This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3535569

Facial 3D Regional Structural Motion Representation Using Lightweight Point Cloud Networks for Micro-Expression Recognition

Ren Zhang, Jianqin Yin, *Member, IEEE,* Chao Qi, Yonghao Dang, Zehao Wang, Zhicheng Zhang, and Huaping Liu, *Senior Member, IEEE*

***Abstract*—Human-computer interaction (HCI) relies on understanding and adapting to users' emotional states. Micro-expressions (MEs), a critical component of emotional perception, are characterized by their spontaneity, rapidity, subtlety, and difficulty to control. They often reveal an individual's true emotions. A comprehensive and detailed representation of motion is necessary to capture the nuances of facial dynamics effectively. Presently, motion representation methods are predominantly confined to 2D analysis within RGB images, overlooking the critical role of facial structure and its movements in conveying emotions. To overcome this limitation, we introduce an innovative facial motion representation that encompasses 3D facial structure, regionalized RGB and structural motion features. Furthermore, we segment the face into eight distinct regions, selecting only the most significant motion points to delineate the primary motion characteristics of each area. To model the interactions among crucial facial motion regions, we employ an advanced, lightweight point cloud and graph convolution network (Lite-Point-GCN). Comprehensive testing on the $\mathrm{CAS(ME)}^3$ dataset, using leave-one-subject-out (LOSO), demonstrates that our method outperforms existing state-of-the-art methods.**

***Index Terms*—Micro-expression, Facial expressions, GCN, Objective class**

I. Introduction

FACIAL EXPRESSION RECOGNITION (FER) in human-computer interaction (HCI) is crucial for creating systems that can intuitively understand users' emotional states, enhancing user experience, communication effectiveness, and overall safety [1], [2]. A specific challenge within HCI is recognizing micro-expressions (MEs), fleeting facial expressions lasting between 1/25 and 1/5 of a second. These brief displays can reveal a person's genuine emotions [3], [4], even when they attempt to mask or suppress them [5]. And the micro-expression recognition(MER) is also vital for other fields like multimedia entertainment, film production, psychology, criminal analysis, and business negotiations [6], [7], especially in the realm of lie detection.

Micro-expression (ME) recognition faces two main challenges: first, due to the low intensity and brief duration of MEs, traditional methods often struggle to capture their subtle motion features effectively [8]. Many existing approaches focus on extracting appearance or texture features from 2D facial images to represent motion [9], while neglecting the crucial role of 3D facial structural movements, which are more robust to lighting changes and sensitive to subtle facial dynamics.

Second, the limited number of ME samples further exacerbates the issue, leading to the risk of overfitting. To address this, leveraging prior knowledge of facial structure can reduce reliance on large datasets. Graph Convolutional Networks (GCNs), by treating facial regions as nodes, can effectively model the relationships between motion regions and global features while incorporating structural priors [10], [11]. Therefore, developing a lightweight GCN network to model the interrelations between facial regions is essential, as it helps alleviate the challenges posed by the limited sample size.

Given these insights, we propose a novel 3D facial motion representation method that integrates facial 3D structure with optical flow, preserving spatial information while capturing temporal motion. The stability of 3D structures to lighting and pose variations enables more accurate capture of subtle facial dynamics.

Furthermore, we leverage the relationship between facial structure and expression as prior knowledge, segmenting the face into eight semantic regions. Dong et al. [12] demonstrated that different facial muscle movements correspond to distinct emotional categories. This segmentation helps identify emotion categories by analyzing the motion relationships between regions. By extracting unique motion features from each region, we reduce redundancy and precisely delineate motion areas, providing a feature representation rich in semantic content, as shown in Figure 1. By classifying motion locations into seven types, we achieve more effective regional partitioning. The detailed description of these objective categories is provided in Section IV-E.

To address the scarcity of ME samples, we introduce a lightweight point cloud GCN (Lite-Point-GCN) for MER. This network extracts local features from facial regions using a lightweight PointNet++ [13], combining spatial and motion information, and performs global modeling through GCN, effectively reducing the risk of overfitting.

Our key innovations are summarized below:

- We propose a novel facial motion representation method that integrates 3D facial structure with motion features. By segmenting facial semantic regions, the method pre-

The authors are with the School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhangren@bupt.edu.cn; jqyin@bupt.edu.cn; qichao199@163.com;dyh2018@bupt.edu.cn; 2199586379@qq.com; zczhang@bupt.edu.cn) Department of Computer Science and Technology, BNRist, Tsinghua University, Beijing 100084, China (hpliu@tsinghua.edu.cn). (Corresponding author: Jianqin Yin. e-mail: jqyin@bupt.edu.cn;)

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3535569

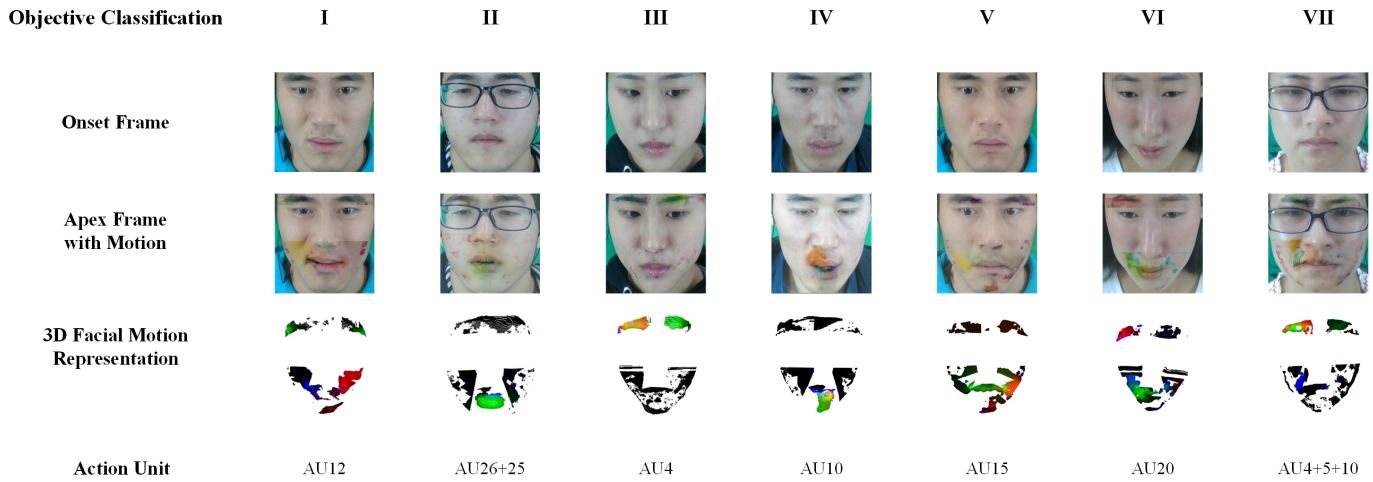


Fig. 1. The motion locations are categorized based on objective criteria. The first row shows the starting frame for each category, followed by the second row with keyframes and corresponding motion labels. The third row illustrates our proposed method for representing facial motion, aligning with the Action Unit (AU) categories. Red represents horizontal motion, green vertical, and blue depth (front-back). Overlapping colors indicate simultaneous motion in multiple directions. As each direction has both forward and backward motion, reverse motion is shown in black.

serves the spatial and temporal features of motion and leverages depth information to achieve more precise motion feature representation;

- We propose Lite-Point-GCN, a lightweight point cloud-based GCN that integrates facial structure and motion features. It effectively models local and global relationships among facial components, linking emotional categories to motion features;
- We conducted extensive experiments on the$\mathrm{CAS(ME)}^3$ dataset, incorporating depth information, to evaluate the proposed motion representation method. The results demonstrate its effectiveness and robustness across different validation methods. Combined with Lite-Point-GCN, our method surpasses existing state-of-the-art (SOTA) method.

II. Related Work

The accurate detection and recognition of ME pose significant challenges due to the limited availability of samples and the minimal range of movement involved. Thus, the effective extraction of motion features from various facial regions is essential [14]. Currently, feature representation is predominantly conducted through two methodologies: manual feature-based and learning-based approaches.

A. Manual feature representation methods

Within the manual feature-based framework, the utilization of geometric features to represent ME movements has seen notable success. Specifically, the optical flow method has been instrumental in accurately tracking the movement changes across facial points, offering a quantifiable approach to detect subtle facial muscle movements.

The Main Directional Mean Optical-flow (MDMO) [15] approach leverages traditional optical flow method to divide the face into sections aligned with action units (AUs), extracting predominant directions and mean optical flow values within each segment for feature representation. This method facilitates emotional category classification of ME by mapping distinct optical flow directions to corresponding facial regions and applying a cost-effective Support Vector Machine (SVM) classification approach.

Conversely, the Neural Micro-Expression Recognition (NMER) [16] strategy employs optical flow data from onset and apex expression frames as neural network inputs, utilizes macro-expression (MaE) samples, and implements methods to amplify ME movements while minimizing those of MaE. This approach effectively bridges domain gaps and enhances MER capabilities.

Building on the foundations of NMER, the Balanced Micro-Expression Recognition (BMER) method [17] refines amplification and reduction methods, dynamically adjusting these parameters in response to facial movement intensities. Further, it incorporates strategies like facial alignment within optical flow contexts, thus improving the clarity and expression of optical flow features as facial motion indicators. Notably, BMER's advancements in optical flow feature articulation significantly elevate recognition performance, underscoring the critical role of precise motion feature representation in MER.

B. Neural Network Feature Representation Method

To mitigate the challenge posed by a limited number of samples, various strategies have been implemented for the self-supervised learning of motion feature. For instance, Fan et al. [18] employed self-supervised learning methods to identify and quantify local motions associated with key movement points, subsequently synthesizing a dense motion field through a weighted amalgamation of these localized movements. This process involved integrating static frames with motion data to create synthetic post-movement frames that closely mimic true post-movement scenarios, thereby ensuring the accuracy of the motion information in reflecting genuine facial movements. In the final step, they leveraged a Transformer model to

categorize emotions based on the synthesized dense motion field.

Lei et al. [11] investigated video-based motion amplification technology and developed an innovative spatiotemporal convolutional network. This network is designed to capture motion representations by learning shape changes of individual vertices in the onset frame and subsequently incorporating selected features of these representations into a graph neural network for further learning. This methodology underscores the significance of GCN in MER. However, it retains only a fraction of facial motion information, thus limiting its ability to comprehensively model the relationship between facial movements and emotions.

Nguyen et al. [19] leveraged the deep bidirectional transformer (BERT) model for facial MER, introducing an innovative Diagonal Micro-Attention (DMA) mechanism alongside a Patch of Interest (PoI) module. These tools are adept at capturing and enhancing subtle facial movements. The DMA mechanism excels in identifying minute discrepancies between frames and pinpointing facial movements accurately. Meanwhile, the PoI module concentrates on identifying and emphasizing crucial ME zones, minimizing background distractions and interference. This targeted localization of motion regions is effectively represented in the network by a vast dataset, demonstrating significant relevance to the network's performance.

In summary, while manual feature representation approaches notably delineate facial movements through optical flow, deep learning methods merge network architecture with features to both overtly and covertly extract ME motion features. These approaches underscore that advanced motion representation methods markedly boost recognition precision. Nonetheless, they are constrained to deriving features from 2D imagery, susceptible to variations in lighting and image quality. Hence, our method integrates texture and structural motion presentation, employing graph convolution to elucidate the interrelations among distinct regions, with the goal of improving both the robustness and accuracy of expression detection.

III. Method

We proposed an innovative approach to representing facial movements, offering a more accurate depiction of ME-related facial actions. Furthermore, we developed a compact point cloud feature fusion network, enhanced with a GCN, dubbed the Lite-Point-GCN network. This methodology utilizes point clouds to represent the facial structure (Section. III-A), integrates 3D motion information as color features alongside facial positional data (Section. III-B), and segments the face into eight regions based on semantic information, focusing on features with significant movements within each region (Section. III-C). A streamlined point cloud network is employed to extract relevant features for each segment, and GCN is applied to delineate the relationships among these features, aiming for precise emotion classification outcomes (Section. III-D). Figure 2 comprehensively illustrates our methodological framework.

A. Preliminaries

Compared to 2D images, 3D facial structures can more accurately capture dynamic changes in MEs, offering better resistance to interference from lighting and pose changes. To accurately derive facial structure information from images, depth maps are utilized to facilitate a precise one-to-one mapping between the facial structure and the images.

Given a ME dataset, $M = \{V_j\}_{j=1}^{n_s}$, where the V_j represents the video sequence of the $j\text{-}th$ sample, and the n_s represents the total number of samples in M. And the $V_j = \{(D_i, I_i)\}_{i=1}^{n}$ with $D_i \in \mathbb{R}^{H\times W}$ and $I_i \in \mathbb{R}^{H\times W\times 3}$, where D_i represent the depth map of the $i\text{-}th$ frame, and I_i represents the corresponding color image, the n represents the total number of frames in V_j, and the $H \times W$ represents the pixel resolution of the image.

The D_i can be transferred from the pixel location and depth information to the 3D point cloud $P_i \in \mathbb{R}^{N\times 3}$, where N represents the number of valid points in the point cloud, corresponding to pixels with valid (non-zero) depth values.

We aim to capture the motion of the face in the video. For each V_j, we extract F_o = (P_o, I_o) and F_a = (P_a, I_a) of the onset frame F_o and the apex frame F_a to capture facial motion, where P_o and P_a represent the 3D point clouds at the onset and apex frames, respectively, while I_o and I_a represent the corresponding 2D images. Together, they provide a comprehensive view of facial motion dynamics across these key frames.

Given the camera's intrinsic parameter $\{(f_x, f_y), (c_x, c_y)\}$, where (f_x, f_y) represent the focal lengths (in pixels), and (c_x, c_y) denote the coordinates of the principal point (the optical center of the image), typically calibrated during the camera manufacturing or calibration process. Based on the pinhole camera model, each pixel (x, y) in the depth map D_i can be mapped to the 3D spatial coordinates (X, Y, Z) in the camera's coordinate system as follows:

$$\begin{aligned} X &= (x - c_x) \times D_i(x, y)/f_x \\ Y &= (y - c_y) \times D_i(x, y)/f_y \\ Z &= D_i(x, y) \end{aligned} \tag{1}$$

where $D_i(x, y)$ represents the depth value at the pixel located at row x and column y in the depth map D_i. Each pixel in the 2D image $I_i(x, y)$ has an associated color value corresponding to the same location in the depth value $D_i(x, y)$.

By associating the color information $I_i(x, y)$ with each 3D point (X, Y, Z), a colorized point cloud can be generated, offering a detailed representation of the visual and motion features of each point. Applying this transformation to all valid pixels in D_i yields a complete point cloud.

B. 3D Spatiotemporal Representation of Facial Dynamics

A significant challenge in analyzing ME lies in the subtlety of facial movements. To describe facial movements accurately and clearly, the optical flow method is employed, to capture the dynamic features of facial pixel movements over time. Unlike network-based approaches for learning facial motion expressions, the optical flow method provides an intuitive

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3535569

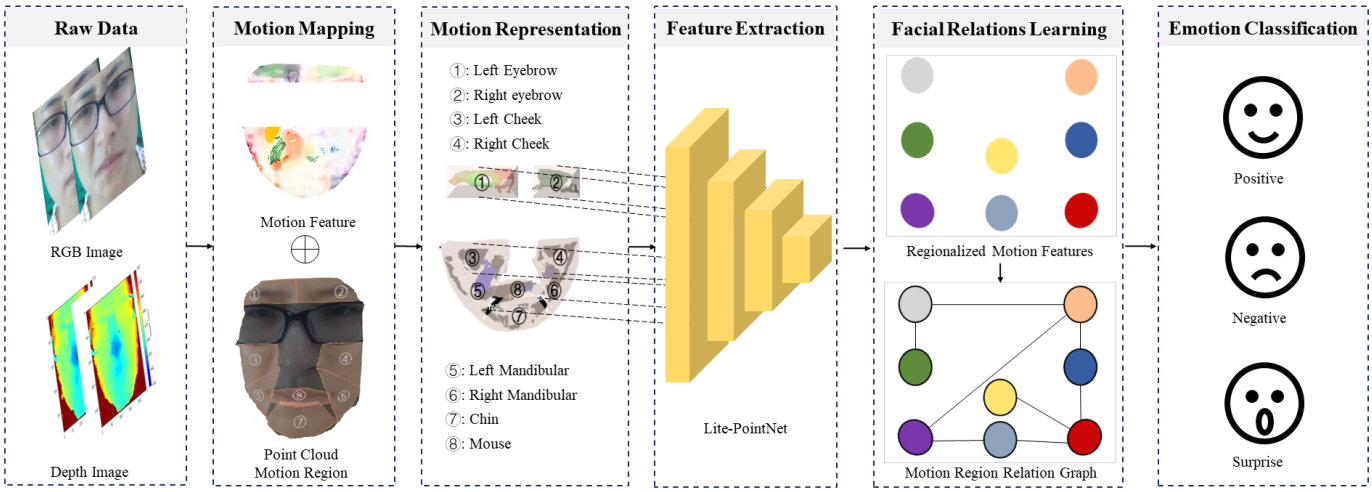


Fig. 2. Facial motion classification acquires spatial and motion information from depth and RGB images. Utilizing facial semantics as a basis, we segment the face into distinct sections, then map and filter the motion information. Motion features are extracted through Lite-PointNet, and GCN models the relationships among facial regions, with SoftMax achieving precise recognition for each ME category.

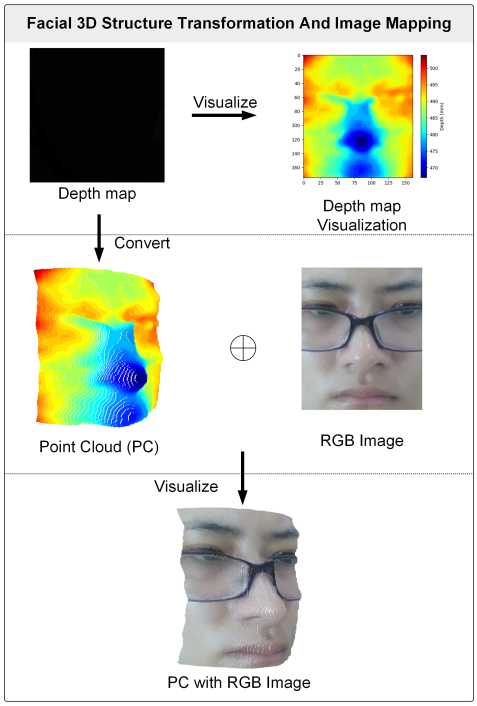


Fig. 3. Upon converting the depth map into a 3D point cloud representation, each point is mapped directly to the corresponding pixel in the RGB image. This process ensures a precise one-to-one correspondence, allowing for the seamless integration of pixel variations depicted by optical flow and structural transformations conveyed by the depth map into the point cloud's individual points.

representation of the temporal dynamics of ME, representing the pixel's motion over time with a 2D vector:

$$\left(u^t, v^t\right) \tag{2}$$

Where, u and v denote the optical flow components in the horizontal and vertical directions at time t, respectively.

Following the alignment and cropping of the face to minimize environmental influences, we adjust the optical flow field by performing an affine transformation on the facial keypoints of the onset frame I_o and the apex frame I_a, to ensure an accurate depiction of optical flow motions focused on the face, ignoring background interference, drawing inspiration from references [17], [15]. We can obtain the displacement of each pixel in the (x, y) direction, namely, the horizontal motion u_i and vertical motion v_i for pixel i.

Nevertheless, given the variability in lighting and environmental factors, the 3D structural motion of the face more accurately conveys facial motion information. It is evident that numerous details of structural motion are not discernible through pixel changes alone. To capture this structural motion information, we employ the following equation:

$$\Delta z_i = z_i^a - z_i^o \tag{3}$$

where z_i^o denotes the z-coordinate of the i^{th} point in P_o, and z_i^a denotes the z-coordinate of the i^{th} point in P_a. Here, Δz_i represents the difference in depth values between these two points, capturing the z-axis displacement due to structural motion.

The motion information of the i^{th} point along the x, y, and z axes is then encapsulated as $\Delta P_i = (u_i, v_i, \Delta z_i)$, where each component describes the displacement in its respective axis. We define the complete motion set $\Delta P = \{\Delta P_i\}_{i=1}^N$, capturing the motion information for all N points. Consequently, the V_j is represented by the point cloud $V_{\text{point}} = (P_a, \Delta P)$, where $P_a \in \mathbb{R}^{N\times 3}$, $\Delta P \in \mathbb{R}^{N\times 3}$ and $V_{\text{point}} \in \mathbb{R}^{N\times 6}$. Each point in V_{point} contains six features, encapsulating both the spatial information from the apex frame and the motion information of the subject within V_j.

The optical flow reveals the motion of pixel points in image sequences, highlighting temporal dynamics, whereas depth flow captures depth variations in the scene, elucidating spatial structure changes. Integrating these insights allows for a more

nuanced detection of subtle motion variations induced by ME. After determining the facial motion regions and directions, the motion magnitude does not affect the expression category [16]. Therefore, we standardize the motion features in the (x, y, z) directions across color channels, normalizing both negative and positive motions to the range $[-1, 1]$. The magnitude of motion for each point in the facial region represents its relative motion. The normalization is employed to decrease overall facial movements and highlight local area movements, while also preserving key motion points in the point cloud.

Additionally, to ensure spatiotemporal consistency between the onset and apex frames, we employed the dlib algorithm to detect and crop the facial region in RGB images, aligning the point clouds of the two frames based on the nasal keypoints. However, due to possible pixel-level deviations inherent in the dlib detection algorithm, the facial point cloud may exhibit an overall displacement. To eliminate this displacement while preserving prominent motion regions, we applied normalization to remove global offset differences. By integrating three-dimensional motion data with spatial positioning, this approach provides a comprehensive spatial motion representation for each point in the apex framework, enabling a detailed depiction of facial motion characteristics.

C. Semantic Regionalization of Facial Motion Representation

The correlation of movements in various facial regions with distinct expressions has been established, yet prior research predominantly targeted eyebrow and mouth motions [12] . Dong et al. [18] highlighted a significant link between the activities of the zygomaticus (smile muscle) and corrugator supercilii (frown muscle) with the conveyance of positive and negative emotions, respectively, in facial expressions. To establish motor relationships between muscles in different facial regions, we segment the face into eight key regions: left and right eyebrows, cheeks, mandibular, mouth, and chin, based on each region's critical role in expressing emotions. This segmentation also accounts for individual variances and external factors like glasses, along with static features like blinking and nose movements, which were deliberately omitted to avoid analysis disruption.

Specifically, we use the 68 facial landmarks detected by dlib, connecting certain key points to define boundaries for facial regions. By combining these landmarks with the spatial coordinates of the point cloud(excluding prominent areas like the nose and glasses), we achieve a semantic segmentation of the face into regions, retaining point clouds from eight facial areas, as shown in Figure 4. The point cloud and motion feature set for the apex frame, covering the eight semantic regions, is defined as $\hat{V_{\text{part}}} = \{(P_{a_k}, \Delta P_k)\}_{k=1}^{8} \in \mathbb{R}^{\hat{N}\times 6}$, where $\hat{N}$ represents the total number of points in each semantic region (with $\hat{N} = \sum_{m=1}^{8} N_m$), and the 6 represents the three spatial dimensions of the point cloud P_{a_k} and the three motion feature dimensions ΔP_k. The details of the specific segmentation method will be described in detail in subsequent Section IV-B2.

While point cloud networks are traditionally geared towards object shape and spatial point distribution, the subtle nature of MEs renders facial shape distinctions too nuanced for effective emotional categorization. To address this, we pivot towards leveraging motion features to prune extraneous elements, thereby accentuating expression-specific disparities. We employ the L2 norm for 3D facial feature points, prioritizing them according to these metrics:

$$\|\Delta P_i(t)\|_2 = \sqrt{u_i^2(t) + v_i^2(t) + \Delta z_i^2(t)} \tag{4}$$

Here, $u_i(t)$, $v_i(t)$, and $\Delta z_i(t)$ represent the changes in x, y, and z directions over time t of point i. In our method, they are defined as the displacements between the onset frame and the apex frame, with $\|\Delta P_i\|_2$ representing the magnitude of this displacement.

Evaluating the magnitude of motion, we prioritize all feature points in each N_m and select the top 1024 point with the most significant motion for network training, effectively focusing on those with the highest motion magnitude.

Ultimately, the V_{part} denotes the filtered representation of each segmented region in $\hat{V_{\text{part}}}$, which encapsulates the distinctive motion characteristics of each region. Formally, $V_{\text{part}} \in \mathbb{R}^{8\times 1024\times 6}$, where the representation encodes critical motion attributes, including magnitude, direction, and spatial positioning. This formulation provides a comprehensive and structured depiction of motion dynamics, ensuring a holistic understanding of regional motion features.

For visualization, the forward motion features across different axes are standardized to the [0,255] range through color channels: red for the X-axis, green for the Y-axis, and blue for the Z-axis. Intensities of these colors correlate directly with the motion's magnitude in their respective directions, as illustrated in Figure 4.

D. Lite-Point-GCN

A significant challenge in ME analysis is the scarcity of samples, which predisposes networks to over-fitting. To mitigate this problem, we introduce a lightweight network tailored for MER, adept at processing both temporal and spatial data within facial point clouds. This network employs a local-to-global strategy to meticulously capture facial dynamics, shown in Figure 5.

1) Local Regional Motion Feature Extraction: The detailed analysis of facial expressions requires a detailed approach, focusing on the extraction of local features from specific semantic facial regions. These regions, crucial for nuanced emotional expression, need targeted attention to precisely capture the subtleties of human emotions. The key to understanding these local dynamics is recognizing that a facial expression is a combination of various region-specific movements. This understanding underpins our methodological framework.

To achieve this objective, we begin by dividing the face into distinct point clouds that represent specific semantic regions, using a lightweight point cloud network for comprehensive analysis. This focus on localized areas allows our network to effectively identify both structural and temporal variations in each segment, thereby deriving motion feature vectors essential for decoding facial expressions.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3535569

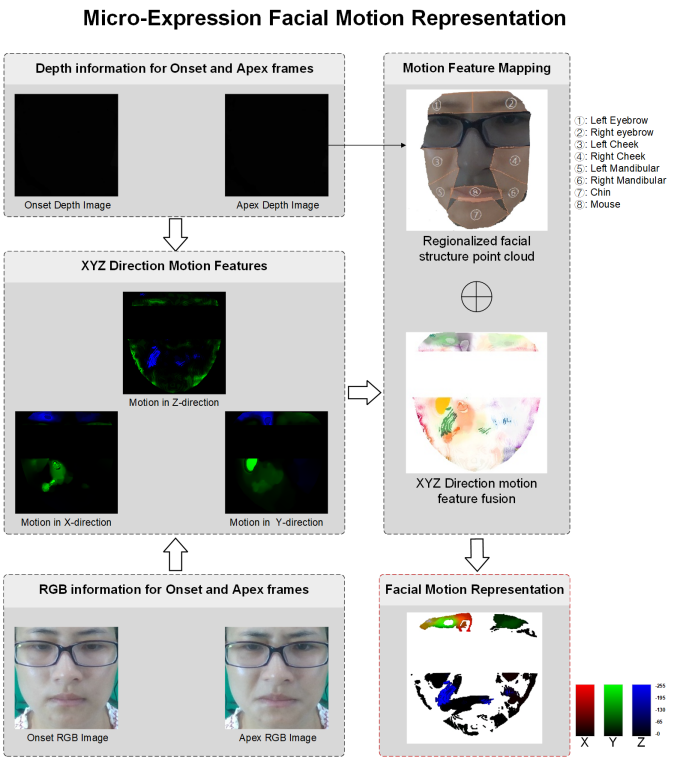


Fig. 4. The motion feature schematic. The optical flow values are initially derived from the RGB image of the apex frame of the onset frame to capture motion information in the X and Y directions. Subsequently, motion values of facial structures are computed from a depth map to ascertain motion in the Z direction. The apex frame from the depth map is then transformed into a point cloud, with motion attributes of each facial region mapped onto the corresponding point cloud segments. By delineating facial regions according to semantic definitions and selectively filtering motion attributes within these areas, a detailed and comprehensive mapping of facial motion regions is achieved.

We first validate the extracted features using the PointNet++ [13], as illustrated in Figure 6. PointNet++ achieves robust feature extraction by constructing a hierarchical organization of points within local regions, progressively abstracting increasingly larger spatial structures along the hierarchy. This approach captures the critical post-motion features in facial regions. Building on the principles of PointNet++, our model incorporates Hierarchical Feature Learning to systematically extract and integrate local, spatial, and motion features at multiple levels.

However, unlike PointNet++, we bypass the computationally intensive processes of point sampling and grouping in local neighborhoods. Instead, our approach implements a more efficient feature extraction mechanism, which retains the benefits of hierarchical feature learning while reducing computational complexity.

Our method reduces reliance on sample numbers by using single-pass global sampling for each facial segment. By jointly modeling point cloud positions and motion characteristics, it effectively integrates spatial and motion features, forming a strong link between motion patterns and expression categories.

Additionally, it captures the intricate interactions between motion amplitude and direction. Through hierarchical abstraction across all segments, unique feature vectors are generated for each region, underscoring the critical role of local feature extraction (as shown in Figure 5: Lite-Pointnet). This focus on localized analysis is essential, not optional, for accurately capturing expressions, enabling each region to contribute its distinctive attributes to the overall dynamics of facial motion.

2) Global Motion Feature Relation Learning: To understand the complex dynamics of facial expressions fully, it is crucial to integrate local feature interactions from various facial components into a cohesive global model. This integration not only clarifies the role of individual facial regions but also demonstrates how these regions work together to create comprehensive expressions.

Consequently, the use of GCN is essential. GCNs excel in merging localized features into a unified global model, capturing the diverse aspects of facial expressions accurately. Their ability to identify complex relationships among data points makes them especially suitable for modeling the detailed interactions between facial regions.

Following this, we utilize GCN to encapsulate the interactions among the eight facial regions. The graph $G = (R, E)$ is constructed, with the node set R denoting the facial regions and the edge set E depicting the interconnections among these regions. Each node feature is characterized by the motion feature vector Φ, derived from the point cloud network. Through convolution operations on this graph, GCN are capable of apprehending complex, higher-order relationships between the facial regions:

$$H = \mathrm{GCN}(G, \Phi) \tag{5}$$

Where, H signifies the feature representation derived from GCN processing, encapsulating the interrelations among various facial regions.

More precisely, GCN processes the node feature matrix $X \in \mathbb{R}^{d \times N_{node}}$ and an adjacency matrix $A \in \mathbb{R}^{N_{node} \times N_{node}}$, where N_{node} indicates the number of nodes, and d signifies the feature description's dimensionality for each motion region node. The culmination of L layers of GCN processing yields embedded nodes, denoted by the final hidden layer X^l. The operation of each GCN layer is described as follows:

$$X^l = \sigma(A \times X^{l-1} \times W^{l-1}) \tag{6}$$

Where, σ signifies a non-linear activation function, while $W^{l-1} \in \mathbb{R}^{d_i \times d_o}$ corresponds to the weight matrix of the $(l-1)^{th}$ layer, with d_i and d_o indicating the input and output dimensions, respectively, of layer l.

To elucidate the interconnections among facial regions, the adjacency matrix A is designated as a learnable parameter, permitting the model to refine the graph's structure throughout training. This adaptability of the adjacency matrix empowers the GCN to more adeptly delineate the structural motion across facial regions. Furthermore, global max pooling is employed to consolidate node features from each graph into a singular graph feature vector, thereby encapsulating global attribute information.

IV. Experiments

Incorporating depth information enables the detection of structural changes. Given that the $\mathrm{CAS(ME)}^3$ dataset stands

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3535569

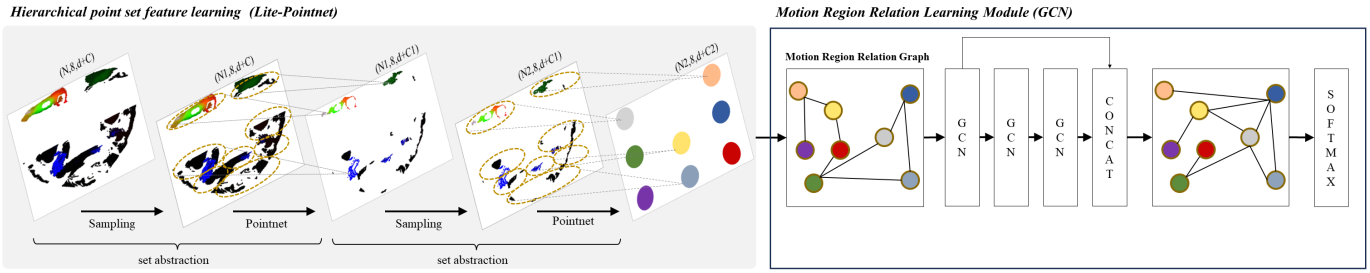


Fig. 5. The architecture of the Lite-Point-GCN network is designed to process facial motion representations by employing downsampling methods, extracting features from local regions, and mapping these to specific feature points within each area. The GCN is then utilized to model the interactions among different motion regions and decipher their interconnections. The classification of the emotion category is ultimately achieved using a Softmax classifier.

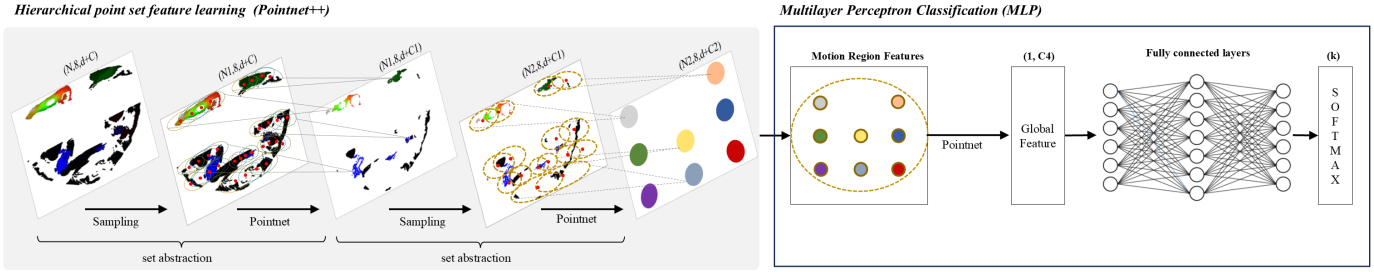


Fig. 6. The PointNet++ network effectively extracts motion features from local regions by utilizing a hierarchical point set feature learning approach. This method saturates the feature extraction from each local region by employing grouping and hierarchical strategies, mapping these features onto specific feature points within the region to represent facial motion. In the figure, the red dots represent sampling centers, and there may be overlapping content in the sampling of the same region. Subsequently, these local features are aggregated into global features through a global pooling layer and processed through fully connected layers, resulting in the output of classification results.

as the sole dataset equipped with depth information, our study aimed to assess the impact of leveraging depth information for MER. Furthermore, ablation experiments were carried out on the network parameters to substantiate the robustness of each component of the model design.

A. Dataset

The CAS(ME)^3 dataset is an advanced multimodal resource designed to enhance research in MEs and emotion recognition. This dataset uniquely includes naturally occurring MEs from simulated crime scenarios, coupled with detailed physiological and acoustic data, providing a rich, ecologically valid platform for in-depth analysis. It features high-definition facial recordings (1280×720) and introduces depth information along with electrodermal activity (EDA) and sound signals, capturing subtle facial changes and enhancing ME analysis precision. The dataset includes 860 MEs and 3,346 MaEs, meticulously annotated into objective [20] and emotional categories to support automated emotion recognition. This rigorous annotation follows the AUs framework, distinguishing seven muscle movement groups. Emotional categorizations are based on expert evaluations within an established framework, categorizing expressions into seven emotions: sadness, anger, disgust, fear, happiness, surprise, and others. The first four are negative, with happiness as a positive emotion, demonstrating the dataset's comprehensive approach to studying facial expressions in emotion research.

B. Experimental details

1) Experimental details of facial motion representation: Given that the dataset samples exhibit errors, redundancies, and, due to occlusions, certain facial expressions are indiscernible, we have purged incorrect samples, resulting in 784 ME annotations categorized under seven emotion labels and four emotion categories: positive, negative, surprise, and other. Furthermore, 657 ME samples were selected for experimentation, excluding those labeled as "other."

To diminish background noise, the dlib algorithm was employed to identify the facial region in the depth and RGB images' onset and apex frames. This facilitated the transformation of pixel position and depth data into two point clouds, utilizing depth camera focal lengths f_x and f_y at 1324.65, centering the image at $(W/2, H/2)$, and applying a scaling factor of 1000. This methodology ensured the retention of comprehensive motion data while omitting point clouds beyond 1.5 times the mean length from the center points, thereby conserving facial structure integrity and minimizing background disturbances.

For motion features on the color channels, we observed that the displacements in the three directions are of the same order magnitude. Therefore, we normalized the motion features in the three directions (u, v and Δz) simultaneously.

For both positive and negative directional motion features, we recorded the maximum and minimum values separately. Using the min-max normalization approach, we normalized positive directional motion features to the range [0,1] and negative directional motion features to the range [-1,0]. By

TABLE I
DEFINITION OF FACIAL REGIONS AND CORRESPONDING KEY LANDMARKS.

Region	Landmarks (Indices)	Description
Left Cheek	1–5, 31	Covers the left cheek area, primarily defined by the left facial contour.
Right Cheek	12–16, 35	Defines the right cheek area, constructed from right facial contour landmarks.
Left Mandibular	4–7, 31, 48	Encompasses the left mandibular region along the left jawline.
Right Mandibular	10–13, 35, 54	Covers the right mandibular region along the right jawline.
Mouth	48–55	Encloses the mouth area, defined by mouth perimeter points.
Chin	6–11, 55–60	Forms the chin region, capturing the main structural area of the chin.
Left Eyebrow	17-21	Left eyebrow area defined above the face's horizontal midpoint and left brow ridge.
Right Eyebrow	22-26	Right eyebrow area, positioned above the face's horizontal midpoint and right brow ridge.

concatenating these, the normalized motion features are constrained to the range [-1,1].

This normalization method enables us to standardize motion features with varying amplitudes across different micro-expressions, eliminating any global offsets and focusing on critical local motion regions in facial features, thereby ensuring analytical consistency.

2) Experimental details of facial region segmentation: Table I outlines the segmentation of facial regions based on specific key landmarks derived from dlib's 68-point facial landmark model. Each facial region, including the cheeks, mandibular areas, mouth, chin, and eyebrows, is defined by a unique set of landmark indices to ensure precise coverage of that area.

The left cheek region is defined by landmarks `1–5` and `31`, which capture the cheek's contour on the left side of the face. The right cheek region is similarly segmented using landmarks `12–16` and `35`. For the left mandibular region, landmarks `4–7`, `31`, and `48` define the area along the left jawline, while the right mandibular region is outlined by landmarks `10–13`, `35`, and `54`. The mouth area is represented by landmarks `48` through `55`, capturing the mouth's complete boundary. The chin region includes landmarks `6–11` and `55–60`, accurately covering the structural area of the chin. Finally, the left eyebrow region is segmented using landmarks `17–21`, while the right eyebrow region is defined by landmarks `22–26`.

This structured segmentation, as presented in Table I, provides a robust foundation for in-depth facial feature analysis, offering well-defined regions consistently referenced in subsequent image processing tasks.

3) Experimental details of network structure: In the Lite-Point-GCN network, the Adam optimizer is deployed with a weight decay of 0.0001, a batch size of 24, and an initial learning rate of 0.001. The network ingests 1024 point clouds as input, encompassing position and color channels that detail motion's position, magnitude, and direction. Network architecture includes intermediary layers designated for point abstraction and aggregation, sized at 512 and 128, respectively. Within the GCN module, a hidden layer dimension of 512 is specified, and to counteract over-fitting, a 4-layer GCN configuration is employed, accommodating 8 nodes.

4) Evaluation metrics: To ensure the accuracy of the test, the traditional leave-one-subject-out approach (LOSO) was used in the testing phase, same as state-of-the-art (SOTA) method. The unweighted F1 score (UF1) and the unweighted average recall (UAR) were used as performance metrics to avoid over-fitting the proposed method to a particular class. The true positives (TP), false positives (FP), and false negatives (FN) is assumed to exceed the number of subjects, and UF1 is calculated as

$$\mathrm{UF1} = \frac{1}{C}\sum\nolimits_{i}^{C}\frac{2 * TP_i}{TP_i + FP_i + FN_i} \tag{7}$$

where C is the total number of MEs, and UAR can be expressed as

$$\mathrm{UAR} = \frac{1}{C}\sum\nolimits_{i}^{C}\frac{TP_i}{S_i} \tag{8}$$

where the S_i is the total number of samples belonging to the i^{th} ME class in the dataset.

C. Experimental results and analysis

1) Comparison to State-of-the-art Methods: To evaluate the efficacy of our facial movement representation, we conducted a comparison with the SOTA method. As presented in Table II of $\mathrm{CAS(ME)}^3$ [21], our results show a marked improvement.

To ensure a fair assessment across different emotion categories, we conducted tests with sets of 3, 4, and 7 emotion categories. Our method achieved a 10% improvement in each category over the comparative method, as highlighted in bold in Table II.

Specifically, in the three-emotion category, our method registered a UF1 score of 68.19% and an UAR of 74.12%, surpassing the μ-BERT model trained on extensive datasets by over 10%. In the four-emotion category, our results were UF1 of 47.64% and UAR of 53.66%, approximately 10% higher than the baseline (+ depth) and on par with μ-BERT. For the seven-emotion category, our method also exceeded the baseline (+ depth) by about 10%, achieving a UF1 of 35.64% and a UAR of 41.59%, reflecting a roughly 3% improvement over the μ-BERT SOTA method.

The μ-BERT approach amalgamates vast quantities of data from various datasets and employs a substantial volume of unlabeled data in a self-supervised manner to guarantee a sufficient sample size. Our methodology significantly outshone in the three-class scenario, while the results in the four-class and seven-class scenarios were less pronounced, likely due to sample scarcity affecting the division into emotional categories. Moreover, the network exhibits heightened sensitivity to movement positions, complicating the classification of the "other" label as an emotion category based on movement positions alone, thereby notably reducing accuracy.

TABLE II
MER ON THE CAS(ME)^3 DATASET.

Method	Classes	UF1 (%)	UAR (%)
FR [22]	3	34.93	34.13
STSTNet [23]	3	37.95	37.92
RCN-A [24]	3	39.28	38.93
μ-BERT [19]	3	56.04	61.25
Ours	**3**	**68.19**	**74.12**
Baseline [21]	4	29.15	29.10
Baseline(+Depth) [21]	4	30.01	29.82
μ-BERT [19]	4	47.18	49.13
Ours	**4**	**47.64**	**53.66**
Baseline [21]	7	17.59	18.01
Baseline(+Depth) [21]	7	17.73	18.29
μ-BERT [19]	7	32.64	32.54
Ours	**7**	**35.64**	**41.59**

2) Comparison to multiple backbone networks: To assess the efficacy of our proposed facial movement representation, we performed evaluations across several leading point cloud networks, such as KP-Conv [25], PointNet [26], Point Transformer (PT-v1) [27], and PointNet++ [13]. Each of these networks embodies a distinct foundational approach.

KP-Conv, a versatile network for point cloud processing, leverages deformable convolutional kernels to capture the local structural nuances of point clouds, thereby enhancing the model's adaptability to non-uniform point distributions. Point Transformer-V1 incorporates a self-attention mechanism, enabling the model to dynamically adjust weights among points within the cloud, thus bolstering its capacity to comprehend complex spatial relationships. PointNet excels at deriving global features from point clouds, demonstrating resilience to permutations of input data and robustness against geometric alterations.

Extending PointNet's capabilities, PointNet++ introduces refined mechanisms for multi-scale grouping and localized feature extraction, significantly boosting the processing of varied point cloud densities and the delineation of intricate local geometric patterns. Our facial movement representation outperformed both the baseline and SOTA benchmarks on these platforms, as documented in Table III, with the initial row highlighting the latest SOTA achievements.

We conducted experiments on multiple point cloud networks, namely KP-Conv [25], PointNet [26], Point Transformer (PT-v1) [27], and Pointnet++ [13]. These experiments aimed to validate the effectiveness of the proposed ME point cloud representation. Interestingly, all of the tested networks outperformed both the baseline and the current state-of-the-art methods. The comparative results are presented in Table III. And the first row in Table III shows the optimal results achieved using the image.

We have demonstrated that our facial movement representation surpasses state-of-the-art achievements across several foundational framework networks. Furthermore, it outperforms prevalent point cloud networks in recognition capabilities. Our findings reveal that within these networks, PointNet++ employs a global modeling strategy to effectively link movement positions with expressions, elucidating the intricate relationship between them.

TABLE III
MULTIPLE MODEL VALIDATION OF ME POINT CLOUD REPRESENTATION EFFECTIVENESS.

Method	Backbone	UF1(%)	UAR(%)
μ-BERT [19]	Image-Trasnformer	56.04	61.25
PointNet [26]	Point-MLP	62.64	66.72
KP-Conv [25]	Point-CNN	62.96	67.97
PT-v1 [27]	Point-Transformer	59.62	64.67
PointNet++ [13]	Point-MLP	63.37	67.87
Ours	**Point-GCN**	**68.19**	**74.12**

Conversely, our network leverages GCN to precisely model these movements, fostering a more streamlined architecture and a superior methodological approach, which culminates in unparalleled MER efficiency, as evidenced by the highlighted data in Table III.

Additionally, our evaluations included KP-Conv, utilizing point convolution technology, and PT-v1, employing Transformer mechanisms. In scenarios involving three emotion categories, KP-Conv registered a UF1 score of 62.96% and UAR of 67.97%, while PT-v1 recorded UF1 of 59.62% and UAR of 64.67%. For PointNet, we achieved UF1 of 62.64% and UAR of 66.72%; PointNet++ exhibited UF1 of 63.37% and UAR of 67.87%.

Graph Convolutional Networks (GCNs) offer significant advantages in global modeling for MER, surpassing the performance of Multi-Layer Perceptrons (MLPs) in PointNet++. By leveraging interconnections among facial regions, GCNs provide a detailed and comprehensive interpretation of facial expressions, resulting in improved recognition accuracy. Furthermore, the flexibility of GCN architectures enables iterative learning and optimization of global models, demonstrating their resilience and adaptability.

In contrast, MLPs fail to capture the intricate network of relationships specific to facial expressions, primarily due to their failure to utilize the topological and relational information in the graph structure. The integration of graph features with topological frameworks is crucial in MER.

Although our global facial movement modeling method shows promising results across various point cloud networks, the scarcity of facial expression samples poses a risk of overfitting in complex models due to limited data availability. Thus, we advocate for the development of lightweight networks to mitigate overfitting risks and enhance the model's generalization capabilities.

D. Ablation studies

1) Ablation studies on the selection and number of points: In the realm of facial feature representation, the quantity and method of point selection are pivotal factors. It is posited that points exhibiting a greater motion more accurately capture the nuances of facial dynamics. Moreover, achieving a comprehensive depiction of facial movements using the minimum number of points is essential.

To this end, a series of ablation studies were undertaken to assess the effects of various selection methods and the number

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3535569

of points on the quality of facial representation. Specifically, we explored three point selection strategies: sorted sampling, random sampling, and central sampling. The evaluation considered sets of 512 and 1024 points, as detailed in Table IV.

TABLE IV
THE SELECTION AND NUMBER OF POINTS.

Point number		Point selection			Evaluation Metrics	
1024	512	Center	Sort	Random	UF1(%)	UAR(%)
✓			✓		**68.19**	**74.12**
	✓		✓		63.87	67.97
✓		✓			42.27	46.87
✓			✓		**68.19**	**74.12**
✓				✓	44.37	49.87

Within the context of three emotion categories, we discovered that employing a sorting method with 1024 points per facial region yielded the highest efficacy, as evidenced by the data highlighted in Table IV, surpassing variations due to individual differences.

However, the attempt to utilize 2048 points revealed limitations in certain facial regions' capacity to fulfill such extensive sampling requirements, resulting in a diminished sample pool and precluding equitable comparisons. Owing to these constraints, a selection of 512 points proved insufficient to encompass all motion areas comprehensively, manifesting in UF1 and UAR scores of 63.87% and 67.97%, respectively, depicted in Figure 7.

Conversely, our method, with the adoption of 1024 points, facilitated enhancements in UF1 and UAR to 68.19% and 74.12%, respectively, achieving optimal recognition performance by maintaining an ideal extent of motion representation, as detailed in Figure 7 (C).

We assessed the efficacy of three point selection strategies we introduced: sorting method, central sampling, and random sampling. The sorting method organizes points based on their motion range, as described in Section III-C, while random sampling involves selecting point clouds randomly within each region, and central sampling selects points nearest to each region's center.

In experiments utilizing 1024 points, both random and central sampling showed comparable outcomes, with central sampling yielding a UF1 of 42.27% and UAR of 46.87%, and random sampling achieving a UF1 of 44.37% and UAR of 49.87%. These findings suggest that point cloud networks prioritize shape over color, despite both central and random sampling maintaining the color channel. The minimal shape differentiation across emotion categories underlines the importance of employing motion range for facial representation.

2) Ablation studies on the structural motion and GCN: Additionally, to contrast our method in motion representation and network architecture with prevailing networks, we examined facial structural motion and global modeling. Common methodologies typically employ optical flow for facial motion representation; conversely, our method amalgamates structural motion details with optical flow and delineates motion information in 3D for each facial region. Further, to demonstrate the comprehensive insights garnered from modeling each part's

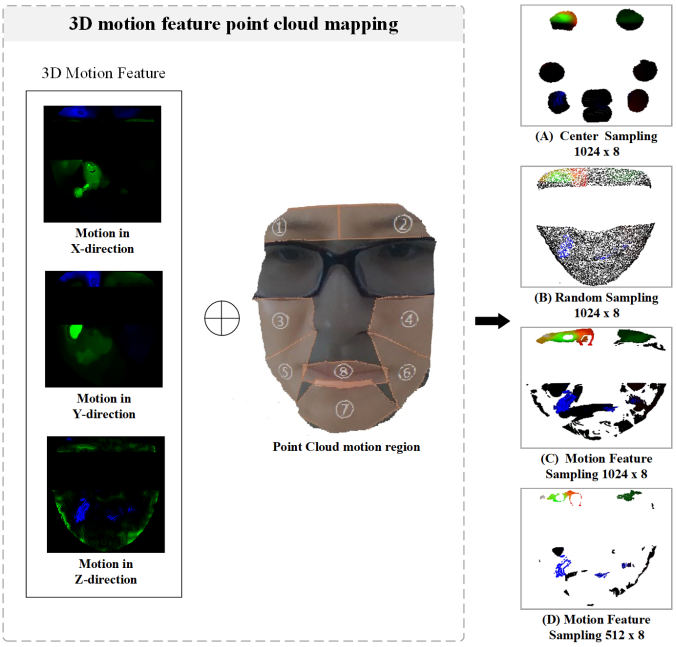


Fig. 7. The choice of sampling methods and the quantity of point clouds significantly influence visualization outcomes. We explored various sampling methods, such as center sampling (A) and random sampling (B), to illustrate the diversity in the geometrical configurations of point clouds associated with different emotions. Nonetheless, representing each region with merely 512 points (D) might lead to an incomplete portrayal of all motion attributes, stemming from an inadequate volume of point clouds. Consequently, to accurately depict the distinctions in shapes across motion categories and preserve facial structure integrity (C), it is crucial to select an optimal number of point clouds.

features via GCN after segmenting facial areas, we showcase experimental outcomes in Table V.

TABLE V
THE IMPACT OF ABLATING FACIAL STRUCTURAL MOTION AND GCN.

Classes	Structural Motion	GCN	UF1(%)	UAR(%)
3			58.70	64.39
3	✓		64.37	68.25
3	✓	✓	**68.19**	**74.12**
4	✓		43.64	51.40
4	✓	✓	**47.64**	**53.66**
7	✓		35.11	41.20
7	✓	✓	**35.64**	**41.59**

We systematically assessed the impact of facial structural motion information and the efficacy of GCN global modeling across three, four, and seven emotion categories. Initially, employing only optical flow for motion depiction without GCN yielded UF1 scores of 58.70% and UAR of 64.93% in the three-emotion category. Introduction of facial structural motion information sans GCN global modeling improved our outcomes to UF1 of 64.37% and UAR of 69.87%. Incorporating both elements led to optimal performance, achieving UF1 of 68.19% and UAR of 74.12%, as highlighted in bold within the Table V, validating the integral role of each component.

To enhance comprehension, we have provided a separate listing of the improvements in Table VI. Our framework exhibits similarities with PointNet++ in terms of feature ex-

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3535569

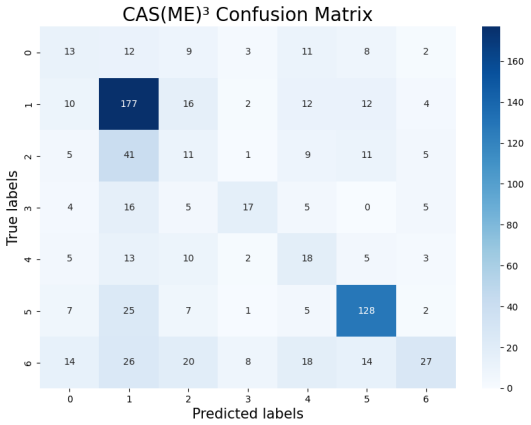


Fig. 8. Confusion matrix for the seven emotion categories. The horizontal axis represents the predicted labels, while the vertical axis represents the true labels. Labels $\{0, 1, 2, 3, 4, 5, 6\}$ correspond to $\{anger, disgust, fear, happiness, sadness, surprise, other\}$, respectively. Darker colors indicate higher quantities.

traction. In the local feature extraction component, we have optimized it by lightweight grouping layer to mitigate the danger of overfitting, resulting in UF1 enhancement of approximately 1%. Regarding the global feature module, we have employed GCN in lieu of MLP and integrated prior knowledge of facial attributes. This enhancement yielded an approximate 4% upsurge in UF1, thereby confirming the rationale behind leveraging facial structural information to augment recognition accuracy.

In the four-emotion category (Table V), the absence of GCN global modeling resulted in UF1 of 43.64% and UAR of 51.40%; for the seven-emotion category, we observed UF1 of 35.11% and UAR of 41.20%. Nonetheless, integrating GCN significantly elevated outcomes, achieving UF1 of 47.64% and UAR of 53.66% in the four-emotion category and UF1 of 35.64% and UAR of 41.59% in the seven-emotion category.

Interestingly, for the seven-emotion category, GCN integration did not markedly alter outcomes. Pursuing potential explanations, we explored modifications in set abstract and sampling methods, network architecture, and sample volumes, yet consistently registered UF1 scores around 35%. This indicates additional constraints within the seven-emotion category beyond network design or sampling strategies.

3) Analyzing MER Challenges based on Confusion Matrix: To explore the challenges in emotion classification accuracy, we first analyzed the confusion matrix for seven emotion categories (see Figure 8), where labels $\{0, 1, 2, 3, 4, 5, 6\}$ represent $\{anger, disgust, fear, happiness, sadness, surprise, other\}$ emotions, respectively. The confusion matrix reveals the following patterns:

First, due to the large sample size and distinct characteristics of "disgust," this category achieves relatively high recognition accuracy. Similarly, the "surprise" emotion has a relatively large sample size, and its unique expressive features make it easier to differentiate from other categories (e.g., positive or negative emotions), leading to a higher recognition rate.

In contrast, negative emotions such as "anger," "disgust," and "sadness" have smaller sample sizes and exhibit minimal inter-category differences, making them more prone to misclassification and thus resulting in lower recognition accuracy. Additionally, the "other" label, despite having a substantial sample size, lacks specific sub-categorization for particular emotions. Its motion characteristics may also resemble those of the first six emotion categories, making it challenging to distinguish based solely on motion features.

Based on these observations, we plan to conduct further experimental research in Section IV-E to explore the relationship between emotion categories and motion features.

TABLE VI
COMPARISON OF LOCAL AND GLOBAL FEATURE IMPROVEMENTS IN NETWORKS.

Method	Local	Global	UF1	UAR
PointNet++ [13]	Pointnet	MLP	63.37	67.87
Ours	Lite-Pointnet	MLP	64.37	68.25
Ours	**Lite-Pointnet**	**GCN**	**68.19**	**74.12**

E. Evaluation of seven emotion classes

In addressing the challenges faced in seven-class classification, we examined the issues through the lenses of label rationality and sample quantity. Initially, concerning label rationality, the direct correspondence between AUs and self-reported emotions may not always be clear-cut [28].

The study by [28] supports this by building a classifier to predict emotions using all available data (without training-testing splits), directly assessing the correlation between AUs and emotions. Their findings show that even with all AUs, predictions are not fully accurate, especially for self-reported emotions.

This suggests a lack of strict one-to-one mapping between AUs and subjective emotional experiences, with $\mathrm{CAS(ME)}^3$ only obtaining a 32.3 UF1 scores for self-reported emotion categories and a perfect 100.0 UF1 scores for objective categories. The meta-study [29] finding also support these results that AUs are better suited for recognizing objective emotional classes but are less effective for subjective self-reported emotions.

This ambiguity is less evident in the three and four emotion categories since they stem from a broader aggregation of the seven emotion categories, where the amalgamation of multiple emotions tends to mask such discrepancies. Therefore, broader categories allow for reduced specificity and enable AUs to capture emotion with less ambiguity, resulting in more balanced classification outcomes. In contrast, the seven emotion categories exhibit higher granularity, increasing the classification difficulty and exposing the limitations of AUs for fine-grained emotion distinctions.

Furthermore, limitations in sample size may predispose the network to over-fitting. To mitigate this, we are exploring the inclusion of MaEs to enrich the dataset and alleviate sample insufficiency.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3535569

1) Introduction of objective class: Drawing inspiration from [28], we endeavored to assess our motion representation within seven emotion categories utilizing the Objective class approach. This approach delineates a classification scheme founded on AUs combinations, harnessing authentic ME motion data alongside feature representation and recognition technologies to mitigate biases inherent in human self-reporting.

Through this methodology, seven emotional categories are established, each mapped to specific AUs. These categories correlate with the six fundamental emotions: happiness, surprise, anger, disgust, sadness, and fear. The seventh category encompasses contempt and additional AUs that the Emotional Facial Action Coding System (EMFACS) [30] does not explicitly link to defined emotions.

It is crucial to note that while these categories are associated with the aforementioned emotions based on prior research, they do not directly correspond to self-reported emotions [30]. Table VII outlines the seven categories alongside their respective assigned AUs.

TABLE VII
EACH CLASS REPRESENTS AUS THAT CAN BE LINKED TO EMOTION.

Class	Action Units
I	AU6,AU12,AU6+AU12,AU6+AU7+AU12,AU7+AU12
II	AU1+AU2,AU5,AU25,AU1+AU2+AU25,AU25+AU26, AU5+AU24
III	A23,AU4,AU4+AU7,AU4+AU5,AU4+AU5+AU7, AU17+AU24,AU4+AU6+AU7,AU4+AU38
IV	AU10,AU9,AU4+AU9,AU4+AU40,AU4+AU5+AU40, AU4+AU7+AU9,AU4 +AU9+AU17,AU4+AU7+AU10, AU4+AU5+AU7+AU9,AU7+AU10
V	AU1,AU15,AU1+AU4,AU6+AU15,AU15+AU17
VI	AU1+AU2+AU4,AU20
VII	Others

2) Experimental result of objective class: In the experiment, we evaluated from two aspects: label categories and sample quantity. As shown in Table VIII.

TABLE VIII
EVALUATION OF EMOTIONAL CATEGORIES AND VALIDATION METHODS.

Train	Test	Categorization	Validation Methods	UF1(%)	UAR(%)
ME	ME	Objective class	LOSO	52.62	62.67
MaE	**ME**	**Objective class**	**Cross-domain**	**67.37**	**66.35**
ME	ME	Emotions	LOSO	35.64	41.59
MaE	ME	Emotions	Cross-domain	27.82	32.64
MaE	ME	Emotions	5-Folds	35.67	36.62

We initiated our investigation by employing objective category verification to assess how sample volume impacts outcomes, utilizing the LOSO for training and evaluation on the ME dataset exclusively. This resulted in UF1 and UAR scores of 52.62% and 62.67%, respectively, revealing significant overfitting and sample distribution imbalances.

However, bypassing any transfer learning approaches and LOSO evaluations, and training solely on MaE datasets while testing on ME samples, we observed UF1 and UAR scores of 67.37% and 66.35% (Bold in Table VIII). This markedly mitigated over-fitting and distribution issues, underscoring the principle that higher-quality samples enhance performance in objective category assessments.

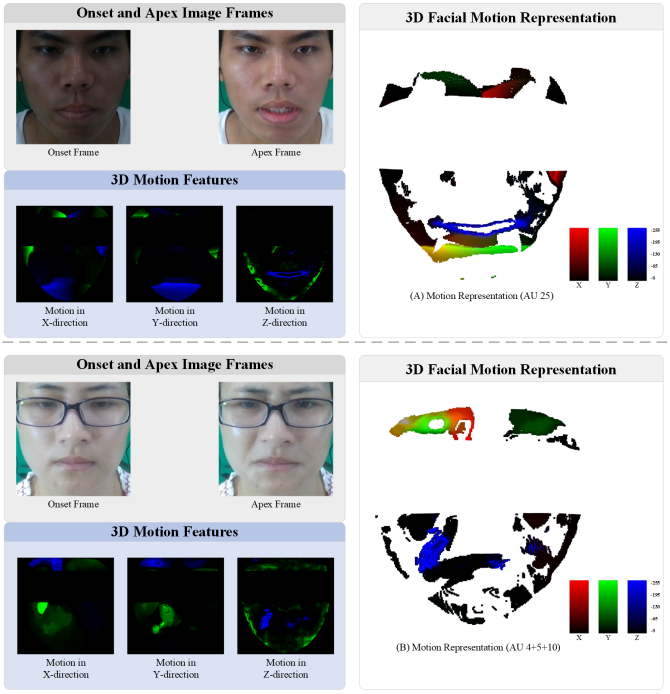


Fig. 9. The interplay between facial structural and texture motions is pivotal. In the image's upper half, variations in lighting conditions render optical flow inadequate for accurately depicting motion's horizontal and directional aspects. However, facial structural details derived from depth maps precisely capture mouth movements. And, in the image's lower half, the amalgamation of pixel value alterations identified through image analysis and structural modifications discerned via depth information provides a comprehensive portrayal of all motion features.

Our strategy, which segments facial motion regions and employs GCN for global relationship modeling, is in sync with the criteria for objective category evaluation.

In subsequent validation within emotion categories, we noted UF1 and UAR scores of 27.82% and 32.64%, reflecting an 8% decrease compared with non-transfer learning, likely due to significant domain discrepancies between MAEs and MEs.

Additional experiments were conducted to probe the domain disparities further. Training on the top 80% of participants across macro and ME samples and evaluating on the remaining 20% of ME participants yielded UF1 and UAR scores of 35.67% and 36.62% through five-fold cross-validation, indicating that the fundamental bottleneck remains unaddressed.

Another validation approach—randomly training on 80% of both macro and ME samples and evaluating on 20% of ME samples—resulted in five-fold cross-validation scores of UF1 43.63% and UAR 43.37%. Given the risk of data leakage [28] with this method, it is advised not to rely on these figures as conclusive evidence.

V. DISCUSSION

A. What the difference between pixel based and 3D structure based motion?

The depth maps have been validated for their efficacy in distinguishing emotional categories as evidenced in [21]. The pixel-based optical flow provides a comprehensive depiction

of motion, while structural motion information offers greater specificity. Integrating texture motion data aligned with spatial structural motion facilitates a nuanced understanding of facial movements (as shown in Figure 9).

This integration has a complementary effect, particularly when RGB images are compromised by lighting conditions; structural motion remains effective in identifying precise movement details (as shown in Figure 9 (A), the blue color represents the structural motion, precisely focused on the mouth). Additionally, these methodologies work synergistically (as shown in Figure 9 (B)).

For example, optical flow precisely captures linear movements of the eyebrows, whereas adjustments in the corners of the mouth are more accurately depicted through structural motion analysis. This dual approach provides a robust representation of facial dynamics.

B. Does the global model must be GCN?

The GCN models the interrelations among these facial regions, facilitating accurate emotional category classification. However, we believe that there exist multiple methods for global modeling.

Initially, PointNet++ [13] employed MLP to model globally after extracting features from point clouds. However, MLP tends to process input features independently, limiting its capacity to integrate complex local structural information. Research has extensively documented that the Transformer model functions as a graph neural network [31], [32]. Consequently, we explored using the Transformer for global modeling.

Specifically, we evaluated the performance of both the traditional Transformer [33] and the PT-V1 Transformer layer [27] in modeling inter-facial region relationships. In the seven emotional classes MER tasks, we assessed the impact of MaE migration.

The UF1 scores for the traditional Transformer and the PT-V1 Transformer layer were approximately 35.17% and 35.32% respectively, showing no significant difference from the scores obtained using GCN and MLP. In the absence of macro-expression migration, the UF1 score decreased to 27.41% and 27.26%, underscoring the severe over-fitting due to network complexity and differences in sample size.

Consequently, we selected GCN as our preferred method for global modeling. Nonetheless, we believe additional methods exist that could effectively describe relationships between regional motions.

VI. Limitation

Despite the undeniable advancements, current representation methods and networks exhibit limitations.

Firstly, while 3D structural information offers a comprehensive framework for capturing facial dynamics, its applicability is constrained by the absence of 3D structures in some datasets, thus limiting the breadth of applicable samples.

Secondly, utilizing prior knowledge of facial muscle distribution enhances facial expression recognition accuracy but encounters challenges due to variations in anatomical accuracy and individual differences, which may affect the precision of facial area segmentation.

Thirdly, optical flow technology, despite its ability to capture facial movements comprehensively, faces accuracy and robustness issues in rapid movements or complex backgrounds, necessitating advanced preprocessing and alignment methods for effective information processing.

Furthermore, in our method, we excluded point clouds in the eye and nose regions. We acknowledge the critical role of eye movements in ME analysis, as they provide essential cues for specific emotions, such as surprise or intense focus.

In this study, we concentrated on key facial regions, including the eyebrows, mouth, and chin, as these areas exhibit stable and prominent emotion-related motions. Eye movements were excluded primarily to minimize interference from external factors such as glasses or lighting reflections. However, we recognize that excluding eye dynamics may impact the accuracy of detecting certain MEs that rely heavily on eye-specific movements, particularly those involving eye-widening.

Moreover, in MER, new methods grapple with inherent challenges: the scarcity of samples necessitates an exploration of domain differences between macro and MEs for transfer learning; and the subtle emotional labeling of MEs complicates the analysis of their relationship with emotions, requiring detailed motion annotation and emotion correspondence studies.

VII. Conclusion and Future Work

This study introduces a novel method for representing facial motion features that merges spatial structure with 3D motion attributes of the face. By meticulously segmenting facial regions and efficiently extracting motion features, this approach significantly enhances the accuracy and robustness of MER.

The method leverages depth information and incorporates both GCN and point cloud technology, facilitating a detailed analysis of the dynamic shifts and interactions among facial components, thus enabling the precise identification of subtle facial expressions.

Despite the considerable advancements in facial motion feature representation and MER through the integration of GCN, future endeavors should concentrate on broadening and diversifying datasets, refining recognition across individual variances and complex settings, and delving into advanced cross-modal information fusion methods.

Additionally, we plan to incorporate a specialized eye movement network with RGB imaging to capture optical flow and motion data, addressing current limitations by enabling the identification of eye-related expressions and movements. This approach will improve the detection of subtle eye dynamics, leading to a more balanced and comprehensive analysis of MEs across all facial regions.

Furthermore, while our framework is initially designed for depth map-derived point clouds, it inherently adapts to real-scanned facial point clouds characterized by unstructured geometry and sparse textures. Although such data pose challenges in inter-frame alignment and noise robustness, lightweight preprocessing with adaptive denoising can preserve geometric fidelity without compromising subtle motion

features. The synergy of PointNet++'s multi-scale aggregation and graph convolutions further exploits the inherent geometric consistency of real scans to model structural dynamics beyond texture dependency.This dual adaptability positions our method as a unified framework for both depth map and real-world 3D facial analysis, which has critical potential for practical deployment of micro-expression recognition systems.

Finally, enhancing real-time performance and resource efficiency is essential for mobile and real-time applications, paving the way for practical deployment in diverse settings.

ACKNOWLEDGMENTS

This work was supported partly by the National Natural Science Foundation of China (Grant No. 62173045), partly by the BUPT Excellent Ph.D. Students Foundation (Grant No. CX2023115).

REFERENCES

[1] J. P. Lee, H. Jang, Y. Jang, H. Song, S. Lee, P. S. Lee, and J. Kim, "Encoding of multi-modal emotional information via personalized skin-integrated wireless facial interface," *Nature Communications*, vol. 15, no. 1, p. 530, 2024.

[2] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[3] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[4] X.-b. Shen, Q. Wu, and X.-l. Fu, "Effects of the duration of expressions on the recognition of microexpressions," *Journal of Zhejiang University Science B*, vol. 13, pp. 221–230, 2012.

[5] P. Ekman, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205–221, 2003.

[6] B. Allaert, I. M. Bilasco, and C. Djeraba, "Micro and Macro facial expression recognition using advanced local motion patterns," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 147–158, 2019.

[7] P. Lewinski, M. L. Fransen, and E. S. Tan, "Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli," *Journal of Neuroscience, Psychology, and Economics*, vol. 7, no. 1, p. 1, 2014.

[8] R. Zhang, N. He, S. Liu, Y. Wu, K. Yan, Y. He, and K. Lu, "Your heart rate betrays you: multimodal learning with spatio-temporal fusion networks for micro-expression recognition," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 4, pp. 553–566, 2022.

[9] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5826–5846, 2022.

[10] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "AU-assisted graph attention convolutional network for micro-expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2871–2880.

[11] L. Lei, J. Li, T. Chen, and 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*, 2020, pp. 2237–2245.

[12] Z. Dong, G. Wang, S. Lu, L. Dai, S. Huang, and Y. Liu, "Intentional-deception detection based on facial muscle movements in an interactive social context," *Pattern Recognition Letters*, vol. 164, pp. 30–39, 2022.

[13] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] N. H. Ren Zhang, "A review of micro-expression recognition research," *Journal of Computer Engineering & Applications*, vol. 57, no. 1, pp. 1–11, 2021.

[15] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2015.

[16] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. IEEE, 2019, pp. 1–4.

[17] R. Zhang, N. He, Y. Wu, Y. He, and K. Yan, "To balance: Balanced micro-expression recognition," *Multimedia Systems*, vol. 28, pp. 1–11, 2022.

[18] X. Fan, X. Chen, M. Jiang, A. R. Shahid, and H. Yan, "SelfME: Self-supervised motion learning for micro-expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 834–13 843.

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

[20] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *Journal of Imaging*, vol. 4, no. 10, p. 119, 2018.

[21] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, and X. Fu, "$CAS(ME)^3$: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[22] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," *Pattern Recognition*, vol. 122, p. 108275, 2022.

[23] S.-T. Liong, Y. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "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, 2019, pp. 1–5.

[24] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 8590–8605, 2020.

[25] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6411–6420.

[26] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.

[27] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point Transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 259–16 268.

[28] T. Varanka, Y. Li, W. Peng, and G. Zhao, "Data leakage and evaluation issues in micro-expression analysis," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 186–197, 2023.

[29] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

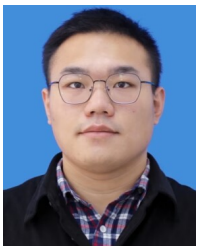
[30] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.

[31] C. Joshi, "Transformers are graph neural networks," *The Gradient*, vol. 12, p. 17, 2020.

[32] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu, "Do transformers really perform badly for graph representation?" *Advances in neural information processing systems*, vol. 34, pp. 28 877–28 888, 2021.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3535569



Ren Zhang received MEng degree from Beijing Union University, China, in 2019. He is currently a Ph.D. candidate with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include affective computing, facial, and micro-expression analysis.

Chao Qi received the Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2023. He works at the School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications. His research interests include deep learning, pattern recognition, and 3D vision.

Zehao Wang received Bachelor degree in computer science and technology from Beijing University of Posts and Telecommunications, China, in 2019. He is currently a master student at the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. His research interests include computer vision, image processing, and deep learning.

Huaping Liu (Senior Member, IEEE) Received the Ph.D. degree in Computer Science and Technology from Tsinghua University, Beijing, China, in 2004. He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include intelligent robot perception, learning, and control.

Jianqin Yin (Member, IEEE) received the Ph.D. degree from Shandong University, Jinan, China, in 2013. She currently is a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include service robot, pattern recognition, machine learning and image processing.

Yonghao Dang received BEng degree in computer science and technology from the University of Jinan, Jinan, China, in 2018. He is currently a Ph.D. candidate with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include computer vision, image processing, and deep learning.

Zhicheng Zhang received the BEng degree in information engineering, the MEng degree in pattern recognition, and the PhD degree in control theory and engineering from Jilin University, China in 2005, 2007, and 2011, respectively. He is currently an associate professor at the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include signal processing, artificial intelligence, and computational intelligence.