Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Facial micro-expression recognition based on motion magnification network and graph attention mechanism

Falin Wu [*], Yu Xia, Tiangyang Hu, Boyi Ma, Jingyao Yang, Haoxin Li

SNARS Laboratory, School of Instrumentation and Optoelectronic Engineering, Beihang University, No. 37, XueYuan Road, HaiDian District, Beijing, 100191, China

ARTICLE INFO

Keywords:
Micro-expression recognition
Graph attention mechanism
Facial motion magnification
Transformer

ABSTRACT

Micro-expression is extensively studied due to their ability to fully reflect individuals' genuine emotions. However, accurate micro-expression recognition is a challenging task due to the subtle motion of facial muscle. Therefore, this paper introduces a Graph Attention Mechanism-based Motion Magnification Guided Micro-Expression Recognition Network (GAM-MM-MER) to amplify delicate muscle motions and focus on key facial landmarks. First, we propose a Swin Transformer-based network for micro-expression motion magnification (ST-MEMM) to enhance the subtle motions in micro-expression videos, thereby unveiling imperceptible facial muscle movements. Then, we propose a graph attention mechanism-based network for micro-expression recognition (GAM-MER), which optimizes facial key area maps and prioritizes adjacent nodes crucial for mitigating the influence of noisy neighbors, while attending to key feature information. Finally, experimental evaluations conducted on the CASME II and SAMM datasets demonstrate the high accuracy and effectiveness of the proposed network compared to state-of-the-art approaches. The results of our network exhibit significant superiority over existing methods. Furthermore, ablation studies provide compelling evidence of the robustness of our proposed network, substantiating its efficacy in micro-expression recognition.

1. Introduction

Facial expressions, including macro-expressions and micro-expressions, play a crucial role in understanding emotions, which are vital in human cognition and human-machine interaction [1,2]. Macro-expressions are easily recognized but can be disguised, while micro-expressions, lasting from 4 milliseconds to half a second, are spontaneous and difficult to conceal, making them a more accurate reflection of genuine emotions. Despite Paul Ekman Group [3] produced the first Micro-Expression Training Tool (METT) in 2003, it is still challenged to recognize micro-expressions with the naked eye. Undergraduate students can achieve a accuracy rate of 40%, while the US Coast Guard can attain 50% accuracy without additional assistance. Therefore, computer-based automatic micro-expression recognition is broadly studied.

In the early stages of MER studies, hand-crafted approaches were employed. Yeasin et al. [4] proposed a two-step classification method that relied on refined optical flow computation from image sequences. Zhao and Pietikainen [5] developed a feature encoder utilizing volume local binary patterns (VLBP) to model textures. Moreover, they employed Local Binary Pattern histograms from Three Orthogonal Planes (LBP-TOP) to discriminate regional texture features. Another hand-crafted approach, introduced by [6],

[*] Corresponding author.
E-mail address: falin.wu@buaa.edu.cn (F. Wu).

https://doi.org/10.1016/j.heliyon.2024.e35964
Received 16 March 2024; Received in revised form 5 August 2024; Accepted 7 August 2024

Available online 12 August 2024

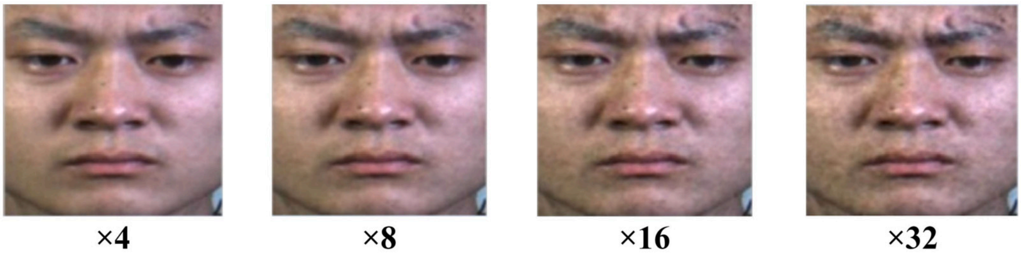


Fig. 1. Magnified results of traditional approaches at different magnification factor.

was the 3D Histogram of Oriented Gradient (3DHOG), which extracted features from muscle movement in three dimensions for MER. Liong et al. [7] presented Bi-Weighted Oriented Optical Flow (Bi-WOOF) to encode the essential expressiveness of vertex frames. Van Quang et al. [8] developed a CapsuleNet specifically designed for MER using individual vertex frames. These studies have made significant contributions to the field of MER through their hand-crafted feature extraction techniques.

In recent years, deep learning has witnessed remarkable advancements, leading to the proposal of numerous deep learning networks that have achieved outstanding performance in MER. Kim et al. [9] introduced ELRCN, which leveraged CNN and LSTM to extract spatial and temporal characteristics from diverse micro-expressions. Khor et al. [10] proposed the Dual-Stream Shallow Network (DSSN) that combined CNNs with inputs based on mixed movement to learn expressive characteristics. Gan et al. [11] presented the Optical Flow Features from Apex frame Network (OFF-ApexNet), which combined features extracted from micro-expressions.

Graph Convolutional Networks (GCNs) have also shown promise in MER tasks by leveraging relationships among landmarks. Lei et al. [12] proposed a graph temporal convolutional network (Graph-TCN), which encodes regional facial movement features of micro-expressions utilizing separate channels for node and edge features. Xie et al. [13] integrated emotion labels and action units in their MER approach, modeling different action units based on relevant information and combining action unit recognition with MER. Kumar and Bhanu [14] introduced the landmark-assisted three-stream graph attention network (GACNN), which dynamically selected varying patch sizes. Graph Convolution Networks (GCNs), which exploit landmark relationships, consider this relationship and outperform CNNs in MER tasks. Recently, Transformers have made remarkable advancements in computer vision [15–18]. The attention mechanism in Transformers plays a critical role in selectively focusing on corresponding sections of face landmarks and extracting profound relationships among them. Therefore, we combine the attention mechanism of Transformers with the facial key region emphasized by GCNs, using the Graph Attention Network (GAT) to improve the accuracy of MER.

MER is a challenging task due to the slight facial muscle movements. Traditional motion magnification algorithms, such as Eulerian [19] and Lagrangian approaches [20], have been used but suffer from noticeable blurring and artifacts, as depicted in Fig. 1. To address this issue, a micro-expression motion magnification network based on the Swin Transformer is proposed. However, increasing the magnification factor amplifies noise from irrelevant movements. To mitigate this issue, key facial landmarks in unrelated areas are eliminated. This paper proposes a graph attention mechanism-based motion magnification guided micro-expression recognition network (GAM-MM-MER). This paper proposes a Swin Transformer-based micro-expression motion magnification network (ST-MEMM) to enhance subtle motions while reducing blurring and artifacts. Additionally, this paper proposes a graph attention mechanism-based micro-expression recognition network (GAM-MER) optimizes facial key area maps and focuses on adjacent nodes to mitigate the influence of noisy neighbors and extract essential feature information. The contributions of this paper are as follows:

1) We propose a Swin Transformer-based micro-expression motion magnification network, the first to employ a Transformer module for magnifying micro-expression motion, which reduces blurring and artifacts while enhancing the magnification factor of micro-expressions.

2) We propose a graph attention mechanism-based micro-expression recognition network, which optimizes facial key area maps and prioritizes adjacent nodes to reduce the influence of noisy neighbors and emphasize essential feature information.

3) We conduct experiment on CASME II and SAMM datasets and achieve better results than the state-of-the-art methods.

The rest of the paper is organized as follows: First, the proposed video motion magnification network ST-MEMM is introduced in Section 2.1. Secondly, the proposed micro-expression recognition network GAM-MER is introduced in Section 2.2. Then, the experiment results and analysis are shown in Section 3. Finally, conclusion and future work are provided in Section 4.

2. Methodology

The overall architecture of our proposed framework is depicted in Fig. 2. The framework comprises two main modules: the video motion magnification network and the micro-expression recognition network. The detailed description and functioning of each network will be presented in Sections 2.1 and 2.2, respectively.

2.1. Video motion magnification network

The proposed ST-MEMM network enhances subtle movements in micro-expression videos and reveals imperceptible facial muscle motion. The ST-MEMM network comprises of three sections: the Encoder module $\mathrm{M_e}$, the Manipulator module $\mathrm{M_m}$, and the Decoder module $\mathrm{M_d}$, as shown in Fig. 3.

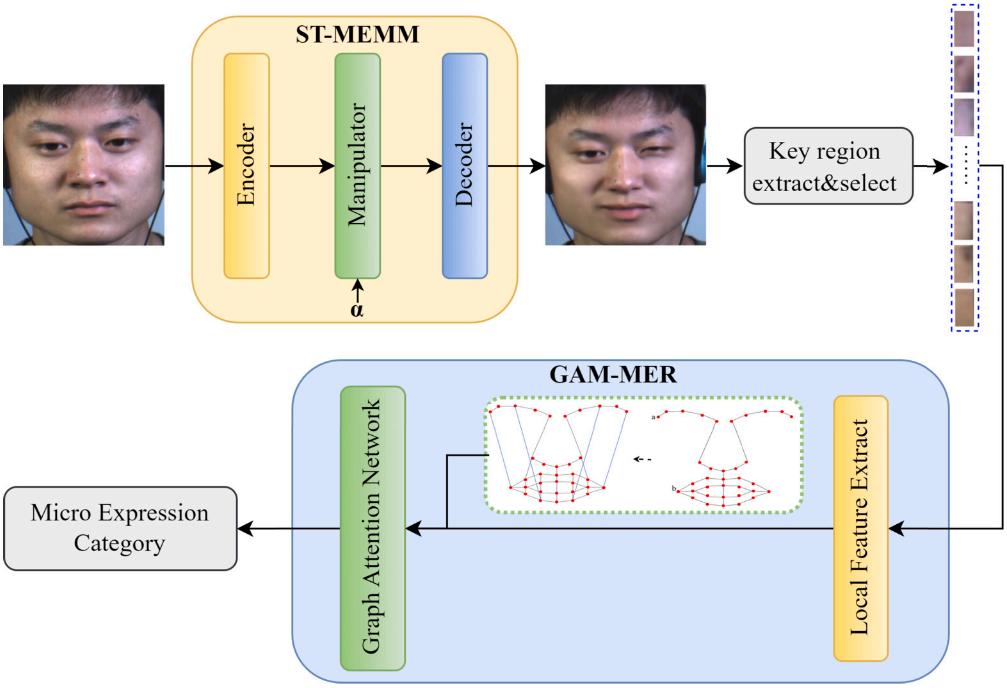


Fig. 2. The overall architecture of GAM-MM-MER.

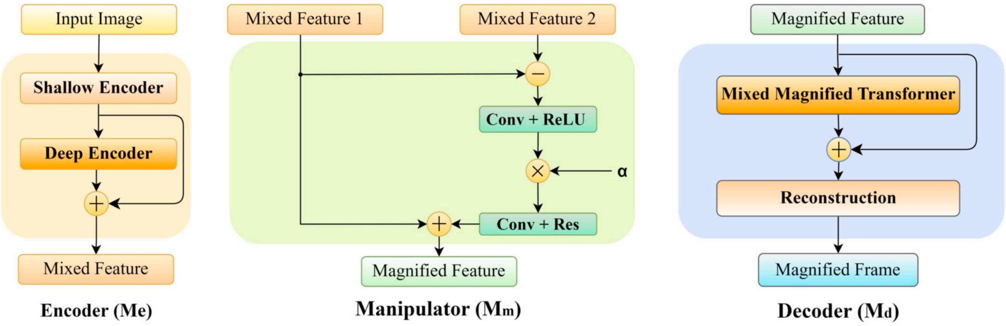


Fig. 3. Swin Transformer-based micro-expression motion magnification network (ST-MEMM) architecture.

First, the Encoder module $\mathrm{M_e}$ extracts shallow features using the Shallow Encoder block $\mathrm{E_S}(\cdot)$ and deep features using the Deep Encoder block $\mathrm{E_D}(\cdot)$. Mathematically, the Encoder module is formulated in Equation (1) and Equation (2):

$$[\mathbf{F}_1(\mathrm{S}),\mathbf{F}_2(\mathrm{S})] = \mathrm{E_S}([\mathbf{I}_1,\mathbf{I}_2]) \tag{1}$$

$$[\mathbf{F}_1(\mathrm{D}),\mathbf{F}_2(\mathrm{D})] = \mathrm{E_D}([\mathbf{I}_1,\mathbf{I}_2]) \tag{2}$$

where $\mathbf{I}_1$ and $\mathbf{I}_2$ are input images. $\mathbf{F}_1(\mathrm{S})$ and $\mathbf{F}_2(\mathrm{S})$ are shallow features extracted with the Shallow Encoder block. $\mathbf{F}_1(\mathrm{D})$ and $\mathbf{F}_2(\mathrm{D})$ are deep features extracted with the Deep Encoder block.

Subsequently, the extracted shallow and deep features are mixed and transferred to the Manipulator module $\mathrm{M_m}$ for further feature manipulation. Mathematically, the Manipulator module is formulated in Equation (3):

$$\mathbf{F}_\mathrm{M} = \mathbf{F}_1(\mathrm{MIX}) + \mathrm{M}_{\mathrm{ConvRes}}(\alpha \cdot \mathrm{M}_{\mathrm{ConvReLU}}(\mathbf{F}_2(\mathrm{MIX}) - \mathbf{F}_1(\mathrm{MIX}))) \tag{3}$$

where $\mathbf{F}_\mathrm{M}$ are amplified features and α is adjustable magnification factor. $\mathbf{F}_1(\mathrm{MIX})$ is combination of $\mathbf{F}_1(\mathrm{S})$ and $\mathbf{F}_1(\mathrm{D})$, and $\mathbf{F}_2(\mathrm{MIX})$ is combination of $\mathbf{F}_2(\mathrm{S})$ and $\mathbf{F}_2(\mathrm{D})$.ConvRes is combination of Convolution and Residual block, and ConvReLU is combination of Convolution and ReLU block.

Finally, after amplifying the difference of features between both frames, the magnified features are transferred to the Decoder module $\mathrm{M_d}$. This module comprises of the Mixed Magnified Transformer module $\mathrm{M_{MMTB}}$ to further magnify and mix the feature, and the Reconstruction module to generate final magnified frame. Mathematically, the Mixed Magnified Transformer module is formulated in Equation (4):

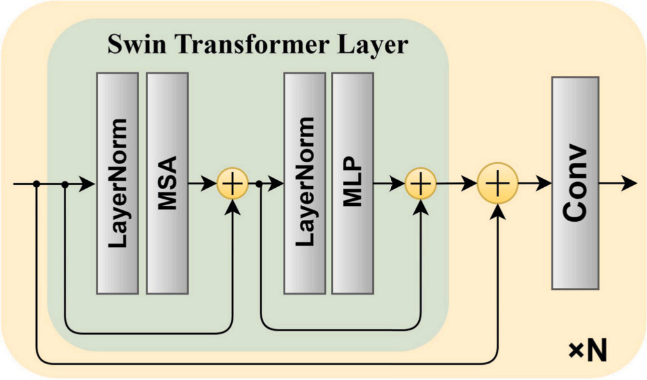


Fig. 4. Residual Swin Transformer Block.

$$\mathbf{F}_{\mathrm{MMTB}} = \mathrm{M}_{\mathrm{MMTB}}(\mathbf{F}_{\mathrm{M}}) \tag{4}$$

where $\mathrm{M}_{\mathrm{MMTB}}$ is the Mixed Magnified Transformer block, and $\mathbf{F}_{\mathrm{MMTB}}$ is the mixed magnified feature. Finally, the combination of magnified features and further magnified features are transferred to the Reconstruction module $\mathrm{R}_{\mathrm{SE}_{\mathrm{R}}}$, which reverses the shallow feature extractor to generate the final magnified frame. Mathematically, the Reconstruction module is formulated in Equation (5):

$$\mathbf{I}_{\mathrm{G}} = \mathrm{M}_{\mathrm{Conv}}(\mathrm{M}_{\mathrm{ConvTrans}}(\mathbf{F}_{\mathrm{MMTB}} + \mathbf{F}_{\mathrm{M}})) \tag{5}$$

where $\mathrm{M}_{\mathrm{ConvTrans}}$ is Transpose Convolution module, and $\mathrm{M}_{\mathrm{Conv}}$ is Convolution module. $\mathbf{I}_{\mathrm{G}}$ is final generated magnified image of Swin Transformer-based micro-expression motion magnification network (ST-MEMM).

As depicted in Fig. 4, the Residual Swin Transformer block (RSTB) is utilized to capture deep features from input frames in the Deep Encoder module M_{e}, and to further magnify features in the Mixed Magnified Transformer module M_{m}. RSTB consists of multiple Swin Transformer layers and a convolution layer, facilitating the processing of multilevel features.

The Swin Transformer layer partitions an input frame of size $\mathrm{H} \times \mathrm{W} \times \mathrm{C}$ into non-overlapping windows of size $\mathrm{M} \times \mathrm{M}$ and calculates local attention within single window. The Multi-head Self Attention mechanism involves multiple iterations of the attention mechanism, and the resulting output is subsequently passed through a Multi-Layer Perception.

2.2. Micro-expression recognition network

Magnified micro-expression images are transferred to the graph attention mechanism-based network for micro-expression recognition (GAM-MER). The architecture of GAM-MER network is shown in Fig. 5.

First, proposed ST-MEMM network $\mathrm{N}_{\mathrm{ST-MEMM}}$ output the amplified micro-expression images. Mathematically, the ST-MEMM network is formulated in Equation (6):

$$\mathrm{I}_{\mathrm{M}} = \mathrm{N}_{\mathrm{ST-MEMM}}(\mathrm{I}_{\mathrm{R}}) \tag{6}$$

where I_{R} is the input raw micro-expression image and I_{R} is the amplified image.

Subsequently, key facial landmark regions are extracted. Furthermore, in MER task, it is unnecessary to focus on all extracted facial landmark regions, as some of these regions may contain feature information that is irrelevant to micro-expressions or even introduce distracting information. For instance, non-voluntary behaviors such as head movements, blinking, and eye movements are unrelated to facial muscle movements caused by micro-expressions. Moreover, these interfering factors can be further amplified by proposed ST-MEMM network. Additionally, facial muscle micro-movements induced by micro-expressions primarily manifest around the eyebrows, nose tip, and mouth, while they rarely affect the muscle movements in the facial contour region. Consequently, we only concentrate on the facial landmark regions that are relevant to micro-expressions. Mathematically, the region select and extract is formulated in Equation (7):

$$\mathrm{I}_{\mathrm{S}} = \mathrm{M}_{\mathrm{select}}(\mathrm{M}_{\mathrm{extract}}(\mathrm{I}_{\mathrm{M}})) \tag{7}$$

where I_{R} is selected facial landmark regions. $\mathrm{M}_{\mathrm{dlib}}$ and $\mathrm{M}_{\mathrm{select}}$ are separately employed to extract key facial landmark regions and select regions relevant to micro-expressions.

Consequently, selected facial landmark regions are transferred to shallow feature extraction module to extract local feature of each region. Due to the small size and the specific requirement for extracting local features from facial key regions, CNN is selected as the basic module for local feature extraction. Mathematically, the local feature extraction module is formulated in Equation (8):

$$\mathbf{F}_{\mathbf{L}} = \mathrm{M}_{\mathrm{LFE}}(\mathrm{I}_{\mathrm{S}}) \tag{8}$$

where $\mathrm{M}_{\mathrm{LFE}}$ is the local feature extraction module and $\mathbf{F}_{\mathbf{L}}$ is extracted local feature of each facial landmark region.

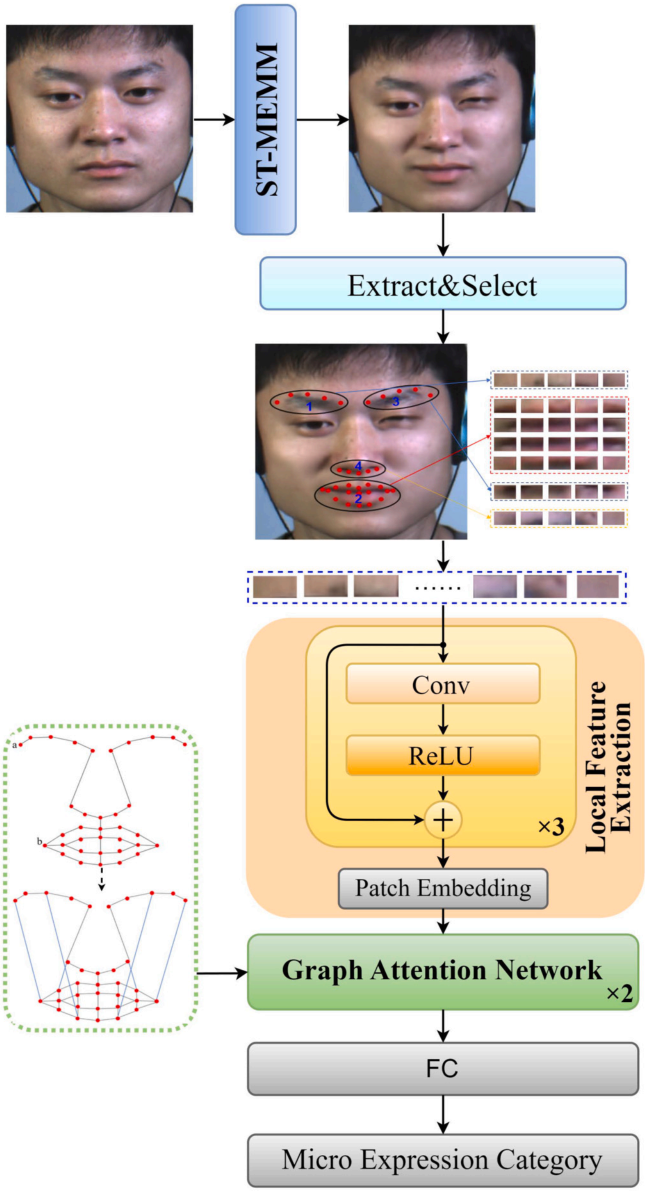


Fig. 5. Graph attention mechanism-based network for micro-expression recognition (GAM-MER) architecture.

Finally, the existing facial key region graph structure in current research is challenging to sufficiently capture the intrinsic connections between distantly located but strongly correlated nodes. Considering prior knowledge, the topological structure of the facial key region graph is optimized. Based on the optimized facial landmark region map, local feature of each region is transferred to the two-layer graph network and output the micro-expression category. Mathematically, the graph attention network is formulated in Equation (9):

$$\mathrm{C}_{\mathrm{MER}} = \mathrm{M}_{\mathrm{GAT}}(\mathbf{F}_{\mathbf{L}}) \tag{9}$$

where $\mathrm{C}_{\mathrm{MER}}$ is the output category of MER and $\mathrm{M}_{\mathrm{GAT}}$ is the module of graph attention network.

3. Experiments

This section presents an extensive analysis of experiments, including datasets, evaluation metrics, experiment detail, experiment results, and ablation study.

3.1. Datasets

3.1.1. Video motion magnification dataset

The Swin Transformer-based micro-expression motion magnification network (ST-MEMM) is trained using the selected dataset introduced by Oh et al. [21]. This dataset has demonstrated effectiveness in generating high-quality models that exhibit strong generalization capabilities, even when applied to scenarios unrelated to the training dataset. Oh et al. [21] constructed their dataset by utilizing 200,000 images from MS COCO dataset [22] as background and incorporating 7,000 segmented objects from the PASCAL VOC dataset [23] as foreground.

3.1.2. Micro-expression recognition datasets

To assess the performance of the graph attention mechanism based micro-expression recognition network (GAM-MER), experiments are conducted on publicly available datasets, namely CASME II [24]. CASME II dataset consists of 247 micro-expression videos captured at a resolution of 640×480 and cropped to a size of 280×340 for the face. The dataset encompasses videos from 35 participants, and it is categorized into six distinct facial micro-expression categories. The videos are provided in RGB format, and the average age of participants is 22.03 years.

Similarly, experiments are also conducted on publicly available datasets SAMM [25,6,26], which comprises 159 micro-expression videos captured at a resolution of 2040×1088, with the face cropped to a size of 400×400. The dataset includes videos from 32 participants, and is divided into eight categories of micro-expressions. The SAMM dataset consists of gray-scale video samples from individuals representing 13 different ethnicities, and the average age of participants is 33.24 years.

Meanwhile, we conduct experiments on the publicly available SMIC-HS dataset [27]. The SMIC-HS dataset comprises 164 micro-expression videos captured at a resolution of 1280×720, with the face cropped to a size of 190×230. The dataset contains videos from 16 participants and is categorized into three distinct micro-expression categories. Notably, the SMIC-HS dataset consists of grayscale video samples from individuals representing two different ethnicities, with an average participant age of 28.1 years.

3.2. Performance metrics

$\mathrm{F}_{1-\mathrm{score}}$ and A_{cc} are common metrics for evaluating MER network. Mathematically, the calculation of A_{cc} is formulated in Equation (10):

$$\mathrm{A}_{\mathrm{cc}} = \frac{\mathrm{P}}{\mathrm{N}} \times 100\% \tag{10}$$

where A_{cc} is calculated by dividing the sum of correct prediction results P by the total number of test data N. Additionally, the accuracy for each category is calculated.

The $\mathrm{F}_{1-\mathrm{score}}$ assigns equal weight to each category of test data. Based on the confusion matrix, True Positives TP_{C}, False Positives FP_{C}, and False Negatives FN_{C} are calculated for each category of micro-expressions. The final balanced $\mathrm{F}_{1-\mathrm{score}}$ is obtained by averaging the $\mathrm{F}_{1-\mathrm{score}}$ of each category. Mathematically, the calculation of $\mathrm{F}_{1\mathrm{C}}$ and UF1 is formulated in Equation (11) and Equation (12):

$$\mathrm{F}_{1\mathrm{C}} = \frac{2 \times \mathrm{TP}_{\mathrm{C}}}{2 \times \mathrm{TP}_{\mathrm{C}} + \mathrm{FP}_{\mathrm{C}} + \mathrm{FN}_{\mathrm{C}}} \tag{11}$$

$$\mathrm{UF1} = \frac{\mathrm{F}_{1\mathrm{C}}}{\mathrm{C}} \tag{12}$$

where $\mathrm{F}_{1\mathrm{C}}$ represents the $\mathrm{F}_{1-\mathrm{score}}$ for each category, and C denotes the number of categories.

3.3. Experiment detail

Motion magnification is applied at various magnification factors (8, 10, 12, and 15) to augment the dataset and mitigate the category imbalance during training. Additionally, data augmentation techniques such as random brightness, contrast, and color transformations are employed to mitigate the issue of overfitting.

The ST-MEMM model utilizes the ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a batch size of 4 and a learning rate of 10^{-5} without weight decay. The GAM-MER model also employs the ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The experiments are conducted on a workstation has Ubuntu 16.04 operating system and three NVIDIA GeForce GTX 3090 GPUs.

3.4. Experiment results

As depicted in Tables 1 and 2, a comparative analysis is conducted to evaluate proposed network in comparison to baseline handcrafted feature extraction approaches and state-of-the-art deep learning-based networks on CASME II, SAMM and SMIC-HS datasets. The evaluation is performed for three and five categories of micro-expressions.

Table 1 provides a detailed comparison between our proposed network and outstanding deep learning-based networks on CASME II and SAMM datasets, focusing on three categories of MER. On CASME II dataset, our proposed GAM-MM-MER network achieves an accuracy of 91.57% representing a 0.79% improvement over the prior leading network (SelfME [17]). However, our $\mathrm{F}_{1-\mathrm{score}}$ is 1.68% lower compared to the prior leading network (SelfME [17]). On SAMM dataset, our proposed GAM-MM-MER network achieves an

Table 1
Comparison with state-of-the-art methods. Experiments are conducted on CASME II, SAMM and SMIC-HS datasets with 3 categories: Negative, Positive, and Surprise.

Method	Feature Extract	CASME II		SAMM		SMIC-HS	
		Acc	F_1	Acc	F_1	Acc	F_1
LBP-TOP [28]	Handcrafted	0.7026	0.7429	0.3954	0.4102	0.2000	0.5280
Bi-WOOF [7]	Handcrafted	0.7805	0.8026	0.5211	0.5139	0.5727	0.5829
CapsuleNet [8]	Handcrafted	0.7068	0.7018	0.6209	0.5989	0.5820	0.5877
SSSN [10]	CNN	0.7119	0.7151	0.5662	0.4513	0.6341	0.6329
DSSN [10]	CNN	0.7080	0.7300	0.5735	0.4664	0.6341	0.6462
OFF-ApexNet [11]	CNN	0.8764	0.8680	0.5409	0.5409	0.6817	0.6695
Graph-TCN [12]	GCN	0.8648	0.8871	0.8050	0.7657	N/A	N/A
AU-GACN [13]	GCN	0.7120	0.3550	0.7020	0.4330	N/A	N/A
GACNN [14]	GCN	0.8966	0.8695	0.9098	0.8463	N/A	N/A
μ-BERT [15]	Transformer	0.9034	0.8914	N/A	N/A	0.8580	0.8384
SelfME [17]	Transformer	0.9078	**0.9290**	N/A	N/A	0.6970	0.7012
GAM-MM-MER (Ours)	Transformer	**0.9157**	0.9122	**0.9125**	**0.9134**	**0.8622**	**0.8710**

Table 2
Comparison with state-of-the-art methods. Experiments are conducted on CASME II dataset with 5 categories: Disgust, Happy, Surprise, Repression, and Other, and SAMM dataset with 5 categories: Anger, Happy, Surprise, Contempt, and Other.

Method	Feature Extract	CASME II		SAMM	
		Acc	F_1	Acc	F_1
LBP-TOP [28]	Handcrafted	0.3968	0.3589	0.3556	0.3589
LBP-SIP [29]	Handcrafted	0.4656	0.4480	N/A	N/A
STLBP-IP [30]	Handcrafted	0.6397	0.6125	N/A	N/A
Bi-WOOF [7]	Handcrafted	0.5789	0.6100	N/A	N/A
CNN-LSTM [9]	CNN	0.6098	N/A	N/A	N/A
SSSN [10]	CNN	0.7119	0.7151	0.5662	0.4513
DSSN [10]	CNN	0.7078	0.7297	0.5735	0.4644
Graph-TCN [12]	GCN	0.7398	0.7246	0.7500	0.6985
AU-GACN [13]	GCN	0.7427	0.7047	0.7426	0.7045
GACNN [14]	GCN	0.8252	0.7517	**0.8971**	0.8365
μ-BERT [15]	Transformer	0.8553	0.8348	0.8386	0.8475
C3Dbed [17]	Transformer	0.7764	0.7520	0.8126	0.8067
GAM-MM-MER (Ours)	Transformer	**0.8609**	**0.8655**	0.8788	**0.8817**

accuracy of 91.25%, representing a 0.27% improvement over the prior leading network (GACNN [14]), and an $F_{1-\text{score}}$ of 91.34%, representing a 6.71% improvement compared to the prior leading network (GACNN [14]). On SMIC-HS dataset, our proposed GAM-MM-MER network achieves an accuracy of 86.22%, representing a 1.30% improvement over the prior leading network (GACNN [14]), and an $F_{1-\text{score}}$ of 87.10%, representing a 3.26% improvement compared to the prior leading network (GACNN [14]).

With an increase of micro-expression categories from three to five, Table 2 provides a detailed comparison between our proposed approach and outstanding deep learning-based networks. On CASME II dataset, our proposed GAM-MM-MER network achieves an accuracy of 86.09% and an $F_{1-\text{score}}$ of 86.55%, representing a 0.56% and a 3.07% improvement over the prior leading network (μ-BERT [15]), respectively. On SAMM dataset, our proposed GAM-MM-MER network achieves an accuracy of 87.88%, 1.83% lower compared to the prior leading network (GACNN [14]), and an $F_{1-\text{score}}$ of 90.24%, representing a 3.42% improvement compared to the prior leading network (μ-BERT [15]).

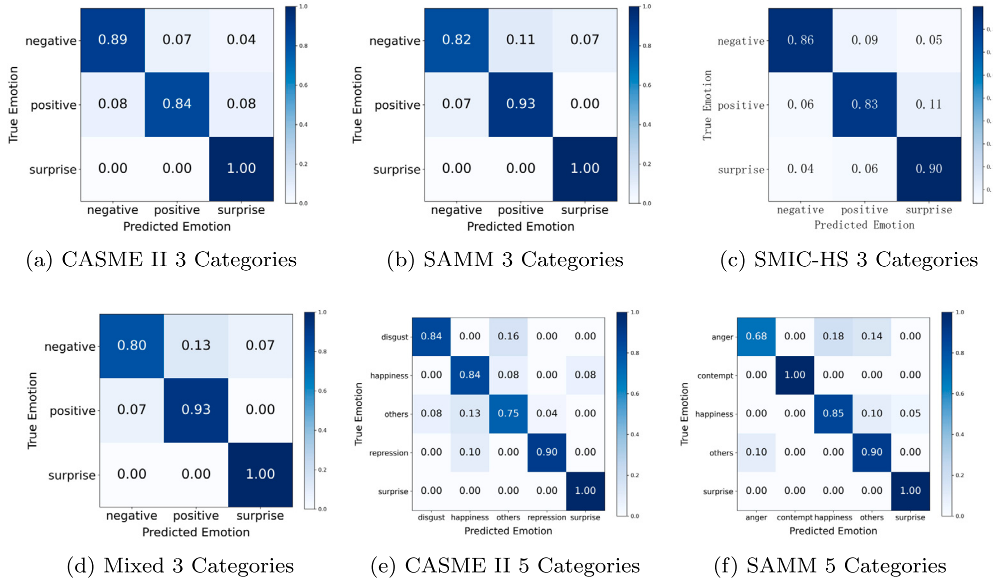
Fig. 6 displays the confusion matrix on the CASME II and SAMM datasets with three and five categories, and on the SMIC-HS and mixed datasets with three categories. Each sub-figure corresponds to a different category and dataset, labeled from (a) to (f).

3.5. Ablation study

To evaluate the effectiveness of proposed network, ablation experiments are performed, specifically focusing on analyzing the improvement of the accuracy and $F_{1-\text{score}}$ with GAM-MER Architecture and magnification factor.

3.5.1. GAM-MER architecture

The ablation study conducted in this research demonstrates that the two-layer graph attention network achieves optimal results, as shown in Table 3. To ensure a fair comparison, we utilize identical magnified frames from the ST-MEMM network and maintain the same architecture of the GAM-MER network while adding one layer of graph attention network or removing one layer of graph attention network. In addition, we employ the graph convolution network instead of the graph attention network to showcase the improvement brought by the attention mechanism. Furthermore, to demonstrate the enhancement achieved by selecting facial key areas closely related to micro-expressions, we employ all facial key areas instead of selected key areas. Moreover, we compare the



(a) CASME II 3 Categories (b) SAMM 3 Categories (c) SMIC-HS 3 Categories

(d) Mixed 3 Categories (e) CASME II 5 Categories (f) SAMM 5 Categories

Fig. 6. Confusion matrix on the CASME II and SAMM datasets with 3 and 5 categories, and on the SMIC-HS and mixed datasets with 3 categories. Each sub-figure represents a different category and dataset, labeled from (a) to (f).

Table 3
Ablation study of GAM-MER architecture on the CASME II and SAMM datasets.

Method	CASME II 3 *Cat.*		CASME II 5 *Cat.*		SAMM 3 *Cat.*		SAMM 5 *Cat.*	
	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
Add one layer GAT	0.8957	0.8916	0.8261	0.8237	0.8750	0.8709	0.8182	0.8198
Remove one layer GAT	0.8795	0.8780	0.8087	0.8068	0.8500	0.8476	0.7913	0.7879
Graph convolution network	0.8535	0.8543	0.7780	0.7819	0.8467	0.8428	0.7692	0.7705
All facial key areas	0.8634	0.8675	0.8183	0.8204	0.8431	0.8406	0.8302	0.8267
Raw facial graph	0.9046	0.9010	0.8495	0.8512	0.9013	0.9034	0.8660	0.8637
Baseline	**0.9157**	**0.9122**	**0.8609**	**0.8655**	**0.9125**	**0.9134**	**0.8788**	**0.8817**

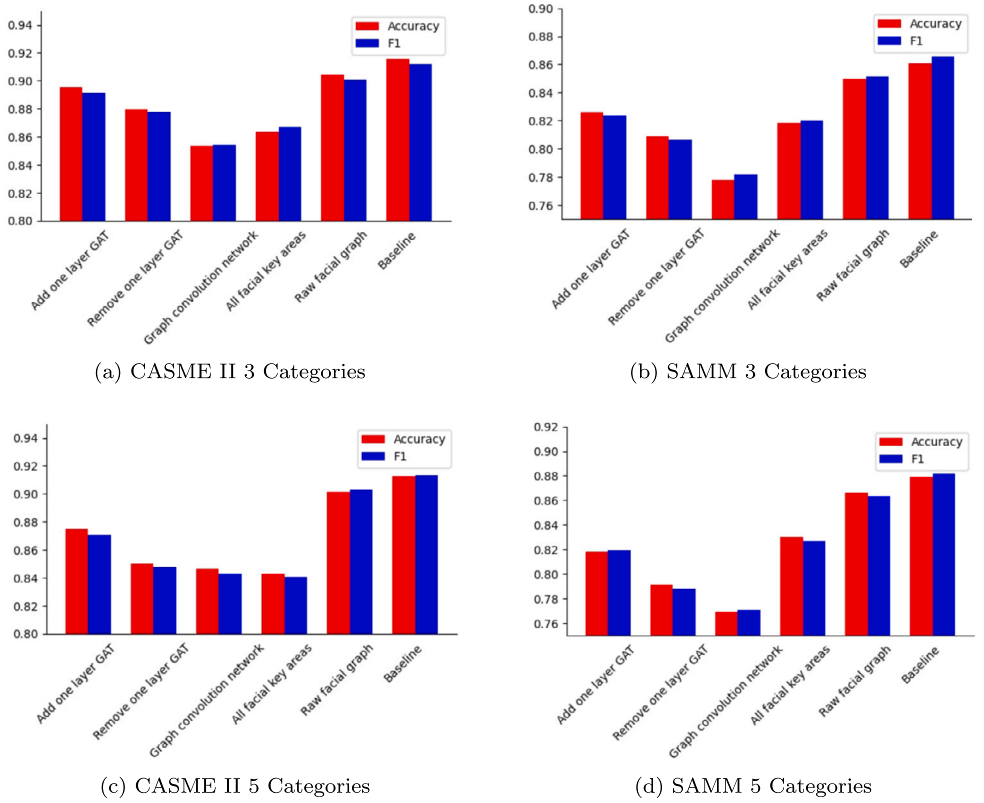
Table 4
Ablation study of magnification factor α on the CASME II and SAMM datasets for 3 and 5 micro-expression categories.

Method	CASME II 3 *Cat.*		CASME II 5 *Cat.*		SAMM 3 *Cat.*		SAMM 5 *Cat.*	
	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
Raw Image	0.7868	0.7902	0.7432	0.7474	0.7946	0.7921	0.7684	0.7706
$\alpha = 20$	0.8349	0.8307	0.7891	0.7936	0.8472	0.8492	0.7989	0.8006
$\alpha = 10$	0.8876	0.8884	0.8257	0.8288	0.8902	0.8913	0.8447	0.8468
$\alpha = 5$	**0.9157**	**0.9122**	**0.8609**	**0.8655**	**0.9125**	**0.9134**	**0.8788**	**0.8817**

performance using the raw facial graph with that using the optimized facial key graph to highlight the improvement achieved by the optimized facial key graph. Fig. 7 represents the results conducted on the CASME II and SAMM datasets, considering three and five categories of micro-expressions to demonstrate improvements of each module in the proposed GAM-MER. Each sub-figure corresponds to a different category and dataset, labeled from (a) to (d).

3.5.2. Magnification factor

The magnified micro-expression of different factor is depicted in Fig. 9. The ablation study demonstrates that the GAM-MM-MER network with magnification factor $\alpha = 5$ achieves optimal results, as shown in Table 4. For fair comparison, we only change the magnification factor. Table 4 represents the results conducted on the CASME II and SAMM datasets, considering three and five categories of micro-expressions to demonstrate improvements when using different magnification factor. Specifically, on the CASME II dataset accuracy is improved by 12.89%, 8.08%, and 2.81% and $F_{1-\text{score}}$ is increased by 12.20%, 8.16%, and 2.38% respectively for three categories. Similarly, for five categories, accuracy is improved by 11.77%, 7.18%, and 3.52% and $F_{1-\text{score}}$ is increased by 11.81%, 7.19%, and 3.67% respectively. On the SAMM dataset, accuracy is improved by 11.79%, 6.53%, and 2.23% and $F_{1-\text{score}}$ is increased by 12.13%, 6.42%, and 2.21% respectively for three categories. Similarly, for five categories, accuracy is improved by 11.04%, 7.99%,



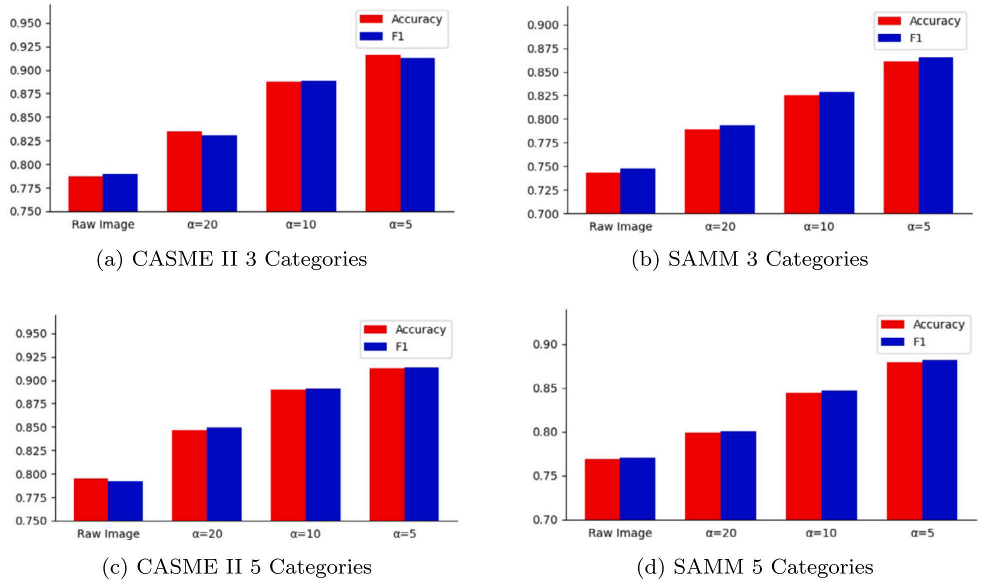
(a) CASME II 3 Categories

(b) SAMM 3 Categories

(c) CASME II 5 Categories

(d) SAMM 5 Categories

Fig. 7. Ablation study of GAM-MER architecture of accuracy and F_1 on the CASME II and SAMM datasets. Each sub-figure represents a different category and dataset, labeled from (a) to (d).



(a) CASME II 3 Categories

(b) SAMM 3 Categories

(c) CASME II 5 Categories

(d) SAMM 5 Categories

Fig. 8. Comparative analysis of different magnification factor on the CASME II and SAMM datasets. Each sub-figure represents a different category and dataset, labeled from (a) to (d).

and 3.41% and $F_{1-\text{score}}$ is increased by 11.11%, 8.11%, and 3.47% respectively. These results demonstrate that The micro expression recognition results are significantly improved compared to the original image with magnification factor $\alpha = 5$. Furthermore, $\alpha = 10$ further enhance the motion amplification effect of the micro expression videos, resulting in superior micro expression recognition results compared to $\alpha = 5$. However, when the magnification factor is further increased to $\alpha = 20$, the amplified micro expression video

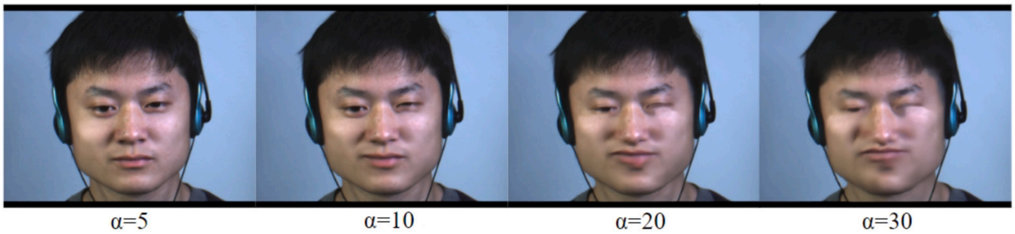


Fig. 9. Magnified micro-expression of different factor α.

Table 5
Model inference time and engineering optimization.

Model	ST-MEMM	GAM-MER
pytorch	156.4 ms	53.7 ms
ONNX	118.3 ms	43.2 ms
TensorRT	62.2 ms	25.3 ms
float 16	43.9 ms	18.5 ms

images start to exhibit a certain degree of distortion, leading to a decrease in micro expression recognition performance. Ultimately, the micro expression recognition results are even worse than those obtained with an enlargement factor $\alpha = 5$. Fig. 8 presents a detailed comparison of the accuracy for each micro-expression category while employing different magnification factor for both three and five categories on the CASME II and SAMM datasets. Each sub-figure corresponds to a different category and dataset, labeled from (a) to (d).

3.6. Time performance and engineering optimization

The Swin Transformer-based micro-expression motion magnification network (ST-MEMM)n and the graph attention mechanism-based micro-expression recognition network (GAM-MER) proposed in Sections 2.1 and 2.2 are tested for inference time on the common GPU inference card A10.

The PyTorch deep learning framework integrates numerous Python libraries for model training and optimization. However, its performance in engineering inference is relatively poor. To enhance the model's performance, this paper employs the ONNX (Open Neural Network Exchange) model format. By converting the PyTorch model to the ONNX model format, the inference performance of the model can be improved. Although converting the model to the ONNX format can improve inference performance to some extent, most video frame rates are not lower than 25 fps, and the existing inference time still cannot meet real-time requirements. To further enhance the model's performance, this paper utilizes TensorRT. TensorRT is a high-performance inference optimizer and runtime engine developed by NVIDIA, specifically optimized for deep learning inference tasks. By converting the ONNX model format to the TensorRT model format, the inference time performance of the model can be further improved to meet real-time inference demands. Furthermore, during the model training stage, this paper employs a 32-bit floating-point precision (float32) for back propagation and weight updates. However, during the model inference stage, using a 16-bit floating-point precision (float16) is sufficient to meet the inference requirements and improve the model's inference time performance. Therefore, this paper further optimizes the model by adjusting the precision of the model's weights and input videos to 16-bit floating-point numbers, thereby further enhancing the model's time performance. The inference times of various models on the A10 inference card using the PyTorch framework, converted ONNX, converted TensorRT, and 16-bit floating-point precision are shown in Table 5.

4. Conclusion

This paper proposes a graph attention mechanism-based motion magnification guided micro-expression recognition network (GAM-MM-MER). First, this paper presents a Swin Transformer-based micro-expression motion magnification network (ST-MEMM), which is the first work to utilize the Swin Transformer to magnify micro-expression motion to reduce blurring and artifacts while improving the magnification effect of micro-expressions. In addition, this paper proposes a graph attention mechanism-based micro-expression recognition network (GAM-MER), which utilizes attention mechanism to focus on key facial landmarks related to micro-expression, which improve the MER result on the public datasets.

For future research, we plan to focus on two main aspects in the field of micro-expression recognition (MER). Firstly, we aim to develop an automated method to dynamically determine the magnification factor for each micro-expression data. This approach will enhance the accuracy of MER tasks by effectively capturing the innermost relationships among facial landmarks. Additionally, we intend to address the computational efficiency of the ST-MEMM network architecture. While the GAM-MER network exhibits a relatively simple architecture, allowing it to perform real-time inference on common GPU devices such as A10, the ST-MEMM network's complexity results in longer inference times on standard machines. To overcome this limitation, we will simplify the architecture of the ST-MEMM network to enable real-time performance, facilitating practical applications of micro-expression recognition. These advancements will contribute to the ongoing progress in the field of micro-expression recognition and pave the way for more efficient and accurate analysis of micro expressions in real-world scenarios.

Additional information

The facial photos used in this article are from the CASME II micro-expression database (©Xiaolan Fu), which are representative photos of the public micro-expression dataset.

Ethics statement

This research involving facial photos is conducted in accordance with established ethical guidelines. The study complies with all relevant regulations and ethical standards regarding the use of human subjects in research.

CRediT authorship contribution statement

Falin Wu: Writing – review & editing, Resources, Project administration, Investigation, Conceptualization. **Yu Xia:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Tiangyang Hu:** Writing – review & editing, Validation, Software. **Boyi Ma:** Software, Methodology, Data curation. **Jingyao Yang:** Validation, Formal analysis, Data curation. **Haoxin Li:** Writing – review & editing, Visualization, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors would like to thank the Chinese Academy of Sciences for the CASME II micro-expression database [24], Davison et al. for the SAMM micro-expression database [25,6,26] and Li et al. for the SMIC-HS micro-expression database [27]. Data are included in the article/supplementary material/referenced in the article.

Funding statement

This research did not receive any specific grant from funding agencies in public, commercial, or no-for-profit sectors.

References

[1] J. Li, Z. Dong, S. Lu, S.J. Wang, W.J. Yan, Y. Ma, Y. Liu, C. Huang, X. Fu, 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 (3) (2023) 2782–2800, https://doi.org/10.1109/TPAMI.2022.3174895.

[2] X. Ben, Y. Ren, J. Zhang, S.J. Wang, K. Kpalma, W. Meng, Y.J. Liu, Video-based facial micro-expression analysis: a survey of datasets, features and algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 44 (9) (2022) 5826–5846, https://doi.org/10.1109/TPAMI.2021.3067464.

[3] Paul Ekman Group, Micro expression training tool and subtle expression training tool, https://www.paulekman.com/, 10 July 2003.

[4] M. Yeasin, B. Bullot, R. Sharma, From facial expression to level of interest: a spatio-temporal approach, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, CVPR 2004, vol. 2, Washington, DC, USA, 2004, pp. II–II, https://doi.org/10.1109/CVPR.2004.1315264.

[5] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928, https://doi.org/10.1109/TPAMI.2007.1110.

[6] A. Davison, W. Merghani, M. Yap, Objective classes for micro-facial expression recognition, J. Imaging 4 (10) (2018) 119, https://doi.org/10.3390/jimaging4100119.

[7] S.-T. Liong, J. See, K. Wong, R.C.W. Phan, Less is more: micro-expression recognition from video using apex frame, Signal Process. Image Commun. 62 (2018) 82–92, https://doi.org/10.1016/j.image.2017.11.006.

[8] N. van Quang, J. Chun, T. Tokuyama, Capsulenet for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 2019, pp. 1–7.

[9] D.H. Kim, W.J. Baddar, Y.M. Ro, Micro-expression recognition with expression-state constrained spatio-temporal feature representations, in: Proceedings of the 24th ACM International Conference on Multimedia, New York, NY, USA, 2016, pp. 382–386.

[10] H.-Q. Khor, J. See, S.-T. Liong, R.C. Phan, W. Lin, Dual-stream shallow networks for facial micro-expression recognition, in: 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 36–40.

[11] Y.S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, L.-K. Tan, Off-apexnet on micro-expression recognition system, Signal Process. Image Commun. 74 (2019) 129–139, https://doi.org/10.1016/j.image.2019.02.005.

[12] L. Lei, J. Li, T. Chen, S. Li, A novel graph-tcn with a graph structured representation for micro-expression recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, New York, NY, USA, 2020, pp. 2237–2245.

[13] H.-X. Xie, L. Lo, H.-H. Shuai, W.-H. Cheng, Au-assisted graph attention convolutional network for micro-expression recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, New York, NY, USA, 2020, pp. 2871–2880.

[14] A.J.R. Kumar, B. Bhanu, Three stream graph attention network using dynamic patch selection for the classification of micro-expressions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022, pp. 2476–2485.

[15] X.-B. Nguyen, C.N. Duong, X. Li, S. Gauch, H.-S. Seo, K. Luu, Micron-BERT: BERT-based facial micro-expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 2023, pp. 1482–1492.

[16] H. Pan, L. Xie, Z. Wang, C3DBed: facial micro-expression recognition with three-dimensional convolutional neural network embedding in transformer model, Eng. Appl. Artif. Intell. 123 (2023) 106258, https://doi.org/10.1016/j.engappai.2023.106258.

[17] X. Fan, X. Chen, M. Jiang, A.R. Shahid, H. Yan, SelfME: self-supervised motion learning for micro-expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023, pp. 13834–13843.

[18] L. Zhang, X. Hong, O. Arandjelović, G. Zhao, Short and long range relation based spatio-temporal transformer for micro-expression recognition, IEEE Trans. Affect. Comput. 13 (4) (2022) 1973–1985, https://doi.org/10.1109/TAFFC.2022.3213509.

[19] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM Trans. Graph. 31 (4) (2012) 1–8, https://doi.org/10.1145/2185520.2185561.

[20] C. Liu, A. Torralba, W.T. Freeman, F. Durand, E.H. Adelson, Motion magnification, ACM Trans. Graph. 24 (3) (2005) 519–526, https://doi.org/10.1145/1073204.1073223.

[21] T.-H. Oh, R. Jaroensri, C. Kim, M. Elgharib, F. Durand, W.T. Freeman, W. Matusik, Learning-based video motion magnification, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, Cham, 2018, pp. 633–648.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755.

[23] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338, https://doi.org/10.1007/s11263-009-0275-4.

[24] W.J. Yan, X. Li, S.J. Wang, G. Zhao, Y.J. Liu, Y.H. Chen, X. Fu, CASME II: an improved spontaneous micro-expression database and the baseline evaluation, PLoS ONE 9 (1) (2014) e86041, https://doi.org/10.1371/journal.pone.0086041.

[25] A.K. Davison, C. Lansley, N. Costen, K. Tan, M.H. Yap, Samm: a spontaneous micro-facial movement dataset, IEEE Trans. Affect. Comput. 9 (1) (2016) 116–129, https://doi.org/10.1109/TAFFC.2016.2573832.

[26] C.H. Yap, C. Kendrick, M.H. Yap, Samm long videos: a spontaneous facial micro- and macro-expressions dataset, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 2020, pp. 771–776.

[27] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: inducement, collection and baseline, in: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–6.

[28] T. Pfister, X. Li, G. Zhao, M. Pietikäinen, Recognising spontaneous facial micro-expressions, in: 2011 International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 1449–1456.

[29] Y. Wang, J. See, R.C.-W. Phan, Y.-H. Oh, Lbp with six intersection points: reducing redundant information in lbp-top for micro-expression recognition, in: D. Cremers, I. Reid, H. Saito, M.-H. Yang (Eds.), Computer Vision – ACCV 2014, Springer International Publishing, Cham, 2015, pp. 525–537.

[30] Y. Zong, X. Huang, W. Zheng, Z. Cui, G. Zhao, Learning from hierarchical spatiotemporal descriptors for micro-expression recognition, IEEE Trans. Multimed. 20 (11) (2018) 3160–3172, https://doi.org/10.1109/tmm.2018.2820321.