scientific reports

Check for updates

OPEN

Leveraging vision transformers and entropy-based attention for accurate micro-expression recognition

Yibo Zhang[1,3], Weiguo Lin[1]✉, Yuanfa Zhang[1], Junfeng Xu[1]✉ & Yan Xu[2]✉

Micro-expressions are difficult to fake and inherently truthful, making micro-expression recognition technology widely applicable across various domains. With the development of artificial intelligence, the accuracy and efficiency of micro-expression recognition systems have been significantly improved. However, the short duration and subtle facial movement changes present significant challenges to real-time recognition and accuracy. To address these issues, this paper proposes a novel micro-expression recognition method based on the Vision Transformer. First, a new model called HTNet with LAPE (hierarchical transformer network with learnable absolute position embedding) is introduced to improve the model's capacity for capturing subtle facial features, thereby enhancing the accuracy of micro-expression recognition. Second, an entropy-based selection agent attention is proposed to reduce the model parameters and computational effort while preserving its learning capability. Finally, a diffusion model is utilized for data augmentation to expand the micro-expression sample size, further enhancing the model's generalization, accuracy, and robustness. Extensive experiments conducted on multiple datasets validate the framework's effectiveness and highlight its potential in real-world applications.

Keywords Vision transformer, Micro-expression recognition, Agent attention

Micro-expressions are involuntary facial expressions of extremely short duration (0.04 to 0.2 seconds), acting as subtle yet powerful indicators of real emotions and intentions that individuals often attempt to conceal[1]. Unlike macro-expressions, which are usually consciously controlled and may not accurately reflect a person's true emotions, micro-expressions provide a more reliable and distinct indication of emotion[2]. Accurately detecting these fleeting expressions has significant implications for various fields, including law enforcement, security, psychological research, and professional negotiations. Micro-expressions can help physicians observe subtle changes in a patient's mood, which is particularly crucial for the early diagnosis of mental health conditions[3]. For instance, patients with mood disorders such as anxiety and depression may unconsciously display several typical micro-expressions, which can serve as diagnostic aids[4]. Certain neurological disorders, such as Parkinson's and Alzheimer's diseases, can affect patients' facial muscle movement, thereby altering their facial expressions. By analyzing changes in micro-expressions, doctors can detect the condition at an earlier stage or monitor its progression. Micro-expressions also have value in pain assessment[5], particularly in patients unable to speak or express themselves, such as infants, Alzheimer's patients, or those in a coma. Changes in facial micro-expressions can provide valuable information regarding pain levels, aiding physicians in administering more precise treatments. During hemodialysis[6], the use of micro-expression recognition technology can rapidly detect abnormalities in the patient's condition, enable early intervention in unforeseen circumstances, and enhance operational efficiency. In doctor-patient interactions, micro-expression analysis can help doctors understand the patient's true feelings or confusion, thereby improving the treatment process, enhancing patient trust, and facilitating communication. During surgery, the micro-expressions of surgical team members may reflect emotions such as nervousness, anxiety, or confidence. Timely detection of these changes can help the team regulate emotions and ensure smooth operations.

[1]School of Computer and Cyberspace Security, Communication University of China, Beijing 100024, China. [2]Department of Nephrology, Emergency General Hospital, Beijing 100028, China. [3]School of Computing, North China Institute of Science and Technology, Langfang 065201, China. ✉email: Linwei@cuc.edu.cn; junfeng@cuc.edu.cn; mirocle77@163.com

In the early stages of micro-expression research, research primarily focused on traditional computer vision techniques for feature extraction and classification. The groundwork for micro-expression analysis was laid by the pioneering work of Ekman and Friesen, who developed the Facial Action Coding System (FACS), which established a standardized system for classifying facial actions[7]. Researchers relied on manual feature extraction methods, such as Local Binary Pattern (LBP), to capture texture information in facial images. LBP is effective in highlighting local texture variations, therefore enabling effective discrimination of facial expressions[8]. However, it lacks the ability to incorporate temporal information, which is essential for accurate micro-expression analysis. To solve this problem, Zhao et al.[9] introduced LBP-TOP, an extension that operates on three orthogonal planes and combines temporal information, which increases computational load but offers a more comprehensive feature set. Wang et al.[10] further improved the method by proposing LBP-SIP, which effectively reduces redundancy and enhances computational efficiency. Optical flow[11] is a technique for analyzing object motion between image frames, and its application in micro-expression recognition has been extensively explored. Liu et al.[12] used the main direction averaged optical flow (MDMO) to capture regional facial motions. Liong et al.[13] designed the bi-weighted oriented optical flow (Bi-WOOF) to weigh local and global motion cues, both of which have proven effective.

The emergence of deep learning[14–17] signified a paradigm shift in micro-expression recognition research. Convolutional Neural Networks (CNNs) were one of the first deep learning models applied to this task. Patel et al.[18] addressed the challenge posed by the limited sample size of micro-expression datasets by employing a pre-trained VGGNet for feature extraction through transfer learning. Subsequent work has focused on modifying the network architecture to accommodate the specific challenges of micro-expression data. For instance, Peng et al.[19] mitigated overfitting by reducing the number of ResNet layers. To better capture spatial and temporal information, researchers have introduced hybrid models that integrate CNNs with recurrent neural networks (RNNs) or long short-term memory (LSTM) networks. These models employ CNNs to extract spatial features and RNNs to model temporal dependencies, thereby significantly enhancing recognition accuracy. Three-dimensional CNNs (3D-CNN) have also been used to jointly process spatial and temporal data, where Reddy et al. concentrated on regional 3D-CNNs to enhance computational efficiency. Cakir et .al[20] utilized action units (AUs) to localize the most active facial landmarks and determine the most representative regional scale for each landmark in a detection task. Their study on variable-scale landmark patches for facial action unit (AU) detection, employing a vision transformer (ViT) with a perceptual attention mechanism, achieved significant results.

Recent advances in the field of micro-expression recognition have brought about a paradigm shift with the introduction of visual transformer-based models capable of capturing long-distance dependencies and processing data in parallel. The Visual Transformer (ViT)[21,22] model of Dosovitskiy et al. has had a profound impact by applying the transformer architecture to image classification tasks. The model replaced traditional convolutional operations with a self-attention mechanism and exhibited exceptional scalability and performance on large-scale datasets. Since then, researchers have applied visual transformers to micro-expression recognition, and Liu et al.[23] subsequently proposed a lightweight ViT model, which enhances micro-expression analysis via transfer learning. HTNet, a hierarchical transformer network that combines optical flow features of the facial region and addresses the limitations of previous models by considering the facial structure and local-to-global feature relationships, was introduced by Wang et al[24].

Visual transformer models still face challenges such as high computational requirements and the need for large datasets. However, datasets in micro-expression recognition are usually limited and often struggle to meet these requirements[25]. This motivates continuous research to improve the efficiency and generalization ability of these models and to explore techniques such as data augmentation and adversarial training to enhance the effectiveness of limited datasets. In conclusion, the field of micro-expression recognition has evolved from manual feature extraction to deep learning-based approaches, and the latest visual transformer models show great potential[26]. However, there are still challenges in terms of computational efficiency, dataset limitations, and real-time analysis requirements, which remain the core challenges of current research[27]. To address these challenges, our research delves into the visual transformer-based micro-expression recognition technique and proposes a new approach to improve recognition accuracy and efficiency by taking advantage of the visual transformer.

Our approach introduces a hierarchical transformer network, HTNet, which integrates the optical flow features of specific facial regions to effectively capture features and processes the inherent spatial relationships in facial markers through the incorporation of a multilayer transformer module.[24] To enhance the model's ability to capture subtle features, we propose and implement a Learned Absolute Position Encoding (LAPE) module, which significantly improves the model's ability to recognize subtle details, thereby optimizing the recognition accuracy. In addition, to mitigate the computational overhead associated with LAPE and to simplify the model, we propose an entropy-based selective removal technique for the attention layer and introduce a novel agent attention mechanism[28]. These innovations not only decrease the model parameters and computational requirements but also preserve the model's ability to learn rich features, thereby achieving an effective balance between computational efficiency and representational capability. Finally, to address the limitation posed by the limited sample size of the micro-expression dataset, which constrains the model's generalization ability, we integrate a data enhancement technique based on the diffusion model[29]. This approach enhances the detection accuracy and robustness of the micro-expression recognition model, making it more suitable for practical application scenarios.

In this study, we provide a comprehensive overview of our contributions to the field of micro-expression recognition, focusing specifically on innovations and improvements to existing visual converter-based models. We perform a systematic evaluation of the proposed approach, demonstrate its effectiveness across

multiple datasets, and investigate the potential impact of micro-expression recognition techniques in practical applications.

Results

Experimental methodology

Our experiments aim to assess the effectiveness of the key components of our framework: the LAPE module, the ESAAT module, and the diffusion model-based data augmentation technique. We also compare our model's performance with state-of-the-art methods and evaluate its generalization ability across a diverse dataset.

Experimental implement details

In this paper, we employ cross-entropy as the loss function, with Adam as the optimizer, a learning rate of 5×10^{-5}, and 800 training epochs. The experiments were conducted on a system running Ubuntu 20.04 LTS (Focal Fossa), equipped with an Intel Xeon(R) Gold 6430 processor, an NVIDIA GeForce RTX 4090 GPU (24GB), and 120GB of RAM. The software environment includes Python 3.8 and CUDA 11.3.

Datasets

We employ four widely-used micro-expression datasets: SMIC[30], SAMM[31], CASME II[32], and CAS(ME)3[33]. These datasets provide a comprehensive range of spontaneous micro-expressions from various subjects, covering a range of emotional responses. The SMIC dataset contains 164 micro-expression sequences with three categories: positive, negative, and surprise. The SAMM dataset consists of 133 sequences with similar emotional categories. The CASME II dataset includes 145 sequences with a focus on spontaneous micro-expressions. The CAS(ME)3 dataset is the largest, containing 673 sequences and providing a more diverse and ecologically valid set of expressions.

Experimental metric

Owing to the imbalanced distribution of micro-expressions across the three categories in the micro-expression dataset, we employ the unweighted F1 score (UF1) and the unweighted average recall (UAR) as evaluation metrics for the model to objectively assess its performance.

UF1 evaluates the overall performance of the model across all categories by averaging the F1-scores of individual categories. Similar to the conventional macro-averaged F1-score (Macro F1-score), UF1 calculates the F1-score for each category and performs an unweighted average to prevent underrepresented categories from being overlooked. Specifically, for each category *c*, the F1-score is computed as follows:

$$F1_c = \frac{2 \times P_c \times R_c}{P_c + R_c} \tag{1}$$

where, P_c (Precision) is defined as:

$$P_c = \frac{TP_c}{TP_c + FP_c} \tag{2}$$

and the R_c (Recall) is defined as:

$$R_c = \frac{TP_c}{TP_c + FN_c} \tag{3}$$

where TP_c represents the number of true positive instances (True Positives) for category *c*, FP_c denotes the number of false positive instances (False Positives) for category *c*, and FN_c refers to the number of false negative instances (False Negatives) for category *c*. The F1-scores of all categories are then averaged as follows:

$$UF1 = \frac{1}{C} \sum_{c=1}^{C} F1_c \tag{4}$$

The UF1 is well-suited for handling category imbalance, as it prevents the metric from being disproportionately influenced by categories with larger data volumes. This metric evaluates the overall performance of the model across all categories while ensuring that the classification performance of underrepresented categories is not overlooked. UF1 ranges from 0 to 1, with values closer to 1 indicating better overall model performance.

Unweighted Average Recall (UAR) calculates the recall for each category and then averages these values. This metric assesses the model's ability to recognize all categories while preventing the overall score from being skewed by categories with larger data volumes. UAR is defined as follows:

$$UAR = \frac{1}{C} \sum_{c=1}^{C} R_c \tag{5}$$

where R_c represents the recall (recall rate), and C denotes the total number of categories. UAR quantifies the model's ability to recognize all categories and is particularly suitable for datasets with imbalanced category distributions. This metric focuses solely on recall, reflecting the model's effectiveness in recognizing samples

Model	Metric	SMIC	SAMM	CASME II
LBP-TOP	UF1	0.2000	0.3954	0.7026
	UAR	0.5280	0.4102	0.7429
Bi-WOOF	UF1	0.5727	0.5211	0.7805
	UAR	0.5829	0.5139	0.8026
OFF-ApexNet	UF1	0.6817	0.5409	0.8764
	UAR	0.6695	0.5392	0.8681
STSTNet	UF1	0.6801	0.6588	0.8382
	UAR	0.7013	0.6810	0.8686
MobileViT	UF1	0.7141	0.7428	0.7251
	UAR	0.7356	0.6781	0.6997
MMNet	UF1	–	0.8391	0.9494
	UAR	–	–	–
Micro-BERT	UF1	0.8550	0.8386	0.9034
	UAR	0.8384	0.8475	0.8914
HSTA	UF1	0.8470	0.8470	0.9250
	UAR	0.7800	0.8390	0.9220
Ours	UF1	0.8203	0.8392	0.9676
	UAR	0.8137	0.8306	0.9613

Table 1. Quantitative experiments compare the proposed method with representative approaches across three datasets. The evaluation metrics used are UF1 and UAR.

	CAS(ME) 3	
Model	UF1	UAR
STSTNet	0.3795	0.3792
Micron-BERT	0.5604	0.6125
HSTA	0.5930	0.6180
Ours	0.5976	0.5981

Table 2. Quantitative experiments evaluate the proposed method against representative approaches in a cross-dataset setting. The evaluation metrics used are UF1 and UAR.

	CAS(ME) 3	
Model	UF1	UAR
w/o data augmentation	0.5976	0.5981
w/ data augmentation	0.6172	0.6143

Table 3. Ablation study on data augmentation modules.

from underrepresented categories. UAR ranges from 0 to 1, with higher values indicating greater average recall across all categories.

Comparative experiments

We compare the performance of our model with several state-of-the-art micro-expression recognition models, such as LBP-TOP[10], Bi-WOOF[13], OFF-ApexNet[34], STSTNet[35], MobileViT[36], MMNet[37], Micron-BERT[38] and HSTA[39]. We conduct the experiments with K-fold cross-validation, with the final results presented in Table 1.

Generalization experiments

To evaluate the capacity for generalization of our model, we performed experiments on the CAS(ME) 3 dataset, which is known for its diversity and ecological validity. We employed two evaluation strategies to assess model performance, and the experimental results are presented in Tables 2 and 3.

Cross-dataset validation: We performed K-fold cross-validation on the CAS(ME) 3 dataset to evaluate the model's capacity to generalize to unseen data.

Impact of Data Augmentation: We compared the model's performance with and without diffusion model-based data augmentation to quantify its effectiveness in improving generalization.

Model	Metric	SMIC	SAMM	CASME II	Composite
w/o LAPE	UF1	0.8049	0.8131	0.9532	0.8603
	UAR	0.7905	0.8124	0.9516	0.8475
w/ LAPE	UF1	0.8165	0.8251	0.9643	0.8713
	UAR	0.8020	0.8174	0.9537	0.8507

Table 4. Ablation study on LAPE modules.

Model	Metric	SMIC	SAMM	CASME II	Composite
w/o ESAAT	UF1	0.8165	0.8251	0.9643	0.8713
	UAR	0.8020	0.8174	0.9537	0.8507
w/ ESAAT	UF1	0.8203	0.8392	0.9676	0.8852
	UAR	0.8137	0.8306	0.9613	0.8665

Table 5. Ablation study on ESAAT modules.

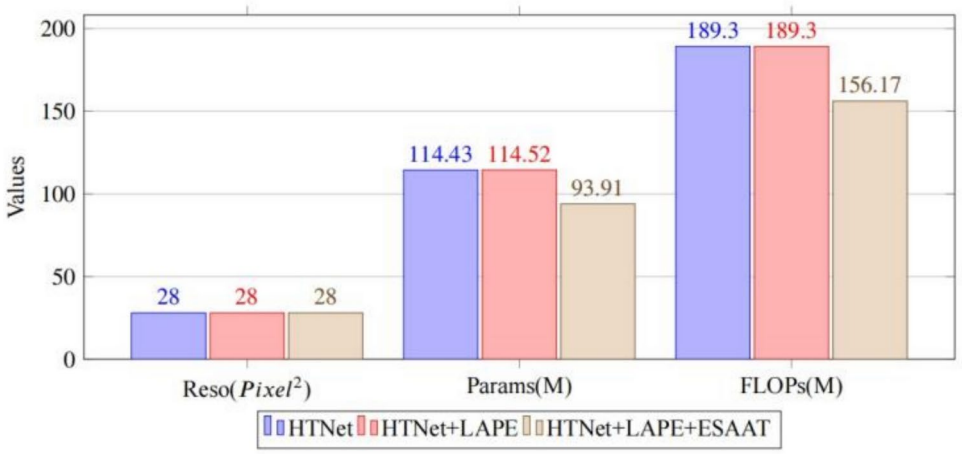

Fig. 1. Comparative analysis of different module combinations in terms of resolution, parameters, and FLOPs.

Ablation studies

An ablation study was performed to evaluate the respective contributions of the LAPE and ESAAT modules to the overall performance of our model. The experimental results are presented in Tables 4 and 5.

LAPE module ablation: We compared the performance of the model with and without the LAPE module to assess its role in capturing spatial relationships in microexpressions.

ESAAT module ablation: We analyzed the influence of the ESAAT module in reducing computational complexity while maintaining accuracy. Additionally, we examined the reduction in model parameters and its effect on recognition accuracy.

From the results presented in tables and Fig. 1, it is evident that, compared to the micro-expression recognition model without the LAPE or ESAAT modules, the number of model parameters is reduced by approximately 18% with its inclusion, while the model demonstrates slight performance improvements across different datasets. These results indicate that the LAPE and ESAAT modules play a crucial role in balancing computational efficiency and expressive power by reducing computational overhead while enhancing the model's representational capacity.

Results and discussion

The results of our experiments reveal the following key findings:

Comparative Experiments: Our model outperforms or matches the state-of-the-art methods in both accuracy and efficiency. The integration of the ESAAT module and data augmentation technique provides a competitive advantage, particularly in handling diverse and complex expressions.

Generalization Experiments: The generalization experiments on the CAS(ME) 3 dataset demonstrate that our model generalizes well to new data, with the data augmentation technique significantly enhancing performance.

Ablation Studies: The LAPE module significantly enhances the model's ability to capture spatial relationships, resulting in higher recognition accuracy. The ESAAT module efficiently reduces the model's computational complexity while maintaining accuracy.

The experimental results demonstrate the effectiveness of our proposed framework for micro-expression recognition. The LAPE and ESAAT modules, when integrated with diffusion model-based data augmentation, not only boost the model's accuracy and efficiency but also substantially enhance its generalization capabilities. These findings underscore the potential of our framework for real-world applications that require accurate and robust micro-expression recognition.

Discussion

The paper concludes that the proposed micro-expression recognition framework, which combines HTNet with LAPE and ESAAT modules as well as diffusion model-based data augmentation, significantly improves the accuracy and efficiency of micro-expression recognition. The framework's performance on multiple datasets demonstrates its potential for practical applications. Future work will focus on enhancing the model's real-time inference capabilities and extending its multi-modal fusion capabilities. In the future, we will focus on leveraging Vision Transformers (ViTs) for multimodal fusion with adaptable patch sizes.

Methods

Method details

The methodology proposed for micro-expression recognition is a comprehensive framework that integrates advanced deep learning techniques, innovative attention mechanisms, and data augmentation strategies. Figure 2 presents the overall architectural diagram of the proposed method. This section provides a detailed explanation of the three core components of our approach: the Learnable Absolute Position Embedding (LAPE) module, the Entropy-based Selection Agent Attention (ESAAT) module, and the diffusion model-based data augmentation technique.

Learnable absolute position embedding module

The LAPE module is designed to enhance the model's ability to capture the spatial dependencies within facial expressions. Traditional Vision Transformer models rely on fixed position embeddings, which may not be fully effective in capturing the nuances of micro-expressions. Our LAPE module introduces learnable position embeddings that adapt to the specific spatial features of facial movements.

The LAPE module functions as follows: 1) For each image patch, a unique position embedding is learned during the training process. 2) These embeddings are added to the patch embeddings, supplying the model with information regarding the relative positions of different facial regions. 3) The position embeddings are optimized alongside the rest of the model, enabling the network to better capture the spatial hierarchy of facial expressions.

Mathematically, the LAPE can be formulated as:

$$LAPE(x_i) = x_i + PE(p_i) \tag{6}$$

where x_i is the embedding of the i-th patch, p_i is its position, and $PE(p_i)$ is the learnable position embedding vector for that position.

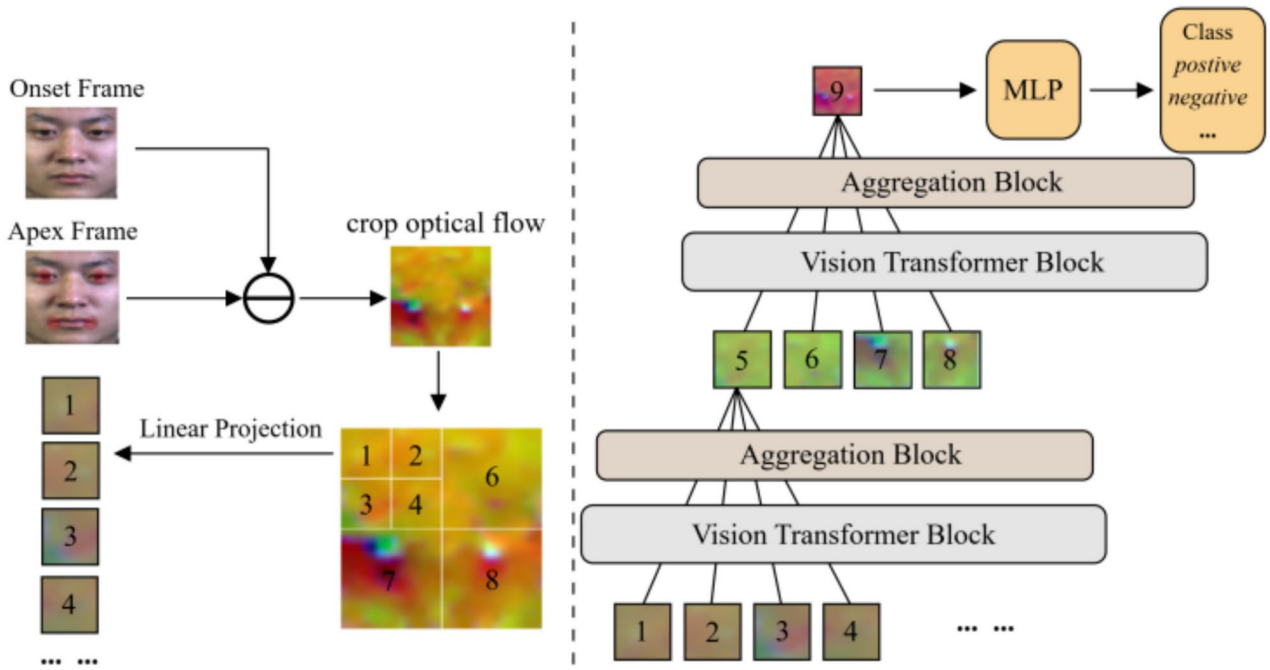

Fig. 2. The overview architecture diagram of the proposed model.

Entropy-based selection agent attention module

The ESAAT module addresses the computational inefficiency of traditional attention mechanisms by selectively removing less relevant attention layers based on entropy measures. This approach reduces the model's computational complexity without sacrificing performance.

The ESAAT module operates through the following steps: 1) Compute the transfer entropy between each attention layer and the output layer to determine the importance of each layer. 2) Remove attention layers with low transfer entropy, as they contribute less to the final output, based on the transfer entropy values calculated. 3) Integrate a new attention mechanism, Agent Attention, which combines the advantages of softmax and linear attention to balance computational efficiency and representational power.

The Agent Attention (as shown in Fig. 3) mechanism can be mathematically represented as:

$$Att(Q, K, A, V) = Softmax(QA^T)Softmax(AK^T)V \tag{7}$$

where Q (Query) represents the query matrix, which encodes the query vector of the input data. K (Key) denotes the key matrix, representing the key vector of the input data. V (Value) refers to the value matrix, which encapsulates the value vector of the input data. A (Agent Matrix) serves as an intermediary, regulating the interaction between the query and the keys. The function $softmax(\cdot)$ denotes standard softmax normalization. Compared to the conventional self-attention mechanism, the Agent Attention mechanism introduces the Agent Matrix A, which decomposes the attention computation into two stages. In the first stage, the correlation between Q and A is computed. In the second stage, the correlation between A and K is computed. The two-stage attentional weighting process allows the query information to be modulated by the agent matrix before interacting with the key-value pairs. Additionally, this approach enhances flexibility. While traditional attention mechanisms compute the relationship between the query and key directly. The agent Attention introduces an intermediary mapping through the Agent Matrix A, enabling the model to capture more intricate attention patterns and operate in higher-order feature spaces.

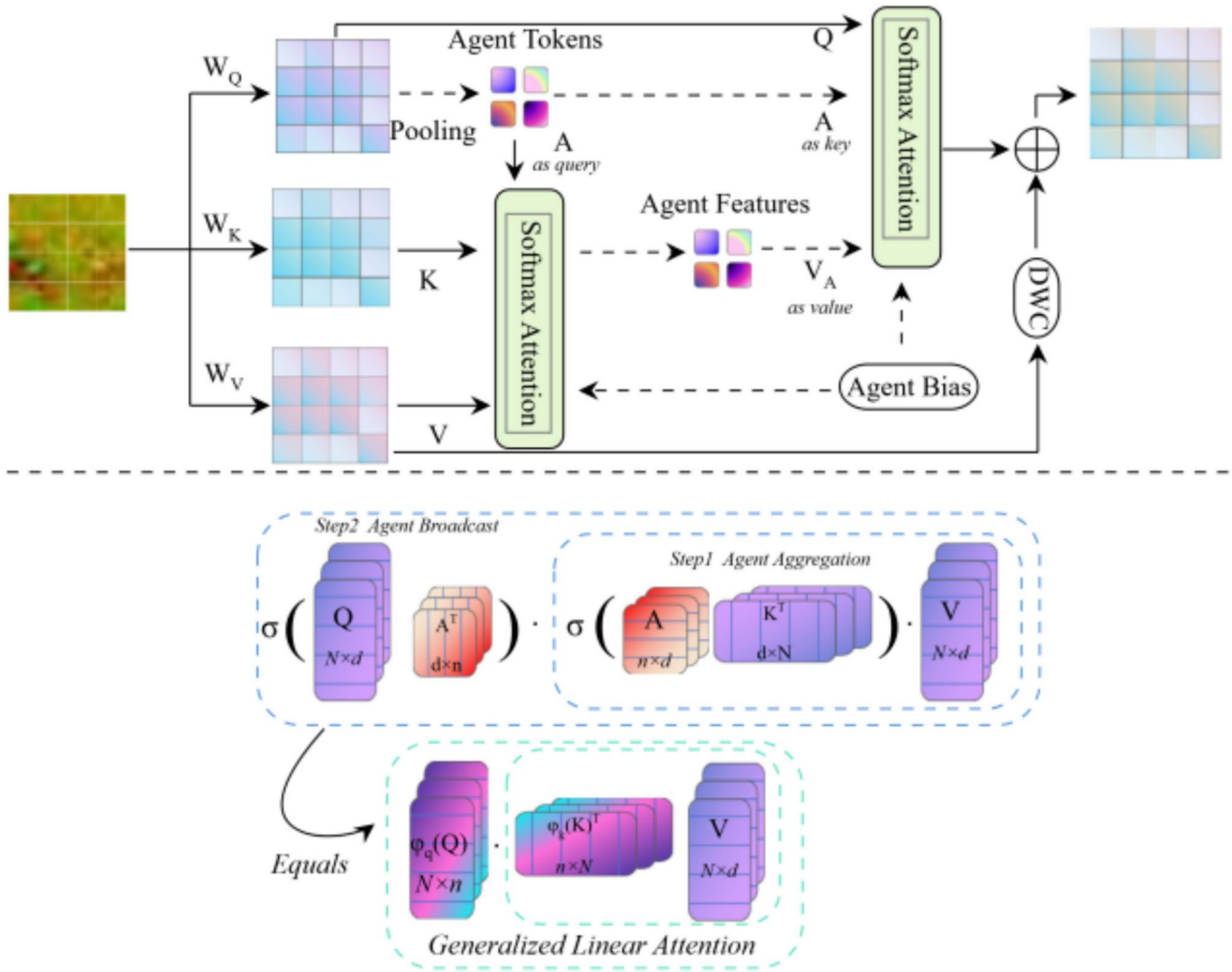

Fig. 3. The schematic diagram of agent attention.

Diffusion model-based data augmentation

The micro-expression datasets suffer from category imbalance and limited data distribution. To overcome the limitations posed by the small and imbalanced nature of micro-expression datasets, we employ a diffusion-model-based data augmentation technique. This method introduces diversity into the training data by gradually adding noise to the images and then training the model to reverse this process, generating new samples that resemble the original data while incorporating diverse expressions.

Specifically, an original micro-expression image x_0 is first initialized, followed by the selection of a diffusion time step T and the definition of a noise schedule β_t, which typically follows a cosine incremental strategy to ensure varying noise intensity at each time step. A sequence of Gaussian noise samples $\epsilon_t \sim \mathcal{N}(0, I)$ is then generated to progressively perturb the image. Subsequently, at each time step ttt, the perturbation process is executed according to a predefined noise scheduling rule, which is mathematically formulated as follows:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon_t \tag{8}$$

where $\alpha_t = \prod_{i=1}^{t} (1 - \beta_i)$ represents the cumulative noise attenuation coefficient. This formula indicates that, at each step, the contribution of the original image x_0 gradually diminishes, while the influence of the noise ϵ_t progressively increases, ultimately resulting in pure Gaussian noise at $t = T$. Finally, a denoising model (e.g., a diffusion model with a U-Net architecture) is trained to predict either the noise ϵ_t or the clean image x_0 directly. The optimization is performed using a mean square error (MSE) loss function:

$$\mathcal{L} = \mathbb{E}_{x_0, \epsilon, t}[||\epsilon - f_\theta(x_t, t)||^2] \tag{9}$$

The clear image is gradually restored by reverse denoising during inference, using the following update rule:

$$x_{t-1} = \frac{1}{\sqrt{1 - \beta_t}} (x_t - \beta_t f_\theta(x_t, t)) + \sigma_t z \tag{10}$$

where σ_t represents the coefficient associated with noise intensity, and $z \sim \mathcal{N}(0, I)$ denotes the random noise used for sampling.

Integration of components

The final model integrates the LAPE, ESAAT, and data augmentation techniques to establish a robust micro-expression recognition framework. The LAPE module provides the model with enhanced spatial awareness, the ESAAT module optimizes the attention mechanism for efficiency, and the data augmentation technique expands the dataset, thereby enhancing the model's generalization ability.

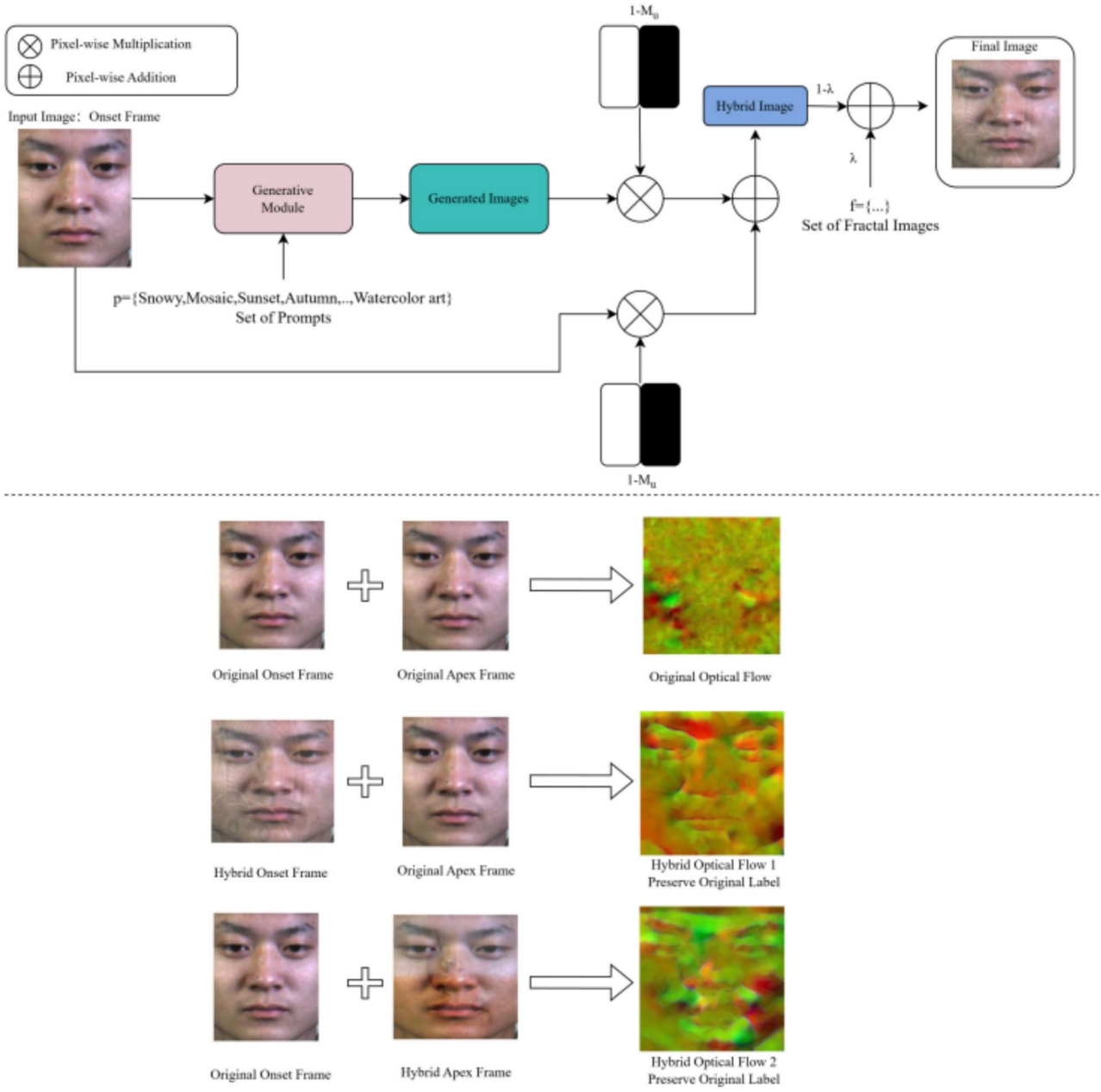

Fig. 4. Data augmentation based on diffusion model.

Data availability

The datasets described in the experiments of this paper are all publicly available datasets. The following statement contains information about the datasets and compared algorithms used in this paper: **Software and Algorithms:** (1) LBP-TOP is available at the URL: https://github.com/estrm/lbptop-emotion-recognition. (2) Bi-WOOF is available at the URL: https://github.com/christy 1206/biwoof. (3) STSTNet is available at the URL: https://github.com/christy1206/STSTNet. (4) MobileViT is available at the URL: https://github.com/wilile26811249/MobileViT. (5) MMNet is available at the URL: https://github.com/hyperconnect/MMNet. (6) Micro-BERT is available at the URL: https://github.com/uark-cviu/Micron-BERT. **Dataset:** (1) SMIC is available at the URL: https://www.oulu.fi/cmvs/node/41319. (2) SAMM is available at the URL: http://www2.docm.mmu.ac.uk/STAFF/M.Yap/dataset.php. (3) CASME II is available at the URL: http://casme.psych.ac.cn/casme/c2. The source of images given in Figs. 2 and 4 are sourced from CASME II. (4) CAS(ME) 3 is available at the URL: http://casme.psych.ac.cn/casme/e4.

Received: 14 February 2025; Accepted: 14 April 2025
Published online: 21 April 2025

References

1. Ekman, P. Darwin, deception, and facial expression. *Ann. N. Y. Acad. Sci.* **1000**, 205–221 (2003).
2. Yan, W.-J., Wu, Q., Liang, J., Chen, Y.-H. & Fu, X. How fast are the leaked facial expressions: The duration of micro-expressions. *J. Nonverbal Behav.* **37**, 217–230 (2013).
3. Wu, F. et al. A micro-expression recognition network based on attention mechanism and motion magnification. *IEEE Trans. Affect. Comput.* **6**, 66 (2024).
4. Zhao, M., Gong, L. & Din, A. S. A review of the emotion recognition model of robots. *Appl. Intell.* **55**, 1–33 (2025).
5. Yang, P., Liu, Y. & Zhou, Y. Research on intelligent intensive care system based on micro-expression tracking and automated Rass scoring. In *Proceedings of the 2024 International Conference on Smart Healthcare and Wearable Intelligent Devices* 179–185 (2024).
6. Hu, J. et al. An effective model for predicting serum albumin level in hemodialysis patients. *Comput. Biol. Med.* **140**, 105054. https://doi.org/10.1016/j.compbiomed.2021.105054 (2022).
7. Ekman, P. & Friesen, W. V. Nonverbal leakage and clues to deception. *Psychiatry* **32**, 88–106 (1969).
8. Ojala, T., Pietikainen, M. & Harwood, D. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition* vol. 1 582–585 (IEEE, 1994).
9. Zhao, G. & Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 915–928 (2007).
10. Wang, Y., See, J., Phan, R. C.-W. & Oh, Y.-H. Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In *Computer Vision—ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1–5, 2014, Revised Selected Papers, Part I 12* 525–537 (Springer, 2015).
11. O'Donovan, P. Optical flow: Techniques and applications. *Int. J. Comput. Vis.* **1**, 26 (2005).
12. Liu, Y.-J. et al. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Trans. Affect. Comput.* **7**, 299–310 (2015).
13. Liong, S.-T., See, J., Wong, K. & Phan, R.C.-W. Less is more: Micro-expression recognition from video using apex frame. *Signal Process. Image Commun.* **62**, 82–92 (2018).
14. Ozdemir, B. & Pacal, I. A robust deep learning framework for multiclass skin cancer classification. *Sci. Rep.* **15**, 4938 (2025).
15. Ozdemir, B., Aslan, E. & Pacal, I. Attention enhanced inceptionnext based hybrid deep learning model for lung cancer detection. *IEEE Access* **6**, 66 (2025).
16. Bayram, B., Kunduracioglu, I., Ince, S. & Pacal, I. A systematic review of deep learning in mri-based cerebral vascular occlusion-based brain diseases. *Neuroscience* **6**, 66 (2025).
17. İnce, S., Kunduracioglu, I., Bayram, B. & Pacal, I. U-net-based models for precise brain stroke segmentation. *Chaos Theory Appl.* **7**, 50–60 (2024).
18. Patel, D., Hong, X. & Zhao, G. Selective deep features for micro-expression recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)* 2258–2263 (IEEE, 2016).
19. Peng, M., Wu, Z., Zhang, Z. & Chen, T. From macro to micro expression recognition: Deep learning on small datasets using transfer learning. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* 657–661 (IEEE, 2018).
20. Cakir, D., Yilmaz, G. & Arica, N. Enhanced facial action unit detection with adaptable patch sizes on representative landmarks. *Neural Comput. Appl.* **37**, 3777–3791 (2025).
21. Dosovitskiy, A. *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929 (2020).
22. Pacal, I., Ozdemir, B., Zeynalov, J., Gasimov, H. & Pacal, N. A novel cnn-vit-based deep learning model for early skin cancer diagnosis. *Biomed. Signal Process. Control* **104**, 107627 (2025).
23. Liu, Y. et al. Lightweight vit model for micro-expression recognition enhanced by transfer learning. *Front. Neurorobot.* **16**, 922761 (2022).
24. Wang, Z., Zhang, K., Luo, W. & Sankaranarayana, R. Htnet for micro-expression recognition. *Neurocomputing* **602**, 128196 (2024).
25. Zhang, L., Hong, X., Arandjelović, O. & Zhao, G. Short and long range relation based spatio-temporal transformer for micro-expression recognition. *IEEE Trans. Affect. Comput.* **13**, 1973–1985 (2022).
26. Li, Y., Wei, J., Liu, Y., Kauttonen, J. & Zhao, G. Deep learning for micro-expression recognition: A survey. *IEEE Trans. Affect. Comput.* **13**, 2028–2046 (2022).
27. Zhang, F. & Chai, L. A review of research on micro-expression recognition algorithms based on deep learning. *Neural Comput. Appl.* **36**, 17787–17828 (2024).
28. Han, D. et al. Agent attention: On the integration of softmax and linear attention. In *European Conference on Computer Vision* 124–140 (Springer, 2025).
29. Gao, D. et al. Resshift-4e: Improved diffusion model for super-resolution with microscopy images. *Electronics* **14**, 479 (2025).
30. Li, X., Pfister, T., Huang, X., Zhao, G. & Pietikäinen, M. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (fg)* 1–6 (IEEE, 2013).
31. Davison, A. K., Lansley, C., Costen, N., Tan, K. & Yap, M. H. Samm: A spontaneous micro-facial movement dataset. *IEEE Trans. Affect. Comput.* **9**, 116–129 (2016).
32. Qu, F. et al. Cas(me)2: A database for spontaneous macro-expression and micro-expression spotting and recognition. *IEEE Trans. Affect. Comput.* **9**, 424–436 (2017).
33. Li, J. et al. Cas(me)3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 2782–2800 (2022).
34. Gan, Y. S., Liong, S.-T., Yau, W.-C., Huang, Y.-C. & Tan, L.-K. Off-apexnet on micro-expression recognition system. *Signal Process. Image Commun.* **74**, 129–139 (2019).
35. Liong, S.-T., Gan, Y., See, J., Khor, H.-Q. & Huang, Y.-C. Shallow triple stream three-dimensional cnn ststnet for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)* 1–5 (IEEE, 2019).
36. Mehta, S. & Rastegari, M. *Light-Weight, General-purpose, and Mobile-Friendly Vision Transformer* (Mobilevit, 2021).
37. Seo, S. et al. *Towards real-time automatic portrait matting on mobile devices*. arXiv preprint arXiv:1904.03816 (2019).
38. Nguyen, X.-B. et al. Micron-bert: Bert-based facial micro-expression recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).
39. Hao, H. et al. *Hierarchical space-time attention for micro-expression recognition*. arXiv preprint arXiv:2405.03202 (2024).

Author contributions

Yibo Zhang: Writing—original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Weiguo Lin: Supervision, Resources, Project administration, Funding acquisition. Yuanfa Zhang: Writing—review and editing, Validation. Junfeng Xu: Writing—review and editing, Supervision, Resources, Project administration, Funding acquisition. Yan Xu: Writing—review and editing, Validation, Resources.

Funding
This work was supported in part by National Key Research and Development Program of China under Grant 2022YFF0902401, in part by the National Natural Science Foundation of China under Grant (No.62302467, No.62402459, and No.U2436208), in part by the Project of Guangdong Key Laboratory of Industrial Control System Security (2024B1212020010), in part by the Fundamental Research Funds for the Central Universities and the Public Computing Cloud, CUC.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to W.L., J.X. or Y.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.