Facial Expression Localization and Recognition Using MDMO and Transformer for Mental Health Diagnosis Support

Yan Yan[1], Hanjun Li[2*]
[1]Guangdong Polytechnic of Industry and Commerce, Guangzhou 510510, China
[2]Guangzhou College of Commerce, Guangzhou 511363, China
E-mail: nyyanyan2008@126.com, LiHanjun_lhj@outlook.com
[*]Corresponding author

Keywords: expression localization, mental health auxiliary diagnosis and treatment, transformer, MDMO, facial expression recognition

Received: September 25, 2025

With the work and study pressure on people increasing and the importance of mental health problems on the rise, facial expression analysis plays an important role in mental health auxiliary diagnosis and treatment (MHADT). This study proposes a facial expression localization and recognition model that integrates Main Directional Mean Optical Flow (MDMO) with the Transformer architecture. It addresses the problems of insufficient generalization ability and limited temporal modeling ability of traditional methods in facial expression recognition. The study is based on the authoritative Audio/Visual Emotion Challenge 2019 – Depression Detection Sub-challenge (AVEC2019 DDS) dataset in the mental health field, which contains 163 training samples, 56 validation samples, and 56 test samples. With Explicit Shape Regression, Local Binary Patterns, Mnemonic Descent Method, Convolutional Neural Network, and benchmark models as comparison objects, it systematically evaluates the model's performance in error, accuracy, and processing speed. The results show that the model achieves the best performance in both facial expression localization and recognition tasks. The validation set errors are 7.26 and 6.85, the localization accuracy reaches 89.6%, and the recognition accuracy reaches 88.9%, which is significantly better than other methods. At the same time, its image processing time is 55 milliseconds (ms) and 44ms, balancing high precision and real-time performance. The study indicates that the fusion of MDMO and Transformer can effectively capture the spatial and temporal features of facial expressions. Thus, this fusion method provides an efficient, stable, and scalable technical solution for emotion recognition in MHADT. This improves both the FER and localization accuracy and effect. Besides, it provides a novel reference in both approaches and application level for the intelligent development of mental health evaluation.

Povzetek: Predstavjen Model MDMO + Transformer natančno in hitro prepoznava obrazne izraze za podporo oceni duševnega zdravja.

1 Introduction

With the increasing severity of global mental health problems, traditional face-to-face psychotherapy faces challenges such as resource shortage, geographical restrictions and high costs [1]. According to statistics from the World Health Organization, nearly 1 billion people worldwide suffer from mental illnesses, whereas the number of people receiving professional treatment is far lower than the demand. Especially in low- and middle-income countries, the accessibility of mental health services is more limited [2-3]. Against this background, digital mental health intervention methods have emerged as an important way to alleviate this problem. For example, smartphone applications have been proven to effectively reduce symptoms of depression and anxiety and improve users' psychological well-being by providing self-assessment, cognitive behavioral therapy courses, and relaxation training [4]. However, existing digital intervention tools mostly rely on text or voice input; these tools lack real-time perception of users' facial expressions and emotional changes, which limits their application effects in complex emotion recognition and intervention.

Recent years have witnessed fast advancement in artificial intelligence technology, particularly with the emergence of multimodal deep learning (DL) models. These developments are steering digital mental health interventions toward increasingly intelligent and personalized approaches [5-7]. Multimodal DL methods can process multiple input information such as text, voice,

and vision at the same time, realizing comprehensive perception and analysis of users' emotional states. For example, models based on the Transformer architecture have achieved remarkable results in natural language processing tasks; their strong context modeling ability provides new ideas for emotion recognition and psychological intervention [8]. These methods perform well in laboratory environments. However, they still face many challenges in practical applications, such as the real-time performance of the model, generalization ability, and adaptability to small-sample data. Therefore, how to improve the real-time performance and generalization ability of the model while ensuring high accuracy has become a hot and difficult issue in current research.

This study explores the application of the expression localization and recognition model based on Main Directional Mean Optical Flow (MDMO) and Transformer architecture in mental health auxiliary diagnosis and treatment (MHADT). Specifically, this study proposes an expression localization and recognition model combining MDMO and Transformer, thus improving the model's localization and emotion recognition accuracy in MHADT scenarios. Through experiments on the Audio/Visual Emotion Challenge 2019 – Depression Detection Sub-challenge (AVEC2019 DDS) dataset, the model's superiority in terms of localization accuracy, recognition accuracy, and processing time is verified. In addition, this study conducts statistical significance tests to verify the model's reliability and performance improvement. The innovation of this study lies in combining MDMO with the Transformer architecture for the first time. It proposes a new expression localization and recognition model, which provides a new technical means for MHADT. The research results show that the model can effectively improve the accuracy of expression localization and emotion recognition, providing strong support for developing digital mental health intervention tools. This study enriches the application scenarios of multimodal DL in mental health; it also offers a reference for designing and optimizing intelligent mental health intervention systems in the future.

2 Related work

Facial expression recognition (FER) research in the MHADT field has become a hot direction at the intersection of AI and medicine [9]. Domestic and international scholars have extensively explored model optimization, feature extraction, and clinical adaptability, forming rich research results. However, there is still room for improvement in the applicability of complex diagnosis and treatment scenarios.

Foreign research started relatively early and led in both theoretical system construction and technological innovation. Kondal et al. studied the "basic emotion theory" and defined six cross-culturally universal basic expression categories: happiness, anger, sadness, fear, surprise, and disgust. The facial action coding system they developed provided a standardized feature annotation basis for FER by labeling 46 facial muscle movement units. This theory remains the core theoretical support for most FER models to this day. Over the years, with the development of DL technology, Jabbooree et al. constructed an FER model based on a Convolutional Neural Network (CNN). This model achieved a classification accuracy of 89.2% on the FER-2013 dataset. Still, it only concentrated on the judgment of a single expression category and did not involve the associated inference of psychological states [10]. Kim et al. attempted to apply the Transformer to expression analysis and captured global facial features through a self-attention mechanism. Although this approach improved the sensitivity of micro-FER, it did not consider the problem of disguised expressions caused by patients' social concealment psychology, resulting in insufficient robustness in clinical diagnosis and treatment scenarios [11].

Domestic research focuses more on the adaptability of technology to local clinical scenarios and has formed unique advantages in multi-modal fusion and lightweight model development. Zhang and Chai proposed a FER scheme based on "visual-physiological" multimodal fusion. They constructed a multi-feature fusion model by synchronously collecting facial images and heart rate variability signals. The experimental results showed that the accuracy of this model in the expression analysis of patients with depressive tendencies was 15.7% higher than that of the single visual feature model. Their research pointed out that "single visual features could not fully reflect psychological states. Besides, introducing physiological signals could effectively make up for the recognition bias caused by expression disguise", which provided a key idea for multi-modal research in MHADT [12]. In response to the current situation of limited hardware resources in primary medical scenarios in China, Zhang et al. presented a lightweight FER model. Through model pruning and quantization technologies, they compressed the model parameter size from 23.6 megabytes (MB) of the traditional CNN to 4.8MB; meanwhile, it maintained an expression classification accuracy of 82.1%, and successfully realized deployment on primary medical terminals (such as portable diagnostic instruments). However, this model still had distinct limitations and could only complete the single task of "expression classification"; it could not simultaneously realize the localization of core expression areas (such as identifying the frontal muscle movement area corresponding to "frowning") and the inference of psychological states. Its output results were limited to categorical labels such as "anger" and "happiness." This limitation created a serious disconnect from the multi-link decision-making needs of clinical physicians, which involved locating abnormal expression areas, judging

emotional types, and assessing psychological risks. Consequently, the system struggled to function effectively in auxiliary diagnosis and treatment (ADT) contexts [13]. Zhao et al. attempted to construct a multi-task expression analysis model to simultaneously optimize the tasks of expression classification and psychological state assessment. However, this model adopted a traditional multi-task learning framework and did not design a special feature interaction mechanism for mental health scenarios, leading to interference in feature extraction between the two tasks. The Mean Absolute Error (MAE) for psychological state assessment reached 0.87, failing to meet the clinical standards for evaluation accuracy [14].

To sum up, although existing studies have made progress in FER accuracy and model lightweight, there are still obvious limitations. Most models still focus on single-task and single-dimensional analysis, which makes it difficult to meet the comprehensive needs of the "localization - recognition - inference" link in MHADT. In addition, existing methods lack a multi-task optimization framework designed for mental health scenarios, leading to a certain disconnect between model output and clinical practical application. This restricts the practicality and reliability of these methods in ADT. According to this, the study introduces an expression localization and recognition scheme that fuses MDMO and Transformer. It aims to break through the limitations of existing research and improve the accuracy and practicality of expression analysis in MHADT.

3 The expression localization and recognition model using MDMO and transformer in MHADT

This study aims to verify whether combining MDMO with the Transformer structure can significantly improve the accuracy and stability of FER in MHADT scenarios. The study explores whether this fusion model can more effectively capture the facial expressions' dynamic changes and temporal dependencies. On the clinical AVEC2019 DDS dataset, the model's comprehensive performance is assessed across accuracy, error rate, and processing speed through comparisons with Explicit Shape Regression (ESR), Local Binary Patterns (LBP), Mnemonic Descent Method (MDM), CNN, and benchmark models. These evaluations confirm its feasibility and application potential for clinical psychological emotion recognition.

3.1 The MDMO feature and transformer model

(1) MDMO feature

The MDMO feature is a technical method designed for extracting changing features of expressions [15]. By calculating the main direction and average speed of pixel movement in facial regions, it can accurately capture the core temporal information of changes in expressions [16]. It can filter out interference from irrelevant movements such as head shaking and light changes while highlighting key motion features related to emotions [17]. This effectively makes up for the defects of traditional static features (which lose temporal information) and ordinary optical flow technology (which has computational redundancy). Thus, it can provide better real-time feature support for analyzing weak and mixed emotional expressions in MHADT scenarios. It is an important foundation for the subsequent realization of accurate expression localization and recognition by combining with the Transformer model [18].

This study selects the direction interval $B_{\max}$ that contains the largest number of optical flow vectors, and calculates the average value of the optical flow vectors within $B_{\max}$:

$$\bar{u}_i^k = \frac{1}{\left|B_{\max}\right|} \sum_{u_i^k \in B_{\max}} u_i^k(p) \tag{1}$$

k=1,2,…,36 represents the serial number of the Region of Interest (ROI); $u_i^k(p)$ refers to the pixel point belonging to the k-th ROI (ROI_i^k) in the i-th frame; p denotes the optical flow vector. Then, these vectors in each ROI are connected to obtain the i-th frame's feature vector:

$$\psi_i = \left(\bar{u}_i^1, \bar{u}_i^2, \ldots, \bar{u}_i^{36}\right) \tag{2}$$

Next, the average feature vector of the entire micro-expression sequence is calculated and converted into polar coordinates:

$$\bar{\psi} = \left[\left(\bar{\rho}_1, \bar{\theta}_1\right)^T, \left(\bar{\rho}_2, \bar{\theta}_2\right)^T, \ldots, \left(\bar{\rho}_{36}, \bar{\theta}_{36}\right)^T\right] \tag{3}$$

Since the main direction's intensity varies among different micro-expression sequences, further normalization of the magnitude ρ of Ψ is required:

$$\rho_k = \frac{\bar{\rho}_k}{\max\left\{\bar{\rho}_j, j = 1, 2, \ldots, 36\right\}} \tag{4}$$

Finally, the normalized MDMO feature of a micro-expression sequence can be expressed as:

$$\psi = \left[\left(\rho_1, \theta_1\right)^T, \left(\rho_2, \theta_2\right)^T, \ldots, \left(\rho_{36}, \theta_{36}\right)^T\right] \tag{5}$$

To balance the influence of the magnitude ρ and the direction θ, a weight parameter λ is also introduced, and the MDMO feature is rewritten as:

$$\bar{\bar{\psi}} = \left(\lambda P, (1-\lambda)\Theta\right) \tag{6}$$

P denotes the magnitude part: P=(ρ1, ρ2, …, ρ36\); Θ represents the direction part: Θ=(θ1, θ2, …, θ36).

(2) Transformer Model

The overall structure of Transformer mainly consists of an encoder, a decoder, as well as related attention mechanisms and fully connected layers [19]. Its upper layer includes units composed of self-multi-head attention, a fully connected feedforward network, and residual normalization modules, which are used for feature extraction and information processing [20-21]. The middle part is the main body of the encoder and decoder. The encoder is formed by connecting multiple encoder units in series to encode the input content; the decoder consists of many sequentially connected decoder units that progressively transform the encoder's output into the final generated result. At the same time, it transmits auxiliary information between the Encoder and Decoder through "Focused Information". The lower layer includes a fully connected feedforward network, self-multi-head attention, encoder-decoder attention, and residual normalization modules to assist the decoder in achieving accurate output. The overall structure is progressive, and the attention mechanism and residual normalization ensure the efficiency and accuracy of the Transformer in sequence information processing [22]. The Transformer's overall structure is presented in Fig. 1:

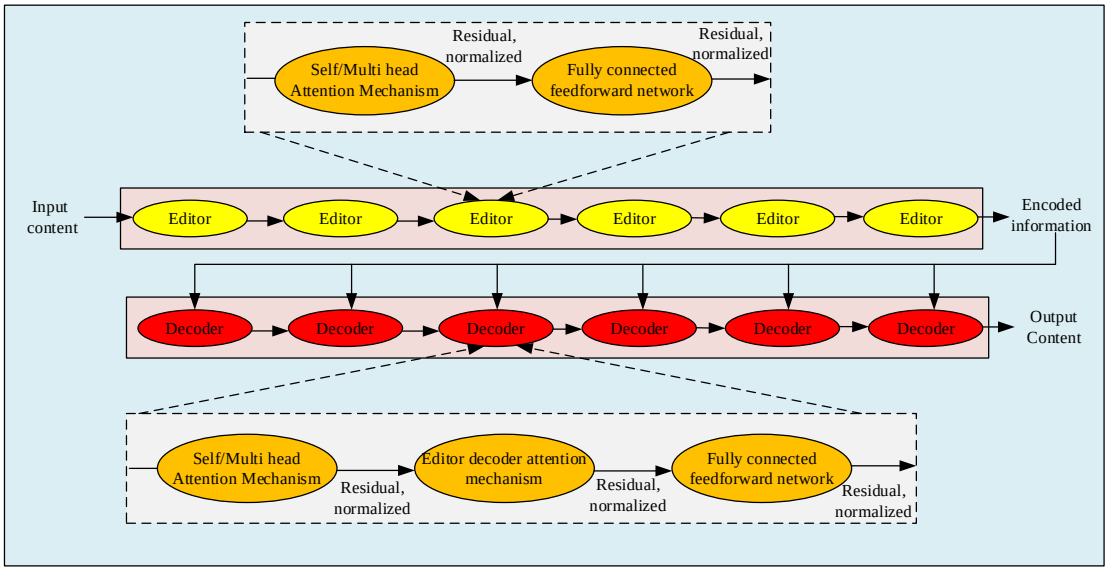

Figure 1: The overall structure of the Transformer

In Fig. 1, the encoder processes the input sequence, while the decoder is responsible for generating the output. The most important component in both the encoder and the decoder is the attention mechanism. This mechanism's implementation is called "Scaled Dot-Product Attention", and its calculation process is as follows:

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

Q, K, and V represent Query, Key, and Value, respectively.

Equation (8) presents the parallel modeling process of multi-head attention:

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^o \tag{8}$$

Equation (9) defines the Feed Forward Network (FNN)-based nonlinear feature transformation:

$$where head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{9}$$

All variables denoted as W in Equations (8) and (9) are parameter matrices; W^o refers to the matrix for merging multiple heads; W indicates the parameter matrix used to obtain the projected Q, K, and V; *Concat* stands for the concatenation operation.

3.2 The expression localization model grounded on MDMO and Transformer in MHADT

The expression localization model based on MDMO and Transformer in the proposed MHADT has a four-layer structure. The input layer receives face images and performs image preprocessing; the feature extraction layer sequentially completes optical flow estimation, optical flow direction calculation, and MDMO feature extraction; the feature modeling layer encodes the extracted features and processes them through a

Transformer Encoder containing self-attention mechanism, multi-head attention, and feedforward neural network; the localization output layer first predicts key facial expression regions and finally outputs the expression localization results. The model is progressive from image input to feature processing and then to localization output; it also presents the complete process of achieving accurate facial expression localization in MHADT based on MDMO and Transformer technologies [23]. The expression localization model's hierarchical structure by the Transformer and MDMO in MHADT is shown in Fig. 2:

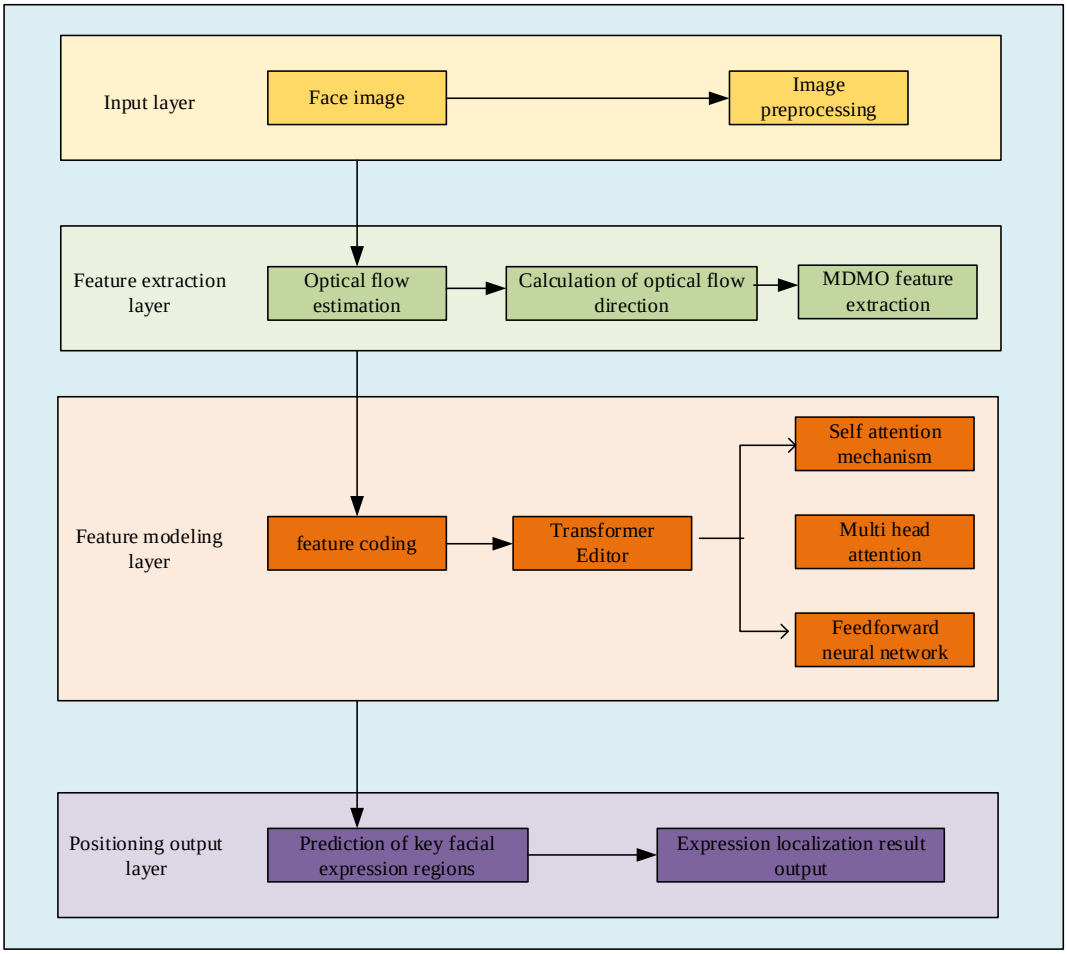

Figure 2: Hierarchical diagram of the Transformer and MDMO-based expression localization model in MHADT

First, this study uses the optical flow method to obtain the target facial movement information and extract directional features. Second, it models the dominant directional features of expression changes through MDMO. Finally, it combines a self-attention mechanism of the Transformer and the FNN to realize the accurate localization and discrimination of key expression regions.

(1)Calculation of optical flow vectors

$$v(x,y,t)=\left(u(x,y,t),v(x,y,t)\right) \tag{10}$$

$v(x,y,t)$ represents the optical flow vector of the pixel point (x,y) at time t, including the horizontal component u and the vertical component v. This step extracts the movement information of facial expressions that change over time.

(2)Calculation of optical flow direction angle

$$\theta(x,y,t)=\arctan\left(\frac{v(x,y,t)}{u(x,y,t)}\right) \tag{11}$$

The movement's main direction information is obtained by calculating the optical flow vector's direction angle θ. This provides a basis for the subsequent construction of the direction histogram and the extraction of MDMO features.

(3) MDMO

$$MDMO_k=\frac{1}{N_k}\sum_{(x,y,t)\in D_k} v(x,y,t) \tag{12}$$

D_k represents the k-th direction interval; N_k stands for the number of optical flow vectors in this interval. In $MDMO_k$, O denotes the mean optical flow feature of this

direction interval. Through Equation (12), the most representative main directional movement features in expression changes can be extracted.

(4) Self-attention mechanism modeling

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (13)$$

Q, *K*, and *V* respectively represent the query, key, and value vectors encoded by MDMO features, and *dk* denotes the feature dimension. Equation (13) is employed to capture the global dependency relationship of expression features among various regions.

(5) Localization prediction output

$$\hat{y} = FFN\left(Attention(Q,K,V)\right) \quad (14)$$

The attention result is subjected to a nonlinear transformation through an FFN; the predicted position or feature representation of the key expression region is output, thereby realizing the accurate localization of expressions.

3.3 The FER model using MDMO and transformer in MHADT

In MHADT, the FER model integrating MDMO and Transformer focuses on realizing multi-level emotion discrimination based on accurately locating key expression regions [24-25]. First, the model fuses the features extracted by MDMO with the change information of local regions to highlight motion patterns related to emotions. Then, through the Transformer's global modeling and multi-head attention mechanism, it effectively integrates the dependency relationships between various facial regions and potential emotional features. Finally, the classification layer completes the discrimination and output of diverse expression categories. This model can maintain sensitivity to subtle expression changes in complex environments, providing technical support with higher accuracy and stronger robustness for the automated assessment of psychological states and clinical ADT.

The proposed FER system based on MDMO and Transformer mainly consists of a presentation layer and a back-end model layer. The presentation layer includes a login module, a local face recognition module, and a camera recognition module, which realize user interaction and data collection functions. The back-end model layer takes the MDMO- and Transformer-based expression recognition model as the core; it performs feature selection, fusion, and weight optimization through the model fusion layer, and finally outputs accurate expression recognition results. The overall system design is progressive from user interaction to model operation, constructing a complete expression recognition process and realizing a full-process closed loop of data collection, processing, and high-precision recognition. The design of the FER system based on MDMO and the Transformer is depicted in Fig. 3:

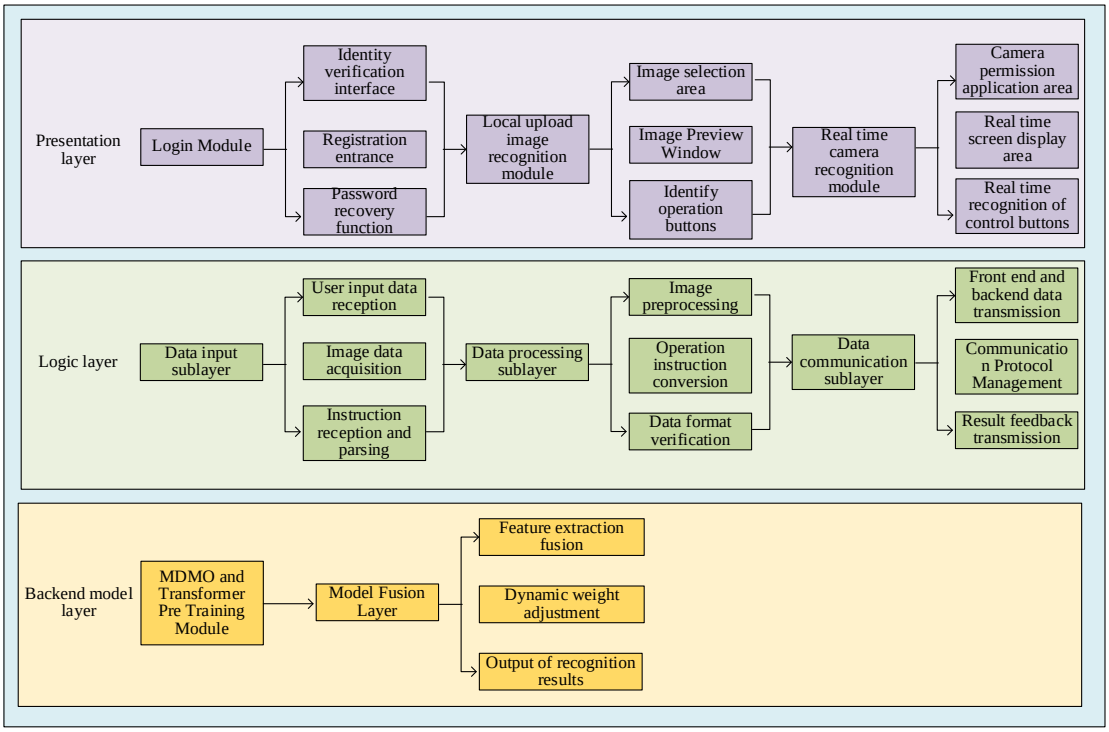

Figure 3: Design of the MDMO- and Transformer-based FER system

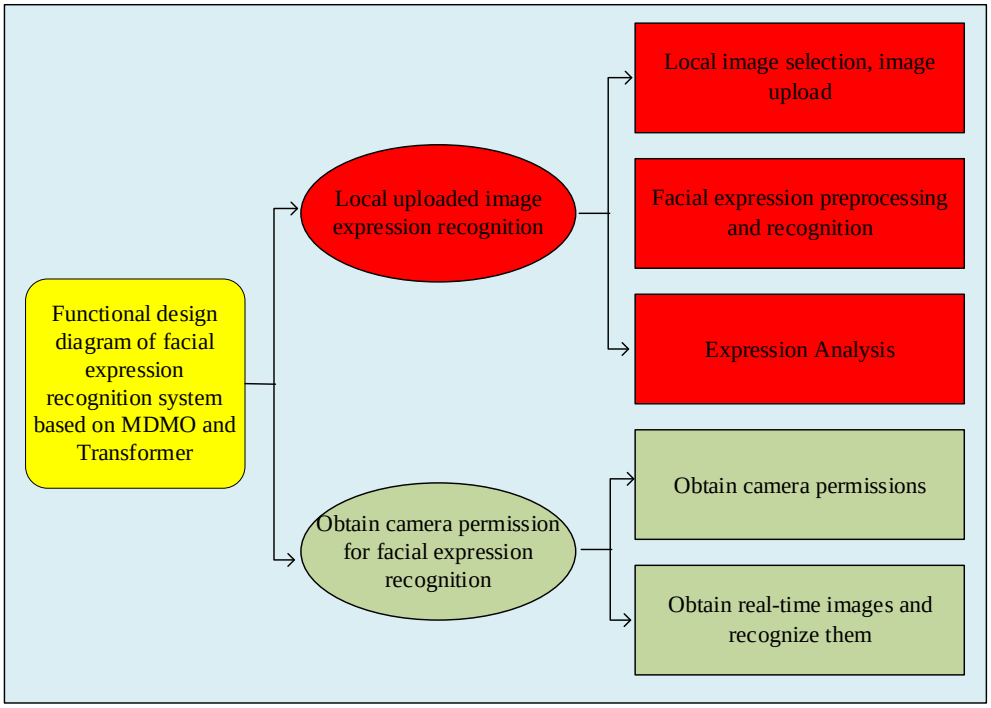

Figure 4: The FER system module architecture based on MDMO and Transformer

Table 1: Dataset splitting strategy

Subset	Sample size	Splitting strategy	Basis for splitting	k-fold cross-validation	Description
Training set	163	Stratified random split	Stratify by individual subjects and emotion categories to ensure balanced proportions across categories	No	Used for model parameter training to ensure all types of expressions are represented
Validation set	56	Independently sampled using the same strategy as the training set	Randomly sample according to subject and emotion category stratification	No	Used for model tuning and early stopping to ensure validation performance reflects generalization ability
Test set	56	Independently sampled, no overlap with training/validation set	Stratify by subject to ensure that subjects in the test set do not appear in training/validation.	No	Used for final model performance evaluation to test generalization and clinical applicability

The proposed FER system mainly includes two modules: expression recognition of locally uploaded pictures and expression recognition of obtaining camera permission. The system module architecture is illustrated in Fig. 4:

4 Application of the proposed expression localization and recognition model based on mdmo and transformer in MHADT

In MHADT, the data used in the application experiment of the proposed model using MDMO and Transformer mainly comes from the AVEC2019 DDS dataset. This dataset is collected and organized from audio-visual records of patients during clinical interviews; it can truly reflect the facial expressions and emotional changes of subjects under different mental health states. Specifically, the AVEC2019 DDS dataset contains 163 training samples, 56 validation samples, and 56 test samples, providing a standardized data foundation for model training, optimization, and performance evaluation. The dataset splitting strategy of this experiment is exhibited in Table 1.

When conducting the proposed expression localization and recognition model's application experiment, ESR, LBP, MDM, CNN, and the baseline model are selected as comparison objects. This selection aims to cover different representatives from traditional feature methods to DL methods, thus ensuring the evaluation's comprehensiveness. Among them, the Baseline model is a lightweight reference model, which consists of two convolution blocks and two fully connected layers. It takes face images as input and outputs key points or expression categories. The training adopts the Adam optimizer and basic data augmentation. Its performance is between traditional methods and CNN, providing a standard baseline for evaluating the MDMO + Transformer model.

The basic configuration of the experiment is outlined in Table 2:

Table 2: Experimental configuration

Item	Configuration instructions
Hardware environment	CPU: Intel Xeon Gold 6226R, RAM: 128 GB
Training epoch	50 epochs
Optimizer	Adam optimizer
Initial learning rate	0.001, adjusted based on performance during training
Batch size	32
Loss function for facial expression localization task	Mean squared error
Loss function for the FER task	Cross-entropy loss

The results of the ablation experiment in Table 3 indicate that the MDMO module mainly contributes to the accuracy of expression localization, and the Transformer module mainly improves the accuracy of expression recognition. Combining the two can provide stable and efficient expression analysis capabilities in the MHADT scenario.

Table 3: Expression localization and recognition ablation experiment based on MDMO and transformer

Model version	Localization accuracy (%)	Localization processing time (milliseconds (ms))	Recognition accuracy (%)	Recognition processing time (ms)
Only MDMO	81.5	52	80.2	55
Only Transformer	84	50	83	50
MDMO + Transformer	89.6	55	88.9	44

For different expression localization models' error comparison, Fig. 5 shows distinct differences in the performance of each model on the training, validation, and test subsets. Traditional methods such as ESR and LBP have relatively high errors on the training subset and show substantial fluctuations on the validation subset, which indicates their limited generalization ability. MDM and CNN perform better on the validation and test subsets, but there are still certain errors. The baseline model outperforms most traditional methods on the validation subset, but its overall stability is insufficient. By contrast, the proposed model integrating MDMO and Transformer achieves the lowest error rates across all three subsets, demonstrating a marked improvement over alternative methods, particularly on the validation subset. This demonstrates its comprehensive advantages in model generalization, stability, and accuracy.

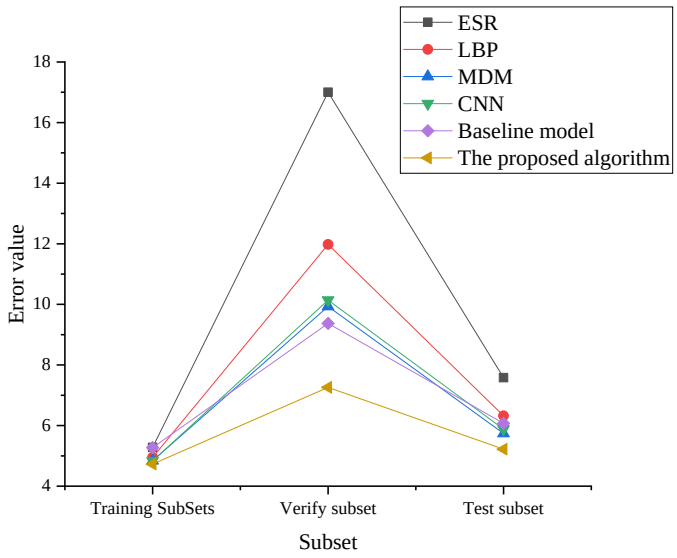

Figure 5: Error comparison of various expression localization models

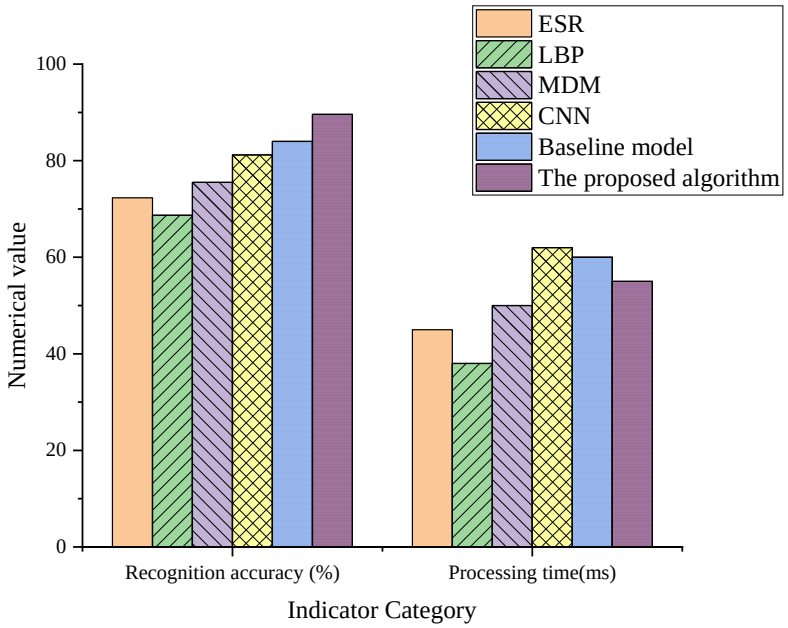

Figure 6: Comparison of localization accuracy and image processing speed of diverse expression localization models

Fig. 6 reveals that in comparing different expression localization models, there are certain differences in localization accuracy and image processing speed. Among traditional methods, LBP has the lowest accuracy, only 68.7%, while MDM has a relatively higher accuracy of 75.5%. CNN and the baseline model further improve the accuracy, reaching 81.2% and 84%, respectively. The model proposed here achieves the highest localization accuracy of 89.6%, notably better than other methods. Regarding processing speed, LBP and ESR take less time, while CNN and the baseline model take relatively longer. The proposed model's processing speed reaches 55 ms, slightly longer than some traditional methods; however, it greatly improves the localization accuracy while maintaining high real-time performance. This shows the optimal balance between precision and efficiency, making it more in line with the dual requirements of accuracy and usability in MHADT.

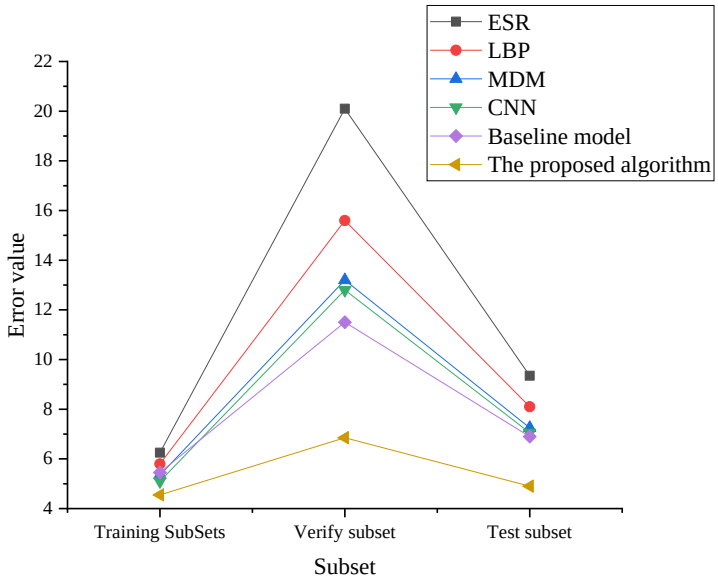

Figure 7: Error comparison of different FER models

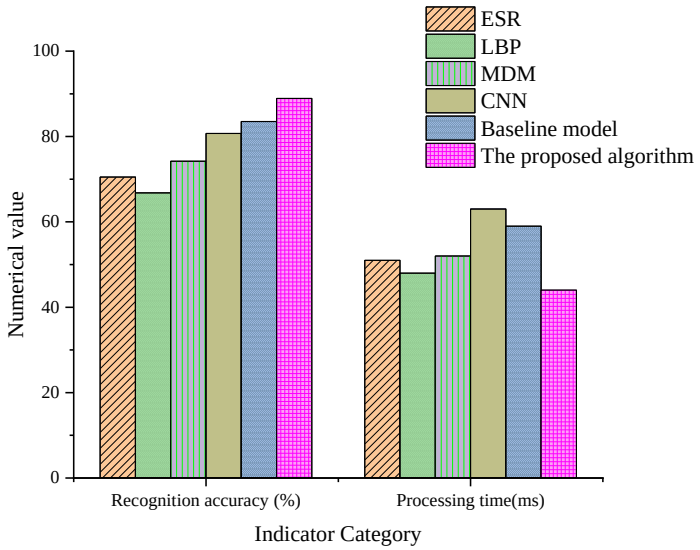

Figure 8: Comparison of recognition accuracy and image processing speed of each FER model

Fig. 7 demonstrates substantial differences in error performance across validation, test, and training subsets among diverse FER models. Traditional methods like ESR and LBP exhibit relatively high errors on validation and test subsets, illustrating limited generalization ability. Although MDM and CNN achieve lower errors on training and test subsets, a considerable error gap persists in the validation subset. The baseline model has lower errors on validation and test subsets, exhibiting greater stability than traditional methods. In contrast, the proposed model achieves the lowest errors across all three subset types. Specifically, its validation subset error is only 6.85, substantially lower than other methods. These results fully confirm the model's remarkable advantages in recognition accuracy and generalization capability, thereby furnishing more reliable technical support for MHADT.

Fig. 8 reveals noticeable differences in image processing speed and recognition accuracy when comparing various FER models. Among traditional methods, LBP achieves the lowest recognition accuracy at merely 66.8%, whereas MDM attains a relatively higher accuracy of 74.2%. CNN and the baseline model demonstrate improvements, reaching 80.7% and 83.5%, respectively. The MDMO- and Transformer-based model introduced in this study attains the highest recognition accuracy of 88.9%, prominently outperforming all comparative methods. Concerning processing speed, the

proposed model requires only 44ms, which is faster than CNN and the baseline model, while superior to traditional approaches such as ESR, LBP, and MDM. These results illustrate the model's dual advantages in real-time performance and recognition accuracy, aligning well with the integrated demands for high accuracy and efficiency in MHADT scenarios.

Table 4 shows that the proposed algorithm achieves the lowest model error on the training, validation, and test sets. The localization and recognition accuracy reach 89.6% and 88.9% respectively, remarkably exceeding ESR, LBP, MDM, CNN, and the baseline model. At the same time, it maintains high efficiency in image processing time, with the best comprehensive performance.

Table 4: Performance comparison of different models

Performance	Evaluation indicators	ESR	LBP	MDM	CNN	Baseline model	The proposed algorithm
Error of the facial expression location model	Training SubSets	5.28	4.95	4.84	4.83	5.27	4.73
	Verify subset	17	11.98	9.93	10.14	9.37	7.26
	Test subset	7.58	6.32	5.74	5.88	6.07	5.22
Localization accuracy and image processing time	Localization accuracy (%)	72.3	68.7	75.5	81.2	84	89.6
	Processing time(ms)	45	38	50	62	60	55
Error of the FER model	Training SubSets	6.25	5.8	5.35	5.1	5.45	4.55
	Verify subset	20.1	15.6	13.2	12.8	11.5	6.85
	Test subset	9.35	8.1	7.25	7.05	6.9	4.9
Recognition accuracy and image processing time	Recognition accuracy (%)	70.5	66.8	74.2	80.7	83.5	88.9
	Processing time(ms)	51	48	52	63	59	44

Table 5 indicates that in the experiment comparing CNN and Baseline models, the results of the statistical significance test show that the proposed model combining MDMO and Transformer achieves significant improvements in both localization and recognition accuracy. Specifically, the facial expression localization and recognition accuracies reach 89.6% and 88.9%. The errors on the validation and test subsets are reduced to 5.22 and 4.90, showing statistically significant differences compared with the CNN and Baseline models ($p < 0.05$ or $p < 0.01$). At the same time, the image processing time is only 44–55 ms, which is also significantly better than the comparison models, balancing real-time performance and high precision.

5 Discussion

Compared with existing studies, the proposed expression localization and recognition model based on MDMO and Transformer shows remarkable performance advantages in the MHADT scenario. Compared with the multi-modal emotion recognition model proposed by Zhang et al., with a recognition accuracy of about 82.1%, the proposed model achieves a recognition accuracy of 88.9% on the AVEC2019 DDS dataset. Meanwhile, the inference time is only 44 ms, achieving higher emotion recognition accuracy while maintaining high real-time performance. The main reasons for the performance improvement are as follows. First, the MDMO feature extraction module can effectively capture the dynamic changes and directional movement information of facial expressions, enhancing the expression of temporal features. Second, the Transformer's global attention mechanism effectively captures long-range dependencies within expression sequences. This capability enables cross-frame information integration and contextual feature enhancement, substantially improving recognition robustness and generalization performance. Zhang et al.'s method enhances overall emotion recognition performance by incorporating multimodal features, including voice, image, and text data. However, the method remains constrained by its reliance on local convolutional features within the visual modality alone. This architectural limitation restricts its capacity to model temporal dynamics, leading to reduced accuracy in scenarios involving complex expression variations.

Table 5: A comparative statistical analysis of the proposed algorithm with CNN and baseline model

Performance indicators	Comparison model	Mean value of the proposed algorithm	Mean value of comparison models	Average difference (Δ)	t-value	p-value	Significance level
Localization accuracy (%)	CNN	89.6	81.2	8.4	3.27	0.014	*(p<0.05)
	Baseline model	89.6	84	5.6	2.89	0.022	*(p<0.05)
Localization accuracy (Test)	CNN	5.22	5.88	-0.66	-2.57	0.028	*(p<0.05)
	Baseline model	5.22	6.07	-0.85	-3.04	0.019	*(p<0.05)
Recognition accuracy (%)	CNN	88.9	80.7	8.2	3.64	0.011	*(p<0.05)
	Baseline model	88.9	83.5	5.4	3.11	0.018	*(p<0.05)
Recognition error (Test)	CNN	4.9	7.05	-2.15	-3.78	0.009	**(p<0.01)
	Baseline model	4.9	6.9	-2	-3.51	0.012	*(p<0.05)
Image processing time (ms)	CNN	44	63	-19	-4.12	0.007	**(p<0.01)
	Baseline model	44	59	-15	-3.56	0.011	*(p<0.05)

Although the proposed model performs well in multiple experimental indicators, it still has certain limitations. First, the AVEC2019 DDS dataset has a small scale and uneven distribution of expression categories. Insufficient samples may lead the model to overfit on certain specific expression categories (such as negative or complex emotions), thereby affecting generalization ability. It may also perform inconsistently among different patient groups. For example, noisy or partially occluded faces, age, gender, or cultural differences may cause some expressions to be easily misclassified. This suggests that the model's results need to be interpreted carefully in clinical applications. Second, the model only uses facial images for recognition; it does not fully integrate multi-source information such as voice or physiological signals, which may limit its performance in multi-scenario psychological assessment. Third, the application of MHADT must focus on patients' informed consent, data privacy protection, and potential demographic biases; therefore, it is necessary to ensure the model's fairness and reliability among different populations.

6 Conclusion

The proposed MDMO- and Transformer-based expression localization and recognition model has undergone systematic application experiments in the MHADT scenario; this model has also been compared with ESR, CNN, MDM, LBP, and the baseline model from multiple dimensions. Concerning expression localization, the model attains the lowest errors on the training, validation, and test subsets, with a localization accuracy of 89.6%, markedly outperforming ESR, LBP, MDM, CNN, and the baseline model. Although its processing speed is 55ms, which is slightly more than that of some traditional methods, it maintains high real-time performance while substantially improving precision. Regarding FER, the model also has the lowest errors on all subsets, with an error of only 6.85 on the validation subset. Its recognition accuracy reaches 88.9%, and the processing time is only 44ms. This is shorter than the baseline and CNN models while surpassing traditional methods like ESR, LBP, and MDM, demonstrating its dual benefits in recognition accuracy and real-time processing capability. The proposed model can efficiently and accurately capture subtle changes in patients' facial expressions. Besides, it can realize reliable assessment of emotional states and mental health indicators, and provide a scientific basis for clinical psychological intervention, treatment course monitoring, and psychological therapy effect tracking. At the same time, the model performs excellently in both processing speed and recognition accuracy, balancing real-time performance and stability, and can support expression analysis for continuous videos or interviews. In addition to the mental health field, the proposed model is also applicable to scenarios such as intelligent HCI, educational assessment, and behavior monitoring. It can provide high-precision technical support for emotion computing, interactive feedback, and behavior analysis. The model demonstrates comprehensive advantages in accuracy, efficiency, and generalization ability, laying a

solid foundation for the application of FER technology in multiple fields.

Future work can draw on ideas such as adaptive inversion control, nonlinear optimal control, and adaptive backstepping control to introduce a more robust adaptive mechanism. Thus, the model's stability and generalization ability can be improved under noisy, occluded, and incomplete data conditions. At the same time, through multi-modal fusion and transfer learning strategies, the model's applicability and reliability in different patient groups and complex clinical scenarios can be further enhanced.

Acknowledgment: This study is supported by Foundation of Guangdong Educational Committee (No. 2025WTSCX200); Research Project on Mental Health Education in Higher Vocational Education (No. XL2025122); Humanity and Social Science foundation of Ministry of Education (No.25JDSZ3164); Foundation of Guangdong Polytechnic of Industry and Commerce (No. 2025-PT-01).

References

[1] Zhou Y, Chen L, Huang T, et al. Dual complementarity transformer for micro-expression recognition. Multimedia Systems, 2025, 31(5): 371. https://doi.org/10.1007/s00530-025-01960-w

[2] Ma Y, Shen J, Zhao Z, et al. What can facial movements reveal? Depression recognition and analysis based on optical flow using Bayesian networks. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 3459-3468. https://doi.org/10.1109/tnsre.2023.3305351

[3] Muetunda F, Sabry S, Jamil M L, et al. AI-Assisted diagnosing, monitoring and treatment of mental disorders: A survey. ACM Transactions on Computing for Healthcare, 2025, 5(4): 1-24. https://doi.org/10.1145/3681794

[4] Bao Y, Wu C, Zhang P, et al. Boosting micro-expression recognition via self-expression reconstruction and memory contrastive learning. IEEE Transactions on Affective Computing, 2024, 15(4): 2083-2096. https://doi.org/10.1109/TAFFC.2024.3397701

[5] Wimbarti S, Kairupan B H R, Tallei T E. Critical review of self-diagnosis of mental health conditions using artificial intelligence. International Journal of Mental Health Nursing, 2024, 33(2): 344-358. https://doi.org/10.1111/inm.13303

[6] Li Y, Liu M, Lao L, et al. Counterfactual discriminative micro-expression recognition. Visual Intelligence, 2024, 2(1): 29. https://doi.org/10.1007/s44267-024-00063-w

[7] Liu Y, Zhang X, Li Y, et al. Graph-based facial affect analysis: A review. IEEE Transactions on Affective Computing, 2022, 14(4): 2657-2677. https://doi.org/10.1109/TAFFC.2022.3215918

[8] Nidhi, Verma B. From methods to datasets: a detailed study on facial emotion recognition. Applied Intelligence, 2023, 53(24): 30219-30249. https://doi.org/10.1007/s10489-023-05052-y

[9] Pereira R, Mendes C, Ribeiro J, et al. Systematic review of emotion detection with computer vision and deep learning. Sensors, 2024, 24(11): 3484. https://doi.org/10.3390/s24113484

[10] Jabbooree A I, Khanli L M, Salehpour P, et al. A novel facial expression recognition algorithm using geometry β–skeleton in fusion based on deep CNN. Image and Vision Computing, 2023, 134: 104677. https://doi.org/10.1016/j.imavis.2023.104677

[11] Kim S, Nam J, Ko B C. Facial expression recognition based on squeeze vision transformer. Sensors, 2022, 22(10): 3729. https://doi.org/10.3390/s22103729

[12] Zhang F, Chai L. A review of research on micro-expression recognition algorithms based on deep learning. Neural Computing and Applications, 2024, 36(29): 17787-17828. https://doi.org/10.1007/s00521-024-10262-7

[13] Zhang P, Wang R, Luo J, et al. Micro-expression recognition algorithm using regions of interest and the weighted ArcFace loss. Electronics, 2024, 14(1): 2. https://doi.org/10.3390/electronics14010002

[14] Zhao S, Tang H, Mao X, et al. Dfme: A new benchmark for dynamic facial micro-expression recognition. IEEE Transactions on Affective Computing, 2023, 15(3): 1371-1386. https://doi.org/10.1109/TAFFC.2023.3341918

[15] Zhang L, Hong X, Arandjelović O, et al. Short and long range relation based spatio-temporal transformer for micro-expression recognition. IEEE Transactions on Affective Computing, 2022, 13(4): 1973-1985. https://doi.org/10.1109/TAFFC.2022.3213509

[16] Pan H, Xie L, Wang Z. C3DBed: Facial micro-expression recognition with three-dimensional convolutional neural network embedding in transformer model. Engineering Applications of Artificial Intelligence, 2023, 123: 106258. https://doi.org/10.1016/j.engappai.2023.106258

[17] Tian Y, Zhu J, Yao H, et al. Facial expression recognition based on vision transformer with hybrid local attention. Applied Sciences, 2024, 14(15): 6471. https://doi.org/10.3390/app14156471

[18] Zhang X, Li M, Lin S, et al. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(5): 3192-3203. 10.1109/TCSVT.2023.3312858

[19] Indolia S, Nigam S, Singh R, et al. Micro expression recognition using convolution patch in vision transformer. IEEE Access, 2023, 11: 100495-100507. https://doi.org/10.1109/ACCESS.2023.3314797

[20] Qin L, Wang M, Deng C, et al. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. arXiv:2308.11509. https://doi.org/10.48550/arXiv.2308.11509

[21] Dai M, Hu J, Zhuang J, et al. A transformer-based feature segmentation and region alignment method for UAV-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(7): 4376-4389. https://doi.org/10.1109/TCSVT.2021.3135013

[22] Ma F, Sun B, Li S. Transformer-augmented network with online label correction for facial expression recognition. IEEE Transactions on Affective Computing, 2023, 15(2): 593-605. https://doi.org/10.1109/TAFFC.2023.3285231

[23] Aloysius N, Geetha M, Nedungadi P. Incorporating relative position information in transformer-based sign language recognition and translation. IEEE Access, 2021, 9: 145929-145942. https://doi.org/10.1109/ACCESS.2021.3122921

[24] Islam M S, Sang Y, Mohammed A A Q, et al. Facial micro-expression recognition from videos through domain adaptation and multi-modal spatio-temporal feature ensemble: Facial micro-expression. Multimedia Systems, 2025, 31(5): 330. https://doi.org/10.1007/s00530-025-01898-z

[25] Munanday A P, Sazali N, Asogan A, et al. The implementation of transfer learning by convolution neural network (CNN) for recognizing facial emotions. Journal of Advanced Research in Applied Sciences and Engineering Technology, 2023, 32(2): 255-276. https://doi.org/10.37934/araset.32.2.255276