for Micro- and Macro-expression Spotting

Ke Xu[†*]
University of Chinese Academy of Sciences, China
The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences,
xuke2021@ia.ac.cn

Kang Chen[†]
Peking University, China
The State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences,
2201210163@stu.pku.edu.cn

Licai Sun
University of Chinese Academy of Sciences, China
The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences,
sunlicai2019@ia.ac.cn

Zheng Lian
The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China
lianzheng2016@ia.ac.cn

Bin Liu
The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, University of Chinese Academy of Sciences, China
liubin@nlpr.ia.ac.cn

Gong Chen
University of Chinese Academy of Sciences, China
The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences,
chengong2021@ia.ac.cn

Haiyang Sun
University of Chinese Academy of Sciences, China
The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences,
sunhaiyang2021@ia.ac.cn

Mingyu Xu
The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China
xumingyu2021@ia.ac.cn

Jianhua Tao
Department of Automation, Tsinghua University, China
Beijing National Research Center for Information Science and Technology, Tsinghua University
jhtao@tsinghua.edu.cn

ABSTRACT

The task of interval localization of macro- and micro-expression in long videos has a wide range of applications in the field of human-computer interaction. Compared with macro-expression, micro-expression has shorter duration, lower intensity, and smaller number of samples, which make them more difficult to spot accurately in long videos. In this paper, we propose a pre-trained model combined with the optical flow method to improve the accuracy and robustness of macro- and micro-expression spotting. Firstly, self-supervised pre-training is performed on rich unlabeled data based on VideoMAE. Then, multiple models are trained on the datasets SAMM-LV and CAS(ME)^3 for macro- and micro-expression with different fine-grains. Finally, different lengths of slices are generated based on the models with different fine-grains, and the optimal matching method through the combination of model fine-grainedness and slice lengths is explored. At the same time, macro- and micro-expression generating regions were spotted using the optical flow method, fused with the model outputs to supplement the spatio-temporal information not captured by the model and to exclude the interference of non-interested regions. We evaluated the performance of our method on the MEGC2023 testset (consisting of 10 long videos from SAMM and 20 long videos from CAS(ME)^3) and won first place in the MEGC2023 Challenge. The results demonstrate the effectiveness of the method.

CCS CONCEPTS

• **Computing methodologies →Computer graphics.**

KEYWORDS: Optical Flow, VideoMAE, Micro-expression, Macro-expression, Pre-training

ACM Reference format:
Ke Xu, Kang Chen, Licai Sun, Zheng Lian, Bin Liu, Gong Chen, Haiyang Sun, Mingyu Xu, and Jianhua Tao. 2023. Integrating VideoMAE based model and Optical Flow for Micro- and Macro-expression Spotting: In *Proceedings of the 31st ACM International Conference on Multimedia (MM'23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA. 5 pages. https://doi.org/ 10.1145/3581783.3612868

1 INTRODUCTION

Micro-expression is a spontaneous and brief facial expression produced when people try to hide their inner emotions. It can

† These authors are equal contribution to this work.

* Corresponding author

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

neither be camouflaged nor suppressed, and can reflect people's true inner emotions [1]. In general, tasks related to micro-expression include two main aspects: micro-expression spotting in a long video and emotion recognition in micro-expression clips [2]. The micro-expression spotting task is to predict the start and end time of micro-expression in a long video, while the micro-expression recognition task is to categorize micro-expression segments, such as positive, negative, neutral, etc. Micro-expression has the characteristics of reflecting people's true inner emotions and are difficult to disguise and inhibit. In recent years, micro-expression spotting and recognition have been widely used in crime investigation [3], communication and negotiation [4], and other fields. Compared with macro-expression, micro-expression has a duration of 1 /25 ~ 1 /3 s and a smaller movement amplitude [5]. It is more difficult to induce micro-expression, manual annotation of micro-expression requires more professional people, so the sample size is smaller. At the same time, in a long video, there will inevitably be blinking and shaking of the head and other factors to interfere with, which makes micro-expression spotting task challenging. Therefore, in order to solve the above problems, we propose to utilize the mutual integration of pre-trained models and optical flow method [6,7] to spot macro-expression and micro-expression segments in long videos. We use a pre-training approach based on VideoMAE [8] to alleviate the overfitting problem due to insufficient sample size of micro-expression. For the small intensity and short duration of micro-expression, the dense optical flow method can effectively capture the optical flow features between different frames, and the division of ROI (regions of interest) can effectively mitigate the interference of non-interested regions. The main contributions of this paper are summarized as follows.

1. We pre-train a VideoMAE based model, and fine-tuning on micro-expression dataset to spot macro- and micro-expression in long videos.
2. We explore the optimal combination method by training multiple models for macro and micro-expression with different fine-grainedness and generating different lengths of expression clips.
3. We propose a fusion strategy and post-processing method to complement the spatio-temporal information not captured by the model and to exclude the interference of non-interested regions.

The remaining sections of this paper are as follows: Section 2 briefly reviews the related studies on micro-expression spotting, Section 3 describes the methodology proposed in this paper, Section 4 gives the evaluation criteria and experimental results, Section 5 analyzes the experimental results, and Section 6 makes conclusions.

2 RELATED WORK

Spotting micro-expression with the naked eye is a great challenge. Although Ekman invented the Micro-Expression Training Tool (METT) [9] to train people to analyze micro-expression, this method is very time-consuming and not very accurate. In the early stages of micro-expression research, manual labeling based on the FACS coding system was used [10]. However, manual coding is laborious and time-consuming. Therefore, in recent years, there has been an increasing demand for automatic spotting of facial micro-expression relying on computers.

Shreve et al. [11,12] used the optical flow method [13] to calculate the optical strain magnitude and discover macro- and micro-expression. Li et al. used local temporal patterns (LTP-ML) [14] for spontaneous micro-expression spotting. Wang et al. proposed convolutional networks for discovering multi-scale micro-expression segments in long videos (MESNet) [15]. Pan [16] employed a deep learning technique in which the authors utilized a BCNN structure to extract global and local features of face regions from each frame of a long video sequence. However, deep learning methods suffer from small sample data. Li et al. [17] proposed a multi-branch self-supervised learning method for micro-expression spotting based on human attention mechanism. As a baseline for the FME Challenge 2021, Yap et al. [18] proposed a 3D-CNN using frame-skipping and contrast enhancement. In the third Micro-Expression Grand Challenge (MEGC2022), Yap et al. [19] proposed a pure deep learning solution, which instead of tracking frame differential motion, but rather compares them through a convolutional model that is used as a baseline for MEGC2021. Among these challenges, much work has been done to spot macro-expression and micro-expression intervals in long video sequences.

3 METHODOLOGY

As shown in the general framework diagram in Figure 1, our method is divided into 3 parts: dataset preprocessing, self-supervised training based on VideoMAE [20], Interval fusion and post-processing strategies. In this section, we describe each part in detail.

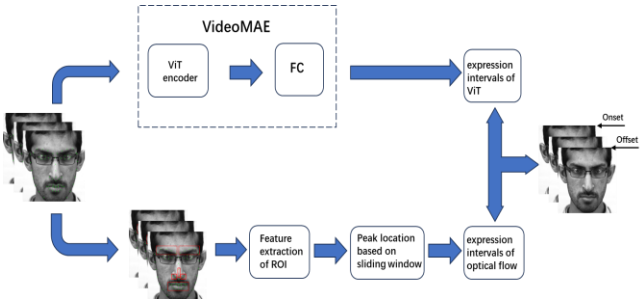

Figure 1: Proposed macro- and micro-expression spotting framework.

3.1 Dataset processing

Since macro- and micro-expression spotting tasks generally focus only on the face region and other regions in the video are disturbances, the dataset is preprocessed to remove some of the perturbations. For the training sets and testsets, we use the Dlib toolkit [21] to spot the facial landmarks in each frame and crop out the faces based on the landmark points. In detail, we select the first image in the folder that has one and only one face, use Dlib toolkit to locate the facial landmarks to determine the appropriate face cropping frame, and use the face cropping frame to crop the subsequent images. This can save computational resources. Figure 2 shows one of the cropped images. During the experiment we found that CAS(ME)^3 part of the data has no human face or two faces, and we eliminated the disturbance of this part of the data by face detector [30].

Figure 2: Face Cropping Alignment

3.2 Self-supervised training based on VideoMAE

As shown in the general framework diagram in Figure 3, VideoMAE [20] is a self-supervised video pre-training method based on a video masking self-encoder [22,23]. It utilizes the temporal dimension of video as the temporal evolution of a still image and addresses semantic redundancy and temporal correlation in video. VideoMAE efficiently captures high-level semantic information in video by employing both high masking ratios and masking pipelines. It provides a promising solution for training video transformers.

Specifically, VideoMAE consists of four main modules: temporal block embedding, pipe masking, high-capacity encoder, and lightweight decoder. Given an original video, a pixel of size 2*16*16 (2 represent two adjacent frames) in a video clip is considered as a pixel block to alleviate the spatio-temporal redundancy of the video data. Then the pipe masking module generates a mask $M \in \{0,1\}$, k with a mask rate p= 90%, and the high-capacity encoder Φ_e simply performs a global spatio-temporal self-concern process by taking only the unmasked token as input. Subsequently, the lightweight decoder Φ_d reconstructs the original video data by combining the encoded visible tokens with the learnable masked tokens. Finally, the mean-square error between the original and reconstructed video at the mask position is computed to optimize the whole model. The above process can be generally formulated as follows:

$$L_{VideoMAE} = MSE\left(\Phi_d\left(\Phi_e(X \odot M)\right),\ V \odot \Psi(1 - M)\right) \quad (1)$$

where Ψ is the function used to obtain the mask position in pixel space. In the downstream task, the lightweight decoder Φ_d is discarded and only the high-volume ViT encoder Φ_e will be fine-tuned.

We use a large audiovisual speaker recognition dataset called VoxCeleb2 [24] for self-supervised pre-training [25] based on VideoMAE. The dataset consists of more than 1 million video clips from more than 6000 celebrities extracted from about 150,000 interview videos on YouTube. It is divided into a development set and a testset. For pre-training, we exclusively use the development set, which contains 10,929 video clips from 145,569 videos. For fine-tuning, we trained the model on macro and micro expressions of different fine-grained sizes (e.g., 2s, 4s), generating different lengths of expression segments (e.g., 2s, 4s) with the same input size of 16x160x160 as in the pre-training phase, and we tried to find the best combination for each subject by combining the models of different fine-grained expressions with the generated clips of different lengths.

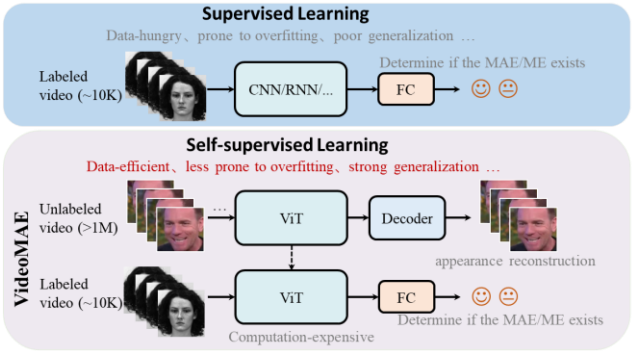

Figure 3: macro- and micro-expression spotting framework based on VideoMAE.

3.3 Interval fusion and post-processing strategies

In the fusion and post-processing stage, the interval overlap is first compared, and if the micro-expression interval is within the range of the macro-expression interval, the macro-expression interval is retained, and if the micro-expression interval is not within the range of the macro-expression interval, the micro-expression interval is directly retained. Then a post-processing strategy is that if the IOU between the expression interval output by the model and the expression interval output by the optical flow method [29] is greater than or equal to a certain value, it is retained. The remaining model output intervals are subjected to blink spotting with dlib to exclude interference caused by pure blinks. Another strategy is that the model output intervals are first subjected to dlib blink spotting to exclude the interference caused by pure blinks, and then fused with the optical flow intervals. Finally, we combined the results after the various post-processing strategies and adopted them.

4 EXPERIMENTS

4.1 Datasets

There are two main training sets for MaE and ME spotting tasks used in this paper: CAS(ME)^3 [26] and SAMM Long Videos [27]. CAS(ME)^3 provides about 80 hours of videos with more than 8 million frames, including 1,109 manually labeled micro-expressions and 3,490 macro-expressions. The large sample size of CAS(ME)^3 can be efficiently performed for micro-expression analysis method validation. expression analysis method validation, which is important for AI-based micro-expression analysis. SAMM Long Videos, abbreviated as SAMM-LV, is an extension of the SAMM dataset. It consists of 147 long videos with 32 themes, including 343 macro-actions and 159 micro-actions. CAS(ME)^3 records at a frame rate of 30 fps and a resolution of 1280 × 720, while SAMM-LV records at a frame rate of 200 fps and a resolution of 2040 × 1088, with a face size of 400 x 400 in the video. The invisible testset (MEGC2023-testset) contains 30 long videos, including ten clips from SAMM-LV and 20 clips cropped from different videos in CAS(ME)^3 (previously unreleased).

4.2 Performance Metrics

The Intersection of Unions (IOU) method was applied to evaluate the performance of the existing proposed methods according to the

guidelines of the competition organizers. The true positive (TP) per interval in one video is first defined based on the intersection between the spotted interval and the ground-truth interval. The spotted interval is considered as TP if it fits the following condition :

$$\frac{W_{spotted} \cap W_{groundTruth}}{W_{spotted} \cup W_{groundTruth}} \quad (2)$$

where k is set to 0.5, represents the ground truth of the macro or micro-expression interval (onset-offset). If the condition is not fulfilled, the spotted interval is regarded as false positive (FP). We consider that each ground-truth interval corresponds to at most one single spotted interval.

Precision is the probability that a positive sample occurs out of all samples predicted to be positive. Recall is the probability of predicting a positive sample from actual positive samples. We use precision, recall, and F1 score to evaluate the performance of the method.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \quad (5)$$

4.3 Parameter settings

During pre-training, we extracted 16 frames from each video clip with a time step of 4. When using cubes, this produces 8 × 10 × 10 input tokens of size 2 × 16 × 16 after the cube embedding. Regarding the hyper-parameters, we mainly followed VideoMAE. Specifically, we used a $\beta 1 = 0.9$ and $\beta 2 = 0.95$ for the AdamW optimizer, with an overall batch size of 128, a base learning rate of $3e$-4, and a weight decay of 0.05. In addition, we use a cosine decay learning rate scheduler. By default, we pretrained the model 50 times and warmed it up 5 times. When using 4 Nvidia Tesla V100 GPU, the pre-training takes about 3-4 days. For fine-tuning, Macro- and micro-expression was modeled by expressions during 0.5-4s, generated expression segments with lengths of 1-1.5s at the same 16x160x160 input size as the pre-training stage. In the fusion and post-processing stage, we set the IOU to 0.5.

4.4 Results and Discussion

We use an unseen testset (MEGC2023-testset) to evaluate the performance of our method. The results are shown in Table 1. For CAS(ME)^3, the F1 score of our method in this paper is 0.22. For SAMM-LV, the F1 score is 0.37. The total score for both datasets is 0.22. Our method performs best in the challenge and outperforms the baseline [28].

Table 1: Experimental results of MEGC2023-testset.

	CAS(ME)^3	SAMM-LV	overall
F1-score	0.21	0.37	0.22
Precision	0.27	0.37	0.28
Recall	0.17	0.36	0.19

The effect of our method on the validation set is shown in Table 2, and it can be seen that the effect on the validation set is higher than that on the testset for both datasets, which is due to the better division of the labeled intervals during the training process. While the testset is not visible with labels and can only generate slices of different lengths according to the models with different granularities, another approach is to use the sliding window method to process the test data.

Table 2: Experimental results of two training sets CAS(ME)^3 and SAMM-LV.

	CAS(ME)^3	SAMM-LV
F1-score	0.27	0.38
Precision	0.38	0.52
Recall	0.21	0.30

5 CONCLUSION

We propose to utilize a pre-trained model combined with an optical flow method to spot macro-expression and micro-expression segments in long videos. The aim of the method is to automatically recognize micro-expression and macro-expression. According to the characteristics of micro-expression and macro-expression of different subjects, multiple models are trained using macro- and micro-expression with different fine-grain sizes and different lengths of expression segments are generated to accomplish the spotting task. The results of the proposed method show that our method is able to achieve the best results of the challenge on both datasets, outperforming the baseline. On the other hand, the method outperforms the CAS(ME)^3 dataset on the SAMM dataset, probably because CAS(ME)^3 is more perturbed than SAMM in the testsets. Therefore, in the future, we hope to investigate methods with greater robustness.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61831022, No.62276259, No.62201572, No. U21B2010, No.62271083), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (No. Z211100004821013), Open Research Projects of Zhejiang Lab (No. 2021KH0AB06), CCF- Baidu Open Fund (No. OF2022025).

REFERENCES

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

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

[3] Owayjan M, Kashour A, Al Haddad N, et al. The design and development of a lie detection system using facial micro-expression//2012 2nd international conference on advances in computational tools for engineering applications (ACTEA). IEEE, 2012: 33-38.

[4] Whitehill J, Serpell Z, Lin Y C, et al. The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing, 2014, 5(1): 86-98.

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

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

[7] Yu J, Cai Z, Liu Z, et al. Facial expression spotting based on optical flow features//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 7205-7209.

[8] Tong Z, Song Y, Wang J, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems, 2022, 35: 10078-10093.

[9] P. Ekman, Micro Expression Training Tool (METT). San Francisco, CA,USA: Univ. California, 2002.

[10] P. Ekman and W. V. Friesen, "Facial action coding system (FACS): a technique for the measurement of facial actions," Rivista Di Psichiatria, vol. 47, no. 2, pp. 126–38, 1978.

[11] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro- and micro-expression spotting in video using strain patterns,'' in Proc. IEEE Conf. Appl. Comput Vis., Dec. 2009, pp. 1–6.

[12] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, ''Macro- and micro expression spotting in long videos using spatio-temporal strain,'' in Proc. IEEE Conf. Auto Face Gesture Recognit., Mar. 2011, pp. 51–56.

[13] M. J. Black and P. Anandan, ''The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields,'' Comput. Vis. Image Understand., vol. 63, no. 1, pp. 75–104, 1996.

[14] J. Li, C. Soladie, R. Seguier, S-. J. Wang and H. Y. Moi. Spotting micro expressions on long videos sequences. IEEE International Conference on Automatic Face & Gesture Recognition (FG). 2019, pp: 1-5.

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

[16] H. Pan. Local bilinear convolutional neural network for spotting macro- and micro-expression intervals in long video sequences. In IEEE International Conference on Automatic Face and Gesture Recognition, 2020.

[17] LI J, DONG Z, LIU Y, et al. Micro-expression spotting method based on human attention mechanism. Advances in Psychological Science, 2022, 30(10): 2143.

[18] Chuin Hong Yap, Moi Hoon Yap, Adrian K. Davison, and Ryan Cunningham. 2021.Efficient Lightweight 3D-CNN using Frame Skipping and Contrast Enhancement for Facial Macro- and Micro-expression Spotting. CoRR abs/2105.06340 (2021).arXiv:2105.06340 https://arxiv.org/abs/2105.06340

[19] Yap C H, Yap M H, Davison A, et al. 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. 2022: 7016-7020.

[20] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In Advances in Neural Information Processing Systems.

[21] Davis E King. 2009. Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research 10 (2009), 1755–1758.

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16000–16009.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).

[24] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. Proc. Interspeech 2018 (2018), 1086–1090.

[25] Sun L, Lian Z, Liu B, et al. MAE-DFER: Efficient Masked Autoencoder for Self-supervised Dynamic Facial Expression Recognition. arXiv preprint arXiv:2307.02227, 2023.

[26] Li, J., Dong, Z., Lu, S., Wang, S.J., Yan, W.J., Ma, Y., Liu, Y., Huang, C. and Fu, X. (2023). 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, no. 3, pp. 2782-2800, 1 March 2023, doi: 10.1109/TPAMI.2022.3174895.

[27] Davison, A. K., Lansley, C., Costen, N., Tan, K., & Yap, M. H. (2016). SAMM: A spontaneous micro-facial movement dataset. IEEE Transactions on Affective Computing, 9(1), 116-129.

[28] Zhang, L. W., Li, J., Wang, S. J., Duan, X. H., Yan, W. J., Xie, H. Y., & Huang, S. C. (2020, November). Spatio-temporal fusion for macro-and micro-expression spotting in long video sequences. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 734-741). IEEE.

[29] Yu J, Cai Z, Liu Z, et al. Facial expression spotting based on optical flow features//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 7205-7209.

[30] Wu, W., Peng, H. & Yu, S. YuNet: A Tiny Millisecond-level Face Detector. Mach. Intell. Res. (2023). https://doi.org/10.1007/s11633-023-1423-y