Multi-Information Hierarchical Fusion Transformer with Local Alignment and Global Correlation for Micro-Expression Recognition

Jinsheng Wei*
Nanjing University of Posts and Telecommunications
Nanjing, Jiangsu, China
weijs@njupt.edu.cn

Jialiang Sun
Nanjing University of Posts and Telecommunications
Nanjing, Jiangsu, China
1024010316@njupt.edu.cn

Guanming Lu
Nanjing University of Posts and Telecommunications
Nanjing, Jiangsu, China
lugm@njupt.edu.cn

Jingjie Yan
Nanjing University of Posts and Telecommunications
Nanjing, Jiangsu, China
yanjingjie@njupt.edu.cn

Dong Zhang
State Grid Huaian Power Supply Company
Huaian, Jiangsu, China
2311799536@qq.com

Abstract

Learning discriminative micro-expression (ME) features from low-intensity facial movements is a key challenge for micro-expression recognition (MER). Although existing research has demonstrated that the appearance, motion and geometric information are distinguishing for MEs, the effectiveness of merging these information is still unclear. Thus, this paper proposes a Multi-information Hierarchical Fusion Transformer (MiHF-Tr) model to fully and effectively aggregate the facial appearance, motion, and geometric information of MEs, exploring a more reasonable way of multi-information fusion. As different information is homology, MiHF-Tr introduces a local and global hierarchy fusion framework to fuse them by modeling their local and global semantic consistency. Considering the bias of different information in feature representation ability, a single-core self-attention is proposed to achieve local multi-information fusion, which focuses on strong information and supplements it with weak information. The experimental results demonstrate that the fusion of appearance, motion, and geometric features is discriminative, and the proposed method can effectively aggregate multiple information, achieving competitive performance.

CCS Concepts

• **Computing methodologies → Image representations**.

Keywords

Multi-Information; Feature Fusion; Micro-Expression Recognition; Semantic Consistency

ACM Reference Format:
Jinsheng Wei, Jialiang Sun, Guanming Lu, Jingjie Yan, and Dong Zhang. 2025. Multi-Information Hierarchical Fusion Transformer with Local Alignment and Global Correlation for Micro-Expression Recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755860

1 Introduction

Micro-Expression Recognition (MER), as an important topic at the intersection of psychology and computer vision, has received widespread attention lately [41], given its value in clinical diagnosis, national security, emotional computing [7, 19, 20], and interrelated research [4, 5, 23]. Micro-Expressions (ME) are short-duration (usually 1/25 to 1/5 second) with weak intensity facial muscle movements. ME typically reflects the true emotions an individual is trying to hide. However, its instantaneous and low intensity traits pose a serious challenge to accurate recognition [18].

The instantaneous problem can be overcome by utilizing high-speed cameras to capture fragments of MEs. However, low-intensity problems are more challenging for using machine learning techniques to recognize MEs. Although action amplification technology alleviates this problem, the image distortion caused by amplification restricts the intensity of amplification. As shown in Figure 1, ME video contains appearance, motion, and geometric information, which have been proven effective in distinguishing micro-expressions [10, 17, 39, 40], especially motion information [1, 36]. Fully mining and fusing the information contributes to learn more discriminative ME features, thereby improving recognition accuracy. However, existing works [2, 11, 14, 46] simply joint the partial information, and do not explore effective ways to fuse them. To solve this problem, this work focuses on the full mining and fusion of facial multi-information from ME videos.

Different types of information data are heterogeneous, namely, their data structure differ. As shown in Figure 1, the facial landmarks-based geometric information data is non-Euclidean, while the optical flow-based motion information data is Euclidean. To overcome the differences in data structures for different information, this work

*Jinsheng Wei is the corresponding author (Emial: weijs@njupt.edu.cn)

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755860

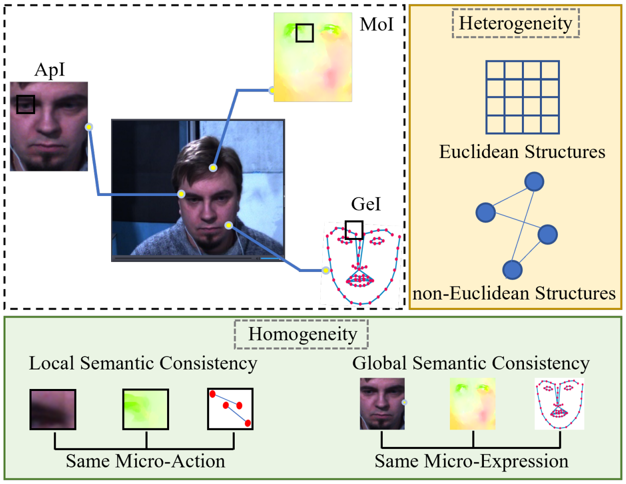


Figure 1: The homogeneity and heterogeneity of multi-information data. ApI, MoI and GeI express appearance, motion, and geometric information, respectively.

utilizes different models to extract different information features, respectively, and align them in local regions.

Different information is also homologous. Micro-actions between different information in local regions are the same, so multi-information has local semantic consistency; As different information is mined from the same facial ME (with consistent categories), multi-information has global semantic consistency. Most existing methods use multi-stream models [14, 16, 18, 35] to aggregate different information. However, these works focus on global semantic consistency of multi-information, while neglecting local semantic consistency. Therefore, this work adopts a hierarchical fusion framework that combines local and global semantic consistency. This framework aligns and aggregates multi-information features at the local region and establishes the global correlation of multi-information features between different local regions.

Due to homogeneity, the correlation between different information is strong. So far, a limited number of works [2, 35, 38] have explored the interaction between different information and improved the fusion effect. However, it is worth noting that these works ignore the differences and complementarity between strong and weak information. Here, strong information and weak information are distinguished by their contributions to classification tasks. Thus, this work focuses on strong information and supplements it with weak information to enhance fusion performance.

Overall, the main contributions of this paper are as follows:

1) This work studies and explores a reasonable way to fuse multiple information from facial ME video, making the first step to aggregate the three types of information: appearance, motion and geometry. Experimental results demonstrate that the proposed fusion method is effective, and the fusion of three types of information is superior to that of any two types of information.

2) To comprehensively aggregate multi-information, a novel Multi-information Hierarchical Fusion Transformer (MiHF-Tr) model is proposed, including Local Alignment Multi-Information Fusion (LAMiF) and Global Correlation Multi-Information Fusion (GCMiF) module. This model introduces a local and global hierarchy framework that joints local and global semantic consistency by local alignment fusion and global correlation fusion, corresponding to LAMiF and GCMiF. The experimental results show that the proposed model can fully aggregate multi-information features.

3) To effectively aggregate multiple information with strong correlation, in the LAMiF module, a Single-Core Self-Attention (SC-SA) is designed to establish a mechanism that combines dominant patterns of strong information with supplementary patterns of weak information. Experimental results demonstrate that SC-SA is an effective and reasonable method to aggregate different information.

2 Related Work

Based on research issues, related works are introduced from two aspects: information fusion and Transformer-based models.

2.1 Information Fusion

ME video clips contain rich appearance, motion, and geometric information that can represent facial micro-action features. So far, many research works [3, 10, 21, 25, 43] have calculated optical flow from video frames and designed feature extraction algorithms to extract motion features from the optical flow. These works have demonstrated that optical flow-based motion information is effective for MER tasks. Li et al. [17] explored the advantages of the apex frame. The experimental conclusion indicates that apex frame-based appearance information contains discriminative information for MER. Recently, to explore more compact micro-action representation methods, Wei et al. [40] designed a new graph convolutional network to model facial landmarks and achieved competitive results. They explored the effectiveness of geometric information based on facial landmarks, and the conclusion was positive [37]. Thus, so far, a large number of works have attempted to extract ME features from three types of information to represent facial micro-actions. However, a single type of information makes it difficult for the model to fully learn discriminative ME features.

Some researchers aggregated different information to enhance the representational ability of features. As a typical case, the dual-stream model [11, 35] is constructed to extract motion features and appearance features from optical flow and RGB frames, respectively. Kumar et al. [14] constructed a dual-stream graph network model to extract motion and geometric features from optical flow and facial landmarks, respectively. The above works demonstrated the advantages of information fusion. However, only two types of information were considered to represent ME. Most notably, these works did not explore the fusion mechanism, only using a simple fusion way, such as concatenation and addition, based on a dual-branch structure. Different from these methods, the proposed method aggregates three types of information to more comprehensively represent facial micro-actions, and presents a multi-information local and global hierarchy fusion framework.

2.2 Transformer-based Model

Recently, Transformer has achieved good performance in correlation modeling. Zhu et al. [48] designed a sparse-based Transformer to extract sparse features from optical flow. Zhang et al. [47] proposed a novel spatio-temporal transformer to enhance long-range

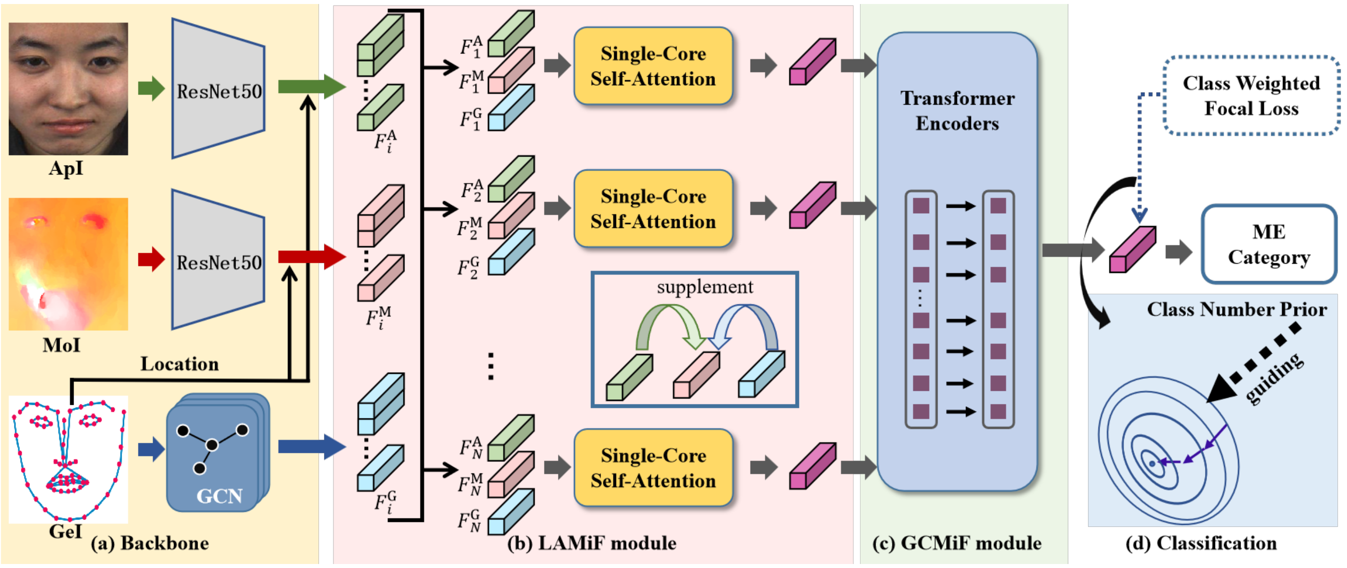


Figure 2: Flowchart of the proposed method. (a) Backbone: extracting the multi-information feature; (b) LAMiF module: alignment fusion of local multi-information; (c) GCMiF module: correlation fusion of global multi-information; (d) classification: CW Focal loss and softmax classifier.

spatial features. Zhai et al. [46] divided facial images into multiple sub-blocks and introduced a multi-head self-attention mechanism to fuse these sub-blocks locally and globally. By integrating overall and detailed information, Ma et al. [27] designed a dual-branch classification network. The two branches are responsible for capturing overall motion and detailed motion, respectively, and Swin Transformer [26] is adapted to focus on the region of interest. Different from these methods, on one hand, in terms of the architecture and purpose, this paper focuses on multi-information fusion and designs a novel MiHF-Tr model that achieves alignment fusion of local multi-information and correlation fusion of global multi-information; On the other hand, in terms of self-attention, a new SC-SA is proposed for multi-information feature fusion, which enhances strong information with weak information.

3 Method

Figure 2 presents the framework of the proposed MiHF-Tr. First, the multi-information backbone extracts multi-information feature using ResNet50 [12] and Graph Convolutional Network (GCN) [30, 42]; second, LAMiF module aligns local multi-information and fuses them using SC-SA; next, GCMiF module achieves the global correlation fusion of multi-information; finally, the classification part adopts Class Weighted Focal Loss (CW-FLoss) to constrain the model for solving sample imbalance, and employs softmax as a classifier.

3.1 Multi-Information Backbone

In this paper, the apex frame, optical flow and facial landmarks are extracted as the appearance, motion and geometric information. The apex and onset frames are used to calculate optical flow, and the facial landmarks are detected in the apex and onset frames.

As we know, the apex frame and optical flow belong to Euclidean data, while facial landmarks belong to non-Euclidean data [40]. ResNet50 and GCN can process Euclidean and non-Euclidean data, respectively. Thus, ResNet50 is employed to extract the appearance and motion feature maps, while GCN is employed to extract the geometric features. To facilitate local alignment of the three features, the feature map preserves the spatial information. Specifically, the apex frame (ApF) and optical flow (OF) are processed by the first l layers of ResNet50 to obtain the appearance feature map (AF) and motion feature map (MF):

$$\begin{aligned} \mathrm{AF} &= \mathrm{ResNet50}^{l}(\mathrm{ApF}) \\ \mathrm{MF} &= \mathrm{ResNet50}^{l}(\mathrm{OF}) \end{aligned} \tag{1}$$

where $\mathrm{ResNet50}^{l}(\cdot)$ expresses the output feature map of the l-th layer of ResNet50, i ranges from 1 to 50; The sizes of ApF and OF are 224×224, that compatible with ResNet50; The sizes of AF and MF are $h_0 \times h_0$, and h_0 can be divided by 224.

For geometric information, 51 facial landmarks are selected from 68 facial landmarks, and the landmarks of facial contours unrelated to ME are discarded. Then, facial landmarks (FL) are processed to get geometric features (GF):

$$\mathrm{GF} = \mathrm{GCN}(\mathrm{FL}) \tag{2}$$

where $\mathrm{GF} \in \mathbb{R}^{51\times d}$, d is the feature dimension of output node; $\mathrm{FL} \in \mathbb{R}^{51\times 4}$, where 4 is the input node feature dimension, representing the coordinate values of apex and onset frames.

3.2 Local Alignment Multi-Information Fusion Module

This module includes two parts: local multi-information alignment and SC-SA. The details are as follows:

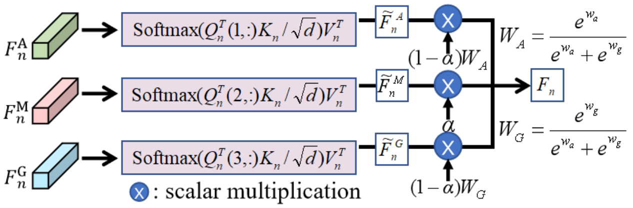

F_n^A
F_n^M
F_n^G
Softmax($Q_n^T(1,:)K_n/\sqrt{d})V_n^T$
Softmax($Q_n^T(2,:)K_n/\sqrt{d})V_n^T$
Softmax($Q_n^T(3,:)K_n/\sqrt{d})V_n^T$
$\tilde{F}_n^A$
$\tilde{F}_n^M$
$\tilde{F}_n^G$
$(1-\alpha)W_A$
α
$(1-\alpha)W_G$
$W_A = \frac{e^{w_a}}{e^{w_a}+e^{w_g}}$
$W_G = \frac{e^{w_g}}{e^{w_a}+e^{w_g}}$
F_n
: scalar multiplication

Figure 3: Illustration of SC-SA.

3.2.1 Local Multi-Information Alignment. Since all information is mined from facial videos, they have homology. The multi-information features within the same local area represent the same micro-actions, that is, different information has local semantic consistency. Therefore, the proposed method utilizes the facial spatial position information provided by facial landmarks to align local features of different information, facilitating subsequent learning of fusion features with consistent semantics.

Assuming the n-th facial landmark of apex frame is $\mathrm{P}_n(x_n, y_n)$, where n is from 1 to 51. Then, the corresponding point in the AF and MF is:

$$\mathrm{PF}_n = (x_n \times \frac{h_0}{224}, y_n \times \frac{h_0}{224}) \tag{3}$$

In AF, we clip an image block of $B \times B$ centered on point PF_n and flatten it to obtain the n-th local appearance feature $F_n^A \in \mathbb{R}^{1\times d}$, where d is feature dimension; The same method can be used to obtain the n-th local motion feature $F_n^M \in \mathbb{R}^{1\times d}$. The n-th node feature of geometric feature GF is the local geometric feature $F_n^G \in \mathbb{R}^{1\times d}$. As a result, for the n-th facial landmark, we can obtain a multi-information local feature set: $F_n^{'} = (F_n^A; F_n^M; F_n^G) \in \mathbb{R}^{3\times d}$.

3.2.2 Single-Core Self-Attention (SC-SA). Based on previous works [1, 14] and the pre-experiments, it turns out that compared to the other two information, optical flow-based motion information has strong discriminability. This is mainly because MEs are a dynamic process, while motion information can more directly represent facial micro-actions, which is consistent with the dynamic nature of ME. However, the apex frame lacks time-domain information, and geometric information only represents the geometric changes around facial landmarks. Therefore, we define optical flow as dominant information and the other two as weak information.

If strong and weak information are handled fairly, the latter may interfere with strong information. Thus, the proposed method focuses on motion information as the single core, with other information as auxiliary, that is, using motion features as the dominant pattern and the other two features as supplementary patterns.

Self-Attention (SA) [26] has powerful capabilities in various tasks [6, 45]. It effectively aggregates different features by establishing dependency relationships between them. Therefore, this paper introduces the above ideas into SA and proposes a SC-SA that effectively fuses appearance, motion, and geometric local features. Specifically, for the n-th local region, multi-information local features F_n^A, F_n^M and F_n^G in $F_n^{'}$ are fused to obtain a multi-information local fusion feature F_n. First, Query Q_n, keys K_n, and values V_n are calculated by $Q_n = W_Q F_n^{'T}$, $K_n = W_K F_n^{'T}$ and $V_n = W_V F_n^{'T}$, respectively, and both $\in \mathbb{R}^{d\times 3}$. W_Q, W_K and W_V are the learable weights and both $\in \mathbb{R}^{d\times d}$; the superscript T represents transposition.

Then, a multi-information fusion feature F_n is calculated by:

$$F_n = \alpha \tilde{F}_n^M + (1-\alpha)(w_a \tilde{F}_n^A + w_g \tilde{F}_n^G) \in \mathbb{R}^{1\times d} \tag{4}$$

where w_a and w_g are learnable values to weight $\tilde{F}_n^A$ and $\tilde{F}_n^G$, respectively; and $\tilde{F}_n^A$, $\tilde{F}_n^M$ and $\tilde{F}_n^G$ are obtained by:

$$\begin{aligned} \tilde{F}_n^A &= \mathrm{Softmax}(\frac{Q_n^T(1,:)K_n}{\sqrt{d}})V_n^T \in \mathbb{R}^{1\times d} \\ \tilde{F}_n^M &= \mathrm{Softmax}(\frac{Q_n^T(2,:)K_n}{\sqrt{d}})V_n^T \in \mathbb{R}^{1\times d} \\ \tilde{F}_n^G &= \mathrm{Softmax}(\frac{Q_n^T(3,:)K_n}{\sqrt{d}})V_n^T \in \mathbb{R}^{1\times d} \end{aligned} \tag{5}$$

Furthermore, in formula 4, w_a and w_g have no constraints, and negative and zero values may occur. Thus, we take probability form as the weights as follows:

$$F_n = \alpha \tilde{F}_n^M + (1-\alpha)(\frac{e^{w_a}}{e^{w_a}+e^{w_g}} \tilde{F}_n^A + c\tilde{F}_n^G) \tag{6}$$

So, the coefficients of $\tilde{F}_n^A$ and $\tilde{F}_n^G$ are limited between 0 and $1-\alpha$. Also, to ensure that $\tilde{F}_n^M$ is the dominant pattern, $\alpha \geq 0.5$. For every local region, the multi-information local features are fused, resulting in 51 multi-information local fusion features $F_n, n = 1, 2, ..., 51$.

3.3 Global Correlation Multi-Information Fusion Module

Although different information has different data distributions, they all point to the same ME category and have global semantic consistency. In fact, ME categories are closely related to different local micro actions. Based on the local semantic consistency features learned by LAMiF, the proposed method achieves the mapping of local semantic consistency features to global semantic consistency features through correlation modeling.

So far, Transformer is a mainstream model that establishes correlations between different token features. Therefore, this paper designs a GCMiF module, which employs Transformer layers to establish the correlation between local features, learning discriminative ME features, that is:

$$F_{ME} = MP(Transformer(F_1, F_2, ..., F_{51})) \tag{7}$$

where MP expresses the mean pooling.

3.4 Loss and Classifier

Inspired by Focal Loss [2], in this paper, a class-weighted focal loss (CW-FLoss) is designed to mitigate the adverse effects of sample imbalance on multi-information fusion. CW-FLoss uses prior knowledge of the sample number to guide the optimization process of the MiHF-Tr model.

$$CW\ FLoss = \sum_{c=0}^{C} N_c(1-p_c)^{\gamma} \log(p_c) \tag{8}$$

Table 1: Label details of CASME II and SAMM datasets. Po: positive; Ne: negative; Su: surprise; Ha: happiness; Di: disgust; Re: repression; An: anger; Co: contempt.

Dataset	Label Distribution			
CASME II	Label	Po	Ne	Su
	Original Label	Ha	Di&Re	Su
	Number	32	90	28
SAMM	Label	Po	Ne	Su
	Original Label	Ha	An&Di&Co	Su
	Number	26	78	15

where C is the total number of ME classes; N_c is the normalization sample number of c-th class ; p_c is the probability of predicting as c-th class; γ is the focus factor. In addition, the classifier employs a softmax function following the mainstream.

4 Experiment

The experimental results are reported in this section. The experiments evaluate the performance of the proposed components and method, and verify the rationality of the proposed multi-information fusion way. First, the effectiveness of jointing appearance, motion, and geometric information is evaluated; Second, we carry out the ablation analysis to evaluate the proposed components, including SC-SA, LAMiF, GCMiF and CW-FLoss; Finally, we compare the proposed MiHF-Tr with state-of-the-art (SOTA) methods.

4.1 Dataset and Evaluation Metric

4.1.1 Datasets. Following the existing works [2, 40], CASME II and SAMM are adopted to evaluate the performance of the proposed method. CASME II contains 255 samples with 26 subjects. All participants are of the same ethnicity. MEs are annotated into seven categories. ME videos were collected by high-speed cameras with 200 fps, and the resolution is 640*480; SAMM contains 159 samples with 32 subjects. All participants are from 13 ethnicities. MEs are annotated into eight categories, and the high-speed camera with 200 fps and 2040*1088 resolution.

Consistent with previous works [2, 9, 28], MEs are classified into three categories, which requires adjusting the original samples and labels. The corresponding relationship before and after adjustment is shown in Table 1.

4.1.2 Evaluation Metrics. All experiments were conducted under Leave-One-Subject-Out (LOSO) cross-validation. Namely, each subject's sample takes turns serving as the test set one time, while the rest samples serve as the training set. Following the recent mainstream works[2, 9, 44], the results of all subjects are accumulated to calculate unweighted F1-Score (UF1) and unweighted average recall (UAR) as the evaluation metrics.

UF1 and UAR are commonly used to evaluate the model performance of multi-classification problems, especially when dealing with imbalanced samples. The calculation formula is:

$$\begin{aligned} \mathrm{UAR} &= \frac{1}{\mathrm{C}} \sum_{c=1}^{\mathrm{C}} \frac{\mathrm{TP}_c}{n_c} \\ \mathrm{UF1} &= \frac{1}{\mathrm{C}} \sum_{i=0}^{\mathrm{C}} \frac{2 \times \mathrm{TP}_c}{2 \times \mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \end{aligned} \tag{9}$$

where C is the total number of ME classes; TP_c, FP_c and FN_c are the true positive, the false positive and the false negative, respectively.

4.2 Implementation Setting

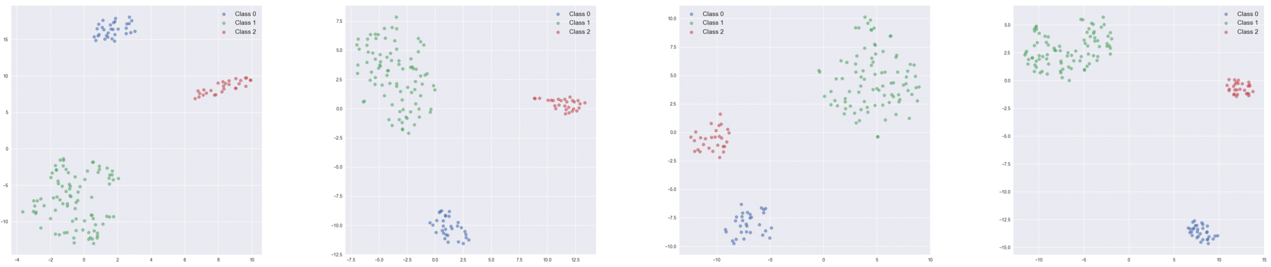
As in existing works, e.g. [15, 24], the onset and apex frames were obtained using database labels, and their detection belongs to another task in ME analysis [33]. Facial micro-actions are magnified using learning-based amplification techniques [29], with a magnification factor of 3, following the works [15, 24]. The Dlib [13] package is employed to detect facial landmarks. Furthermore, the block size $B \times B$ is set to 4×4; α is set to 0.7; l is set to 11, and the corresponding h_0 is 56; d is 256; GCN model extracting geometric features includes two layers.

For the training stage, the model is optimized by an Adam optimizer with an initial learning rate of 0.001, and the learning rate is divided by 10 every 20 epochs. The epoch number and batch size are set to 100 and 64, respectively. All models are trained on a single GTX 1080 GPU (8G) with Pytorch 1.8.1 version.

4.3 The Study on the Effectiveness of Multi-Information Fusion

This paper explores the effectiveness of appearance (Ap), motion (Mo), and geometry (Ge) information fusion. To assess the contribution of each information to the overall performance, we conducted ablation experiments where only two of the three types of information were utilized, as shown in Table 2. According to this table, the proposed method achieves 0.9350 UF1 and 0.9354 UAR on CASME II, and 0.8199 UF1 and 0.7916 UAR on SAMM. It turns out that the fusion of the three information is superior to the fusion of any two information, which demonstrates that aggregating Ap, Mo and Ge information can improve recognition performance. Interestingly, we find that the performance of Ap + Ge information obviously lower than that of other combinations. The main reason is the lack of motion information. It demonstrates that compared with the other two information, Mo information is more discriminative.

By visualizing features using t-SNE [32], we further analyze the effectiveness of Ap, Mo and Ge information fusion. Figure 4 (a), (b) (c) illustrate the visualization of the features aggregating two information. Figure 4(d) illustrates the visualization of the features aggregating three information. It is evident that jointing three information leads to the most distinct and compact clustering of different classes, especially category 1 (green). Specifically, the intra-class compactness and inter-class separability are significantly enhanced compared to the combinations of any two information. The above phenomenon demonstrates that the complementary information provided by Mo, Ap, and Ge features together enables the model to learn more discriminative representations. Furthermore, compared to Ge + Mo and Mo + Ap, the clustering effect of Ge + Ap is the worst, which is consistent with the previous quantitative analysis



(a) Motion & Appearance (b) Motion & Geometry (c) Appearance & Geometry (d) Motion & Appearance & Geometry

Figure 4: Visualization of the features in the layers using t-SNE[32], with different information combination.

Table 2: The comparison of fusing different information.

Information Type			CASME II		SAMM	
Ap	Mo	Ge	UF1	UAR	UF1	UAR
×	✓	✓	0.8915	0.8848	0.7816	0.7471
✓	×	✓	0.7074	0.6861	0.5326	0.5269
✓	✓	×	0.8986	0.8899	0.7489	0.7249
✓	✓	✓	**0.9350**	**0.9354**	**0.8199**	**0.7916**

about Mo discriminability. Also, it can be seen that combinations containing motion information have better clustering effects on class 0 (blue) and class 2 (red).

The average intra-class and inter-class distances (ARa and AEr) are calculated. For the features in Figure 4 (a), (b), (c) and (d), ARas are 0.12, 0.3, 0.32 and 0.1, respectively; AErs are 2.34, 2.62, 2.26 and 2.76, respectively. It demonstrates that the fusion features of Mo, Ap and Ge have the minimum ARa and maximum AEr, which is consistent with the results in Figure 4.

Overall, through quantitative and visual analysis, the following conclusions can be drawn: (1) the fusion of appearance, motion, and geometric information is effective and superior to the fusion of any two types of information; (1) Compared to the other two types of information, motion information is more discriminative.

4.4 The Ablation Analysis

Ablation experiments were conducted on LAMiF, GCMiF, SC-SA and CW-FLoss, and the results and analysis are presented as follows:

4.4.1 The Evaluation on LAMiF and GCMiF. Table 3 shows the ablation result on LAMiF and GCMiF. It can be found that removing either module has a negative impact on performance.

The proposed method (LAMiF + GCMiF) improves UF1 by 0.0205 and 0.0100, compared with LAMiF, on CASME II and SAMM, respectively. Namely, GCMiF can improve performance. It demonstrates that, based on local multi-information fusion, global correlation fusion of multi-information features can effectively learn ME features with global semantic consistency. Also, the removal of LAMiF causes a negative impact that decreases UF1 by 0.0634 and 0.0732, on CASME II and SAMM, respectively. It turns out that directly performing global fusion of multi-information cannot fully aggregate multi-information, while aligning and fusing multi-information features from a local perspective can uncover local micro-motion consistency. Furthermore, compared with GCMiF, LAMiF achieves a more competitive UF1 and UAR on both datasets, which shows that local fusion is more crucial than global fusion. The reasons may be that only global fusion easily ignores the consistency of some local micro-actions.

Table 3: The ablation study on LAMiF and GCMiF.

Module		CASME II		SAMM	
LAMiF	GCMiF	UF1	UAR	UF1	UAR
✓	×	0.9145	0.9026	0.8099	0.7781
×	✓	0.8716	0.8968	0.7467	0.7371
✓	✓	**0.9350**	**0.9354**	**0.8199**	**0.7916**

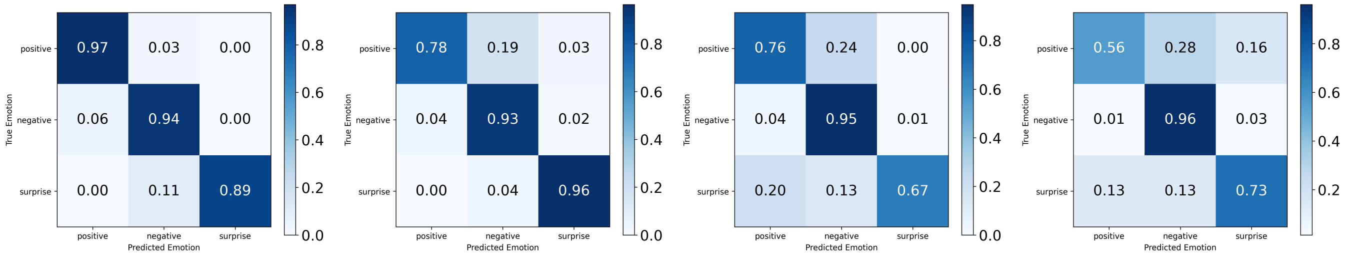
Table 4: The ablation study on LAMiF and GCMiF, under 4 ME categories.

Module		CASME II		SAMM	
LAMiF	GCMiF	UF1	UAR	UF1	UAR
✓	×	0.8677	0.8517	0.7656	0.7409
×	✓	0.8503	0.8720	0.7587	0.7433
✓	✓	**0.8923**	**0.9092**	**0.7866**	**0.7650**

We further divided the negative category into two categories: disgust and repression on CASME II; disgust and anger on SAMM. As a result, MEs are divided into fine-grained 4 categories. As shown in Table 4, the conclusion of the ablation study is also consistent for LAMiF and GCMiF.

Furthermore, we employ the wilcoxon signed-rank test to validate statistical significance. The results on CASME II demonstrate that the p-values for MiHF-Tr over LAMIF (p = 0.048) and MiHF-Tr over GCMIF (p = 0.0146) are both below the significance threshold of 0.05, confirming that the improvements are statistically significant.

Overall, the proposed hierarchical fusion framework from local alignment to global correlation can fully aggregate appearance, motion and geometric information, and the designed LAMiF and GCMiF play a positive role.



(a) CW Focal Loss on CASME II (b) UCW Focal Loss on CASME II (c) CW Focal Loss on SAMM (d) UCW Focal Loss on SAMM

Figure 5: Confusion matrix on CASME II and SAMM, with UCW-FLoss or CW-FLoss.

Table 5: The ablation study on local fusion (SC-SA) and global fusion (SA). Pl means pooling; SA means self-attention.

Type	Method	CASME II		SAMM	
		UF1	UAR	UF1	UAR
Local	SA+MeanPl	0.8531	0.8760	0.7362	0.7351
	SA+MaxPl	0.8961	0.8764	0.7282	0.7363
	Concat+FC	0.8473	0.8512	0.7156	0.7183
	SC-SA	**0.9350**	**0.9354**	**0.8199**	**0.7916**
Global	MeanPl	0.9145	0.9026	0.8099	0.7781
	MaxPl	0.9037	0.9254	0.7922	0.7678
	LSTM	0.8911	0.8913	0.7769	0.774
	SA	**0.9350**	**0.9354**	**0.8199**	**0.7916**

4.4.2 The Evaluation on Local Fusion (SC-SA). To verify the effectiveness of SC-SA in fusing local multi-information, it was compared with other fusion methods. As shown in Table 5, SC-SA achieves the best performance. In fact, SC-SA set a single strong information as the dominant pattern and other weak information as the auxiliary pattern, which can achieve effective information complementarity. It demonstrates that SC-SA avoids the negative interference of weak information on dominant information, improving the local fusion performance of multi-information.

Combining Table 3 and Table 5, an interesting conclusion can be drawn that using SA + mean pooling for local fusion performs worse than not performing local fusion. This may be because meanpooling places different information at the same level, which ignores their differences in feature discriminability, restricting the feature representation of dominant information.

4.4.3 The Evaluation on Global Fusion (SA). Table 5 shows the results of different global fusion methods. On the one hand, pooling is difficult to learn the global correlation between different local fusion features; On the other hand, LSTM is not as good as SA in terms of related modeling. Compared to meanpooling, maxpooling and LSTM, self-attention can more effectively fuse 51 local fusion features, achieving high-performance global fusion.

4.4.4 The Evaluation on CW-FLoss. Table 6 shows the performance comparison of Focal Loss before and after class weighting. CW-FLoss improves UF1 by 0.0420 and 0.0602, on CASME II and SAMM, respectively. It turns out that class weights can improve the optimization process to enhance feature discriminability.

Table 6: The ablation study on CW-FLoss.

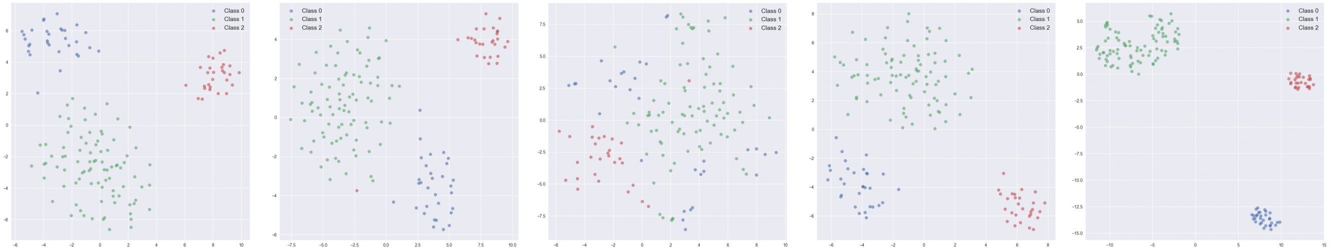
Loss	CASME II		SAMM	
	UF1	UAR	UF1	UAR
UCW-FLoss	0.8930	0.8795	0.7597	0.7515
CW-FLoss	**0.9350**	**0.9354**	**0.8199**	**0.7916**

Table 7: The parameter evaluation on α.

α	CASME II		SAMM	
	UF1	UAR	UF1	UAR
0.1	0.9012	0.8869	0.7764	0.7650
0.3	0.9031	0.8975	0.7820	0.7334
0.5	0.9115	0.8892	0.7705	0.7677
0.7	**0.935**	**0.9354**	**0.8199**	**0.7916**
0.9	0.9197	0.9197	0.8160	0.7804

To further illustrate the advantages of CW-FLoss in dealing with sample bias, we provide the corresponding confusion matrix in Figure 5. According to Table 1, the sample sizes of positive and surprise ME are relatively small. UCW-FLoss (Focal Loss without Class Weighted) performs poorly in recognizing the positive category, and there is a significant difference in performance for recognizing positive and surprise categories. After using CW-FLoss, the performance of recognizing the positive category was significantly improved (0.19 and 0.20 on CASME II and SAMM, respectively). Although the recognition of the surprise category has slightly deteriorated (0.07 and 0.06 on CASME II and SAMM, respectively), the whole performance has improved, and the performance of recognizing positive and surprise categories is more balanced. Overall, CW-FLoss can drive the model to focus on ME classes with fewer samples, improving recognition performance.

4.4.5 The Evaluation on α. From Table 7, which evaluates α from 0.1 to 0.9 with a 0.2 interval, 0.7 is optimal, namely, 0.7 strong features with 0.3 weak features achieve optimal performance gains. As α goes down, the performance goes down obviously; As it goes up, the decrease is slight, which indicates the advantage of motion features. Also, the too-small ratio for the other two causes excessive information loss.



(a) Motion Feature (b) Appearance Feature (c) Geometry Feature (d) LAMiF Output Feature (e) GCMiF Output Feature

Figure 6: Feature visualizations of different layers in the proposed MiHF-Tr model.

Table 8: Comparing with the existing methods.

Method	Year	CASME II		SAMM	
		UF1	UAR	UF1	UAR
LBP-TOP [34]	2015	0.7026	0.7429	0.3954	0.4102
Bi-WOOF [22]	2018	0.7805	0.8026	0.5211	0.5139
MAP-RME [31]	2022	0.8270	-	0.7580	-
SLSTT-LSTM [47]	2022	0.9010	0.8850	0.7150	0.6430
SelfME [8]	2023	0.9078	0.9290	-	-
Micron-Bert [28]	2023	0.9034	0.8914	-	-
3CCWGANAM [9]	2024	0.7230	0.7550	0.7010	0.7480
TFT [35]	2024	0.9070	0.9090	0.7090	0.6560
MFDAN [2]	2024	0.9134	0.9326	0.7871	**0.8196**
CoTDPN [44]	2025	0.7931	0.8015	0.7539	0.7367
MiHF-Tr(Ours)	2025	**0.9350**	**0.9354**	**0.8199**	0.7916

4.5 Visualization

As shown in Figure 6, the feature visualizations of different layers in MiHF-Tr are displayed. Figure 6 (a), (b) and (c) show the feature visualizations of the Mo, Ap and Ge features, respectively. Mo features extracted from optical flow exhibit clearer class separability compared to Ap and Ge features. It can be concluded that motion features show better class separability than Ap and Ge features, validating the analysis in Section 4.4.

Figure 6 (d) shows the results of LAMiF output, and it turns out that the inter-class distance begins to increase, while the intra-class distance also begins to decrease. Figure 6 (e) shows the results of the proposed MiHF-Tr that jointing LAMiF and GCMiF. We can find that the features become more discriminative, and the ME features of different categories have clearer boundaries and lower intra-class distance. It demonstrates that the proposed MiHF-Tr can effectively fuse multi-information to learn discriminative ME features.

4.6 Comparing with Existing Methods

As shown in Table 8, compared to traditional artificial feature methods and traditional deep learning methods, our method has significant advantages; SelfME, Micron-Bert and CoTDPN are transformer-based methods, and compared to these methods, we still maintain a slight performance advantage; MFDAN uses optical flow to guide RGB image coding, and also introduces self-attention and Focal Loss. Although MiHF-Tr is slightly inferior in terms of UAR on SAMM compared to MFDAN, in terms of UAR and UF1 on CASME II and UF1 on SAMM, MiHF-Tr achieves better results. Especially on CASME II, MiHF-Tr's UF1 is 0.0216 higher than MFDAN's. Overall, our method achieves a competitive performance

Table 9: The results of interfering landmarks on CAMSE II.

e	UF1	UAR
1	0.9176	0.9160
3	0.9259	0.9204
10	0.9134	0.9100

5 Discussion

The proposed method effectively integrates three types of information, inevitably introducing more computational costs. However, compared with some current methods, such as Micron-bert (parameter size: 13.760512M), MiHFTr has a smaller computational cost (parameter size: 3.419074M). In addition, facial landmarks are crucial for the proposed method as they affect the feature discrimination of multi-information. As shown in Table 9, we conducted random interference on the facial landmarks of all samples (randomly deviating from the coordinate by e pixels). It turns out that the interference of facial landmarks leads to a decrease in recognition performance, but the magnitude of the decrease is limited. The reason may be that the proposed model utilizes ResNet50 to extract initial feature maps. Due to the receptive fields of convolution, even if there are errors for landmarks, the feature maps cropped by landmarks still contain effective information.

6 Conclusion

This paper explored the effectiveness of appearance, motion and geometric information, and proposes a novel MiHF-Tr model with a local and global hierarchy fusion framework. Several components were designed to improve performance. Extensive experiments were conducted, including quantitative analysis and visualization analysis. The experimental results demonstrate that aggregating appearance, motion and geometric information is effective, and MiHF-Tr can fully fuse multi-information by learning local and global semantic consistency. Also, the ablation analysis demonstrates that LAMiF, GCMiF, SC-SA and CW-FLoss are effective for MER task. Finally, compared with SOTA methods, the proposed method achieves competitive performance.

Acknowledgments

This work was supported in part by the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (Grant No.24KJB520022), in part by the Nanjing Science and Technology Innovation Foundation for Overseas Students (Grant No. RK002NLX23004), in part by Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No.NY223030).

References

[1] Yongtang Bao, Chenxi Wu, Peng Zhang, Caifeng Shan, Yue Qi, and Xianye Ben. 2024. Boosting micro-expression recognition via self-expression reconstruction and memory contrastive learning. *IEEE Transactions on Affective Computing* 15, 4 (2024), 2083–2096.

[2] Wenhao Cai, Junli Zhao, Ran Yi, Minjing Yu, Fuqing Duan, Zhenkuan Pan, and Yong-Jin Liu. 2024. Mfdan: Multi-level flow-driven attention network for micro-expression recognition. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[3] Bin Chen, Kun-Hong Liu, Yong Xu, Qing-Qiang Wu, and Jun-Feng Yao. 2023. Block Division Convolutional Network With Implicit Deep Features Augmentation for Micro-Expression Recognition. *IEEE Transactions on Multimedia* 25 (2023), 1345–1358. doi:10.1109/TMM.2022.3141616

[4] Haoyu Chen, Xin Liu, Xiaobai Li, Henglin Shi, and Guoying Zhao. 2019. Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–8.

[5] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. 2023. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision* 131, 6 (2023), 1346–1366.

[6] Haoyu Chen, Hao Tang, Radu Timofte, Luc V Gool, and Guoying Zhao. 2023. Lart: Neural correspondence learning with latent regularization transformer for 3d motion transfer. *Advances in Neural Information Processing Systems* 36 (2023), 43742–43753.

[7] Paul Ekman. 2009. Lie Catching and Microexpressions. In *The Philosophy of Deception*, Clancy Martin (Ed.). Oxford University Press, New York, NY, 118–133.

[8] Xinqi Fan, Xueli Chen, Mingjie Jiang, Ali Raza Shahid, and Hong Yan. 2023. Selfme: Self-supervised motion learning for micro-expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13834–13843.

[9] Chun-Ting Fang, Tsung-Jung Liu, and Kuan-Hsien Liu. 2024. Micro-Expression Recognition Based On 3DCNN Combined With GRU and New Attention Mechanism. In *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2466–2472.

[10] Wenjuan Gong, Yue Zhang, Wei Wang, Peng Cheng, and Jordi Gonzalez. 2023. Meta-MMFNet: Meta-learning-based multi-model fusion network for micro-expression recognition. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2 (2023), 1–20.

[11] Zhuoyao Gu, Miao Pang, Zhen Xing, Weimin Tan, Xuhao Jiang, and Bo Yan. 2024. Facial Micro-Motion-Aware Mixup for Micro-Expression Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8060–8064. doi:10.1109/ICASSP48485.2024.10446492

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1867–1874.

[14] Ankith Jain Rakesh Kumar and Bir Bhanu. 2021. Micro-expression classification based on landmark relations with graph attention convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1511–1520.

[15] Ling Lei, Jianfeng Li, Tong Chen, and Shigang Li. 2020. A novel graph-tcn with a graph structured representation for micro-expression recognition. In *Proceedings of the 28th ACM international conference on multimedia*. 2237–2245.

[16] Ke Li, Yuan Zong, Baolin Song, Jie Zhu, Jingang Shi, Wenming Zheng, and Li Zhao. 2019. Three-Stream Convolutional Neural Network for Micro-Expression Recognition. *Aust. J. Intell. Inf. Process. Syst.* 15, 3 (2019), 41–48.

[17] Yante Li, Xiaohua Huang, and Guoying Zhao. 2020. Joint local and global information learning with single apex frame detection for micro-expression recognition. *IEEE Transactions on Image Processing* 30 (2020), 249–263.

[18] Yante Li, Jinsheng Wei, Yang Liu, Janne Kauttonen, and Guoying Zhao. 2022. Deep learning for micro-expression recognition: A survey. *IEEE Transactions on Affective Computing* 13, 4 (2022), 2028–2046.

[19] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. 2025. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. *ICML* (2025).

[20] Zheng Lian, Haiyang Sun, Licai Sun, Lan Chen, Haoyu Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, et al. 2025. Open-vocabulary Multimodal Emotion Recognition: Dataset, Metric, and Benchmark. *ICML* (2025).

[21] Sze-Teng Liong, Yee Siang Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. 2019. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. IEEE, 1–5.

[22] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C-W Phan. 2018. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication* 62 (2018), 82–92.

[23] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10631–10642.

[24] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. 2019. A neural micro-expression recognizer. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. IEEE, 1–4.

[25] YongJin Liu, JinKai Zhang, WenJing Yan, SuJing Wang, Guoying Zhao, and Xiaolan Fu. 2015. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing* 7, 4 (2015), 299–310.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[27] Bingyang Ma, Lu Wang, Qingfen Wang, Haoran Wang, Ruolin Li, Lisheng Xu, Yongchun Li, and Hongchao Wei. 2025. Entire-Detail Motion Dual-Branch Network for Micro-Expression Recognition. *Pattern Recognition Letters* (2025).

[28] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. 2023. Micron-bert: Bert-based facial micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1482–1492.

[29] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. 2018. Learning-based video motion magnification. In *Proceedings of the European conference on computer vision (ECCV)*. 633–648.

[30] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. 2020. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2669–2676.

[31] BoKai Ruan, Ling Lo, HongHan Shuai, and WenHuang Cheng. 2022. Mimicking the annotation process for recognizing the micro expressions. In *Proceedings of the 30th ACM International Conference on Multimedia*. 228–236.

[32] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[33] Su-Jing Wang, Ying He, Jingting Li, and Xiaolan Fu. 2021. MESNet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos. *IEEE Transactions on Image Processing* 30 (2021), 3956–3969.

[34] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh. 2015. 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*. Springer, 525–537.

[35] Zebiao Wang, Mingyu Yang, Qingbin Jiao, Liang Xu, Bing Han, Yuhang Li, and Xin Tan. 2024. Two-level spatio-temporal feature fused two-stream network for micro-expression recognition. *Sensors* 24, 5 (2024), 1574.

[36] Zhifeng Wang, Kaihao Zhang, Wenhan Luo, and Ramesh Sankaranarayana. 2024. Htnet for micro-expression recognition. *Neurocomputing* 602 (2024), 128196.

[37] Jinsheng Wei, Haoyu Chen, Guanming Lu, Jingjie Yan, Yue Xie, and Guoying Zhao. 2023. Prior information based decomposition and reconstruction learning for micro-expression recognition. *IEICE TRANSACTIONS on Information and Systems* 106, 10 (2023), 1752–1756.

[38] Jinsheng Wei, Guanming Lu, and Jingjie Yan. 2021. A comparative study on movement feature in different directions for micro-expression recognition. *Neurocomputing* 449 (2021), 159–171.

[39] Jinsheng Wei, Guanming Lu, Jingjie Yan, and Yuan Zong. 2022. Learning two groups of discriminative features for micro-expression recognition. *Neurocomputing* 479 (2022), 22–36.

[40] Jinsheng Wei, Wei Peng, Guanming Lu, Yante Li, Jingjie Yan, and Guoying Zhao. 2023. Geometric graph representation with learnable graph structure and adaptive au constraint for micro-expression recognition. *IEEE Transactions on Affective Computing* 15, 3 (2023), 1343–1357.

[41] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. 2022. An overview of facial micro-expression analysis: Data, methodology and challenge.

IEEE Transactions on Affective Computing 14, 3 (2022), 1857–1875.

[42] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[43] Bing Yang, Jing Cheng, Yunxiang Yang, Bo Zhang, and Jianxin Li. 2021. MERTA: micro-expression recognition with ternary attentions. *Multimedia Tools and Applications* 80, 11 (2021), 1–16.

[44] Jun Yang, Zilu Wu, and Renbiao Wu. 2025. Micro-expression recognition based on contextual transformer networks. *The Visual Computer* 41, 3 (2025), 1527–1541.

[45] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. 2022. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4186–4196.

[46] Zhijun Zhai, Jianhui Zhao, Chengjiang Long, Wenju Xu, Shuangjiang He, and Huijuan Zhao. 2023. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22086–22095.

[47] Liangfei Zhang, Xiaopeng Hong, Ognjen Arandjelović, and Guoying Zhao. 2022. Short and long range relation based spatio-temporal transformer for micro-expression recognition. *IEEE Transactions on Affective Computing* 13, 4 (2022), 1973–1985.

[48] Jie Zhu, Yuan Zong, Hongli Chang, Yushun Xiao, and Li Zhao. 2022. A sparse-based transformer network with associated spatiotemporal feature for micro-expression recognition. *IEEE Signal Processing Letters* 29 (2022), 2073–2077.