Looking into Gait for Perceiving Emotions via Bilateral Posture and Movement Graph Convolutional Networks

Yingjie Zhai*, Guoli Jia*, Yu-Kun Lai, Jing Zhang, Jufeng Yang, and Dacheng Tao, *Fellow, IEEE*

***Abstract*—Emotions can be perceived from a person's gait, i.e., their walking style. Existing methods on gait emotion recognition mainly leverage the posture information as input, but ignore the body movement, which contains complementary information for recognizing emotions evoked in the gait. In this paper, we propose a Bilateral Posture and Movement Graph Convolutional Network (BPM-GCN) that consists of two parallel streams, namely posture stream and movement stream, to recognize emotions from two views. The posture stream aims to explicitly analyse the emotional state of the person. Specifically, we design a novel regression constraint based on the hand-engineered features to distill the prior affective knowledge into the network and boost the representation learning. The movement stream is designed to describe the intensity of the emotion, which is an implicitly cue for recognizing emotions. To achieve this goal, we employ a higher-order velocity-acceleration pair to construct graphs, in which the informative movement features are utilized. Besides, we design a PM-Interacted feature fusion mechanism to adaptively integrate the features from the two streams. Therefore, the two streams collaboratively contribute to the performance from two complementary views. Extensive experiments on the largest benchmark dataset Emotion-Gait show that BPM-GCN performs favorably against the state-of-the-art approaches (with at least 4.59% performance improvement). The source code is released on *https://github.com/zyjwuyan/egait_journal*.**

***Index Terms*—Emotion identification, gait, bilateral posture and movement graph convolutional network, affective constraint.**

I. Introduction

EMOTIONS are biological states associated with feelings, thoughts, behavioral responses, and a degree of pleasure or displeasure. They play an essential role in our lives, witnessing our experiences and reflecting our state of mind about the world and other people [1], [2]. Due to the importance of perceived emotion in daily life, automatic emotion recognition has received increasing attention in many fields, such as human-robot interaction [3]–[5], behavior prediction [6]–[8], and affective computing [9], [10].

In academia, current research mainly falls into leveraging verbal cues such as text [11] and speech [12], as well as non-verbal cues [13], *e.g.*, affective features [14], facial cues [1], [15], and body expressions [16]. There are extensive works using non-verbal cues for understanding the emotions of individuals [17]–[19], in which facial expressions are most commonly utilized [15], [20], due to the massive amount of available data and applications. Nevertheless, in some cases, we can hardly get clear faces, such as on a dark night without sufficient lighting and sometimes facial expressions are not reliable when someone sneers, behaves with mock or makes a face.

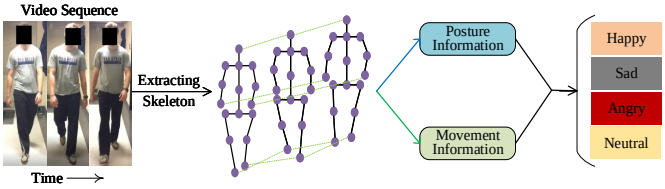

Fig. 1: Motivation of the proposed method. Humans can express their emotions by the posture and movement of multiple parts of their bodies. Based on the skeleton extracted from the gait video, we design a bilateral posture and movement graph convolutional network to identify human emotions.

According to psychology literature [21]–[23], participants can also identify the emotions of the subject by observing their posture. For example, arm swinging is related to positive emotion, collapsed upper body is related to negative emotion [24]. Besides, body movement (*e.g.*, walking speed) also plays an important role in the perception of emotions. [25] demonstrates that the movement is particularly associated with the extent of the emotion. Emotions such as anger and excitement have high arousal, which have closer relation with rapid movements than low arousal emotions, such as sadness and contentment [16]. Therefore, in this paper, we aim to capture human emotions from another perspective of non-verbal cues, *i.e.*, **gait**, which is defined as *a person's walking style* [26]–[28]. Gait has been proven to be significant to human identity recognition [29] and emotion classification [30].

There are some methods that identify human emotions from gait. Early algorithms mainly focus on extracting hand-crafted features such as joint angles [31], [32], covariance descriptors [33], kinematic features [34], and feeding them into classifiers to predict emotions. Recently, researchers began to extract deep features using Long Short-Term Memory networks (LSTMs) [35], [36] and Convolutional Neural Networks (CNNs) [37]–[39]. However, the current methods have the following limitations: First, most of them [30], [40], [41]

Yingjie Zhai and Guoli Jia contribute equally to this work.
Yingjie Zhai, Guoli Jia, and Jufeng Yang are with College of Computer Science, Nankai University. (Email: zhaiyingjier@163.com, exped1230@gmail.com, yangjufeng@nankai.edu.cn)
Yu-Kun Lai is with School of Computer Science and Informatics, Cardiff University. (Email: LaiY4@cardiff.ac.uk)
Jing Zhang and Dacheng Tao are with School of Computer Science, Faculty of Engineering, The University of Sydney (Email: jing.zhang1@sydney.edu.au, dacheng.tao@gmail.com)

ignore the hidden movement information (*e.g*., velocity and acceleration of joints) that is related to human emotions [16], [32] and can help to learn discriminative classifiers. Second, the hand-crafted affective features are directly concatenated with the deep features [30], [35], [42], which makes the deep model hard to understand and utilize these affective features. Additionally, although current action recognition methods can be adapted to identify emotions, they neglect the affective cues that are instructive to bridge the gap between posture and emotion.

To address the above concerns, inspired by the fact that both posture and movement of body parts can express human emotions [43], we propose a Bilateral Posture and Movement Graph Convolutional Network (BPM-GCN) to imitate human emotion perception (see Fig. 1). BPM-GCN maintains two graph convolutional networks simultaneously, namely the posture stream and the movement stream. The posture stream aims to mine emotion information from the state of a person's posture, such as joint positions and the angles between joints. The movement stream utilizes higher-order representations (*i.e*., velocity and acceleration), with which the hidden movement information closely related to human emotions. The two streams address the important problem regarding where we can find more emotional cues from gait. Besides, we also design an affective constraint based on the hand-engineered features in the posture stream, which can distill the prior emotional knowledge into the network and well bridge the gap between posture and emotion. To integrate the features from the two streams, we design a PM(Posture and Movement)-Interacted feature fusion mechanism, which can make one stream obtain externally useful information from the other view, and thus benefits the performance further. To verify the superiority of our model, extensive experiments are performed on the benchmark dataset Emotion-Gait [30], where BPM-GCN achieves State-Of-The-Art (SOTA) performance.

The contribution of this paper is fourfold:

- We propose a novel Bilateral Posture and Movement Graph Convolutional Network, called BPM-GCN, for emotion recognition from gait. BPM-GCN exploits the affective information of gait from two important views, *i.e*., posture and movement, which perceive both state and extent of human emotions.
- We provide an insight into mining useful information in the posture stream by introducing an affective constraint. Such an affective constraint helps to distill the prior affective knowledge into the network, which can guide the training of the network and effectively bridge the gap between posture and human emotions.
- We design a PM-Interacted feature fusion mechanism consisting of a temporal and a spatial attention operation, which can reduce the side-effects from the modality discrepancy and fuse the features from two streams adaptively.
- On the widely used Emotion-Gait benchmark, the proposed BPM-GCN significantly outperforms the SOTA methods with at least 4.59% accuracy improvement, which demonstrates the superiority of our algorithm.

II. Related Work

In this section, we review prior works that are related to this paper. Specifically, we group them into two categories: 1) Gait Emotion Recognition (§ II-A), and 2) Skeleton-based Action Recognition (§ II-B).

A. Gait Emotion Recognition

Affective computing, which involves computer science, psychology, and cognitive science, has been studied for over two decades [14], [44]–[46]. Most of the existing works focus on predicting human emotions from face images, speech, and text [47]–[51]. However, some researchers argue that body posture also plays an important role in human feeling expression [52]–[54]. For example, experiments in [55] show that participants can recognize basic emotions with high accuracy from point-light arm movements of two actors who are instructed to perform drinking and knocking movements with ten different affects. Boone *et al*. [56] find that people with joy tend to open forearms while they will tighten their bodies to be self-protective when they are frightened. Gross *et al*. [57] conduct kinematic analysis on motion-captured data collected from 16 individuals who are asked to walk while experiencing five emotions (joy, contentment, anger, sadness, and neutral), and the experiments indicate that the fastest walking velocity is for joy and anger, and the slowest is for sadness. Besides, when participants are in sadness, they flex the neck and thoracic, but they extend the trunk or depress the shoulder when they are joyful.

Since a person's walking style can reveal certain emotions [23], [52], [53], algorithms have been developed to automatically identify emotions from gait, which is complementary to emotion recognition from other modalities. Most previous approaches are based on hand-crafted features. Karg *et al*. [32] extract efficient features with respect to affects from captured gait data, and then use Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and general discriminant analysis to either reduce the temporal dimension or select relevant features for classification. Blind source separation techniques such as PCA can hardly process complex joint angle trajectories of body movements effectively. To address this issue, Omlor and Giese [58] develop a new non-linear source separation technique. Rather than extracting common features such as stance phase, frequency, and footstep length from gait, Venture *et al*. [31] classify five emotions based on joint angles. Further, Li *et al*. [59] perform the discrete Fourier transform and statistical methods on gait collected from Microsoft Kinect to identify emotions. Chiu *et al*. [60] compare the performance of several supervised models, *e.g*., support vector machines, multilayer perceptron, naive Bayes, and decision trees, trained with features from video frames.

Recently, due to the powerful representation ability of deep networks [61]–[63], researchers started to develop deep models and feed 3D joint positions into them to identify emotions. For example, Randhavane *et al*. [35] concatenate deep features in LSTM and hand-crafted features computed from gait, and then employ a random forest classifier to predict emotions. Following this work, Bhattacharya *et al*. [30] present a new

dataset, called Emotion-Gait, that contains 2,177 human gait and external synthetic gait with annotations of four emotional categories. They also propose a Spatial Temporal graph convolution-based network for Emotion Perceiving (STEP). On the same dataset, Narayanan *et al*. [40] propose a multi-view skeleton model to identify emotions for socially aware robot navigation among pedestrians.

Different from the above-mentioned methods, we propose to recognize gait emotions not only from 3D joint positions but also from the hidden movement information which also carries important emotional cues. To this end, we devise a bilateral posture and movement graph convolutional network to predict emotions, taking both important perspectives into consideration. The two streams can collaboratively contribute to the emotion recognition from gait.

B. Skeleton-based Action Recognition

We aim to identify emotions from gait represented by 3D joint coordinates, which is related to the skeleton-based action recognition task [64] to some extent. Here we briefly review some works that have made great contributions to this field in recent years.

Recurrent Neural Networks (RNNs) are widely used in skeleton-based action recognition. For example, Du *et al*. [65] propose a Hierarchical Recurrent Neural Network (H-RNN), which divides joints into multiple parts, feeds them to corresponding subnets, and further fuses the features extracted from subnets in a hierarchical way. In [66], Wang *et al*. introduce a two-stream RNN to learn both spatial and temporal representations, using rotation and scaling transformations to augment data. In addition to RNNs, as in other computer vision tasks [67], [68], CNNs have also shown great potential for action recognition. Du *et al*. [69] propose to concatenate joint coordinates in a chronological order to convert a skeleton sequence into a matrix, and then feed it into a CNN to extract representations. Kim *et al*. [70] utilize Temporal Convolutional Neural networks (TCNs) for skeleton-based action recognition. They propose a novel model, called Res-TCN, that can explicitly learn spatial-temporal representations.

Recently, skeleton-based action recognition algorithms have been dominated by Graph Convolutional Networks (GCNs). This is intuitive because a human skeleton can be regarded as a graph with joints as nodes. Following this insight, Yan *et al*. [41] propose a Spatial-Temporal Graph Convolutional Network (ST-GCN), which automatically extracts spatial and temporal features from the coordinates of joints. Song *et al*. [71] present a multi-stream GCN where each single stream is forced to explore discriminative features from inactivated joints, which effectively alleviates performance deterioration caused by noisy skeletons. Considering that previous models can only capture local dependencies among joints, Li *et al*. [72] propose an Actional-Structural Graph Convolutional Network (AS-GCN) to learn latent dependencies that are specific to actions. Additionally, to incorporate both the joint and bone information and utilize the relationship between them, Shi *et al*. [73] design a directed graph neural network to predict actions using the information of joints, bones, and their relations.

Note that as the above methods, our BPM-GCN is also based on 3D joint coordinates. However, directly applying these skeleton-based action recognition methods to emotion identification cannot obtain favorable results (see Tab. II) because they neglect the affective cues that are important for perceiving emotions. It is crucial to design a specific architecture for this abstract and challenging task.

III. Methodology

In this section, we first make a definition and present a brief overview of the proposed method in § III-A. Then, we elaborate the proposed bilateral posture and movement GCN (*BPM-GCN*) in § III-B, including the details of the movement stream and posture stream. Besides, we introduce the PM-Interacted feature fusion mechanism, which can effectively fuse the features from both streams, in § III-C.

A. Definition and Overview

A skeleton-based gait is represented by a sequence of 3D joint coordinates. It can be denoted as $C \times T \times N$, where C is the attribute dimension of a joint (*e.g*., if we represent each joint with 3D coordinates, it is 3). T is the length of the temporal sequence, and N denotes the number of joints in a single frame. In this work, we use a spatial-temporal graph [41] to represent the relations of multiple joints and frames. Let graph $G = (\mathcal{V}, \mathcal{E})$ represent a gait, where $\mathcal{V}$ and $\mathcal{E}$ denote the vertices and the edges, respectively. In detail, the vertices $\mathcal{V} = \{v_i^t; i = 1, 2, ..., N, t = 1, 2, ..., T\}$, where v_i^t denotes the i^{th} joint of the t^{th} frame. In our work, $N = 16$ is the number of joints in a frame and $T = 48$ is the number of frames in a gait. Two kinds of edges are considered in the spatial-temporal graph. On the one hand, as shown in Fig. 3(a), within the same frame, each joint is connected with neighboring joints following the connectivity of the human body structure. On the other hand, each joint in one frame is also connected with the same joint in the adjacent frames (see Fig. 3(b)). In other words, the edge set $\mathcal{E}$ contains two subsets. The first set of spatial edges is defined by $\mathcal{E}_s = \{v_i^t v_j^t; (i, j) \in \mathcal{H}\}$, where $\mathcal{H}$ is the connected joint pair set for the human skeleton map. This set represents the intra-frame position connections of the joints. The second set of temporal edges is denoted as $\mathcal{E}_\mathcal{T} = \{v_i^t v_i^{t+1}; i = 1, 2, ..., N,\ t = 1, 2, ..., T-1\}$, which represents the inter-frame temporal connections of joints.

As shown in Fig. 2, the proposed Bilateral Posture and Movement Graph Convolutional Network (BPM-GCN) consists of two streams. One is the posture stream that aims to extract emotional information from the person's posture (*i.e*., joint position, the angles between joints, the distance between joints, and body area), and the other is the movement stream that leverages the velocity and acceleration to model the person's emotions. Both streams imitate human perception from two important views and collaboratively contribute to the performance improvement. Next, we will briefly introduce the basic element (*i.e*., spatial-temporal graph convolution) used in our model, and then introduce the proposed framework (BPM-GCN) in detail.

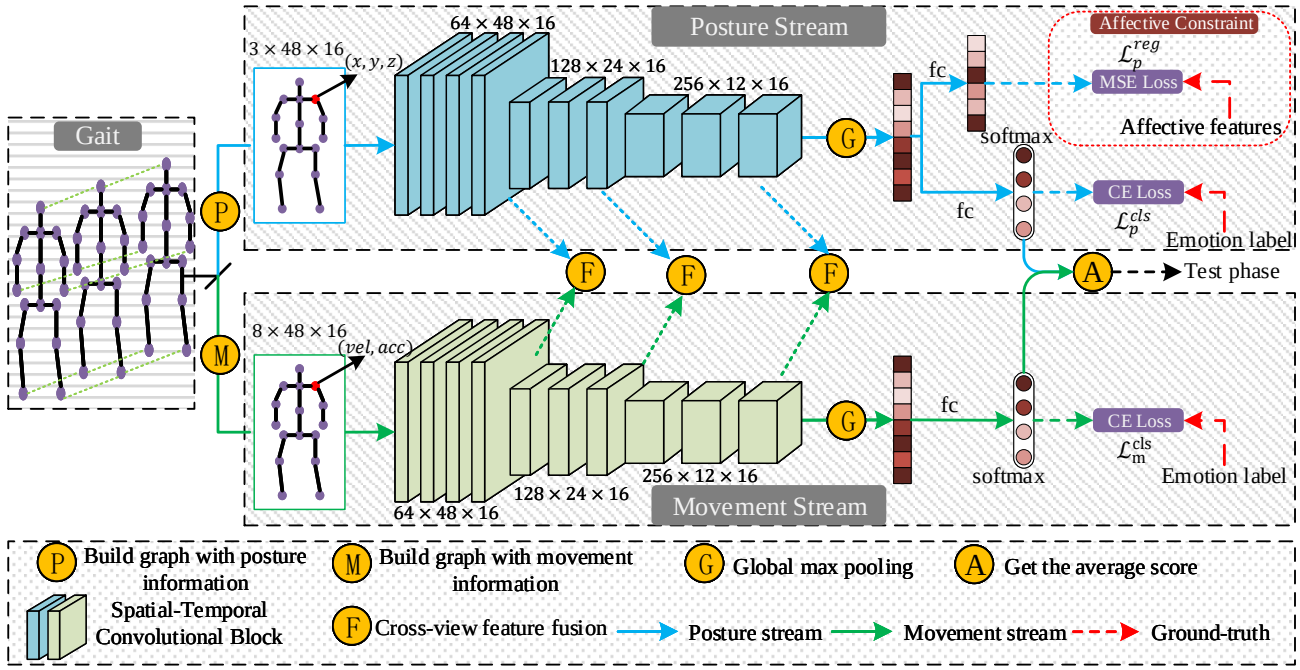

Fig. 2: Pipeline of the proposed algorithm (BPM-GCN), which contains two streams. The posture stream (**top**) constructs graphs based on the joint coordinates, and is optimized by both the classification loss ($\mathcal{L}_p^{cls}$) and the newly designed affective constraint ($\mathcal{L}_p^{reg}$). Meanwhile, the movement stream (**bottom**) constructs graphs using the velocity-acceleration pairs as nodes, and is optimized by the classification loss ($\mathcal{L}_m^{cls}$). Note that (x, y, z) is the 3D coordinates of a joint, and (vel, acc) represents the velocity and acceleration.

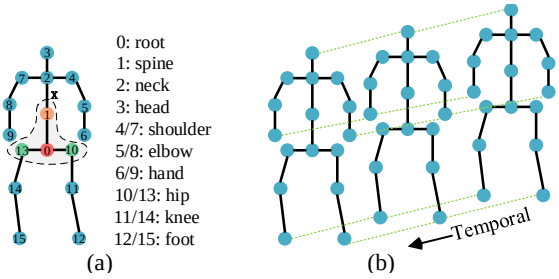

Fig. 3: (a) Illustration of the sampling area for the spatial-temporal graph, where 'x' denotes the gravity center of the body. The area enclosed by the curve is the sampling area, where the red node is the target node, the green nodes represent the subset far from the gravity center, and the orange node denotes the subset close to the gravity center. (b) Illustration of the spatial-temporal sequence for a gait.

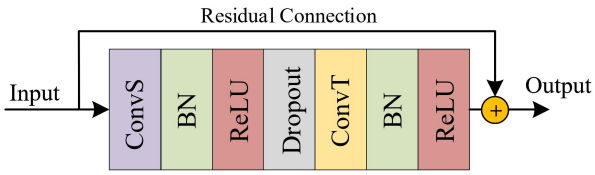

Fig. 4: Illustration of the spatial-temporal convolutional block used in our model. 'ConvS' and 'ConvT' represent the spatial convolution and temporal convolution respectively.

B. Bilateral Posture and Movement GCN

• **Posture Stream and Affective Constraint.** The posture stream takes the joint coordinates based graph as input and outputs the emotion prediction. We divide the posture stream into two branches, *i.e.*, classification branch and regression branch, where they share weights except for the last fully-connected layers. The first branch outputs the classification prediction for the emotions and the second branch distills the knowledge from the posture-based hand-crafted affective features via a regression constraint. Such an affective constraint can bridge the gap between the posture and emotion, and thus benefits the stream to learn more discriminative representations.

Spatial relations of joints is crucial for analyse emotions based on gait patterns [32]. Therefore, we utilize three features including angles, distances, and areas as posture-based hand-crafted affective constraint. First, an **angle feature** is calculated based on three joints. For example, we calculate the angle between the left and right shoulders (see nodes 4 and 7 in Fig. 3(a)) through the person's neck (see node 2 in the same figure). Then, each **distance feature** involves two joints and is calculated as the spatial (Euclidean) distance of the joint locations. For the **area features**, each one involves three joints, and we compute the area of the triangle formed by those joints. To sum up, we can extract 31 posture-based features (*i.e.*, 14 angle, 9 distance, and 8 area features; details can be found in Tab. I).

Let $(G_1, \mathbf{y_1}, \mathbf{y_2})$ be the triplet of a training sample for the posture stream, where $G_1 = (\mathcal{V}_1, \mathcal{E})$ denotes the input gait. $\mathbf{y_1} = (y_1^1, y_1^2, ..., y_1^{C_n})$ is a one-hot vector that represents the emotion label for this sample, in which $y_1^j = 1$ indicates the label is j, and C_n denotes the total number of classes. $\mathbf{y_2} = (y_2^1, y_2^2, ..., y_2^{C_f})$ represents the posture-based affective features, where $C_f = 31$ denotes the dimension of the affective features. Each node $v_{1i}^t \in \mathcal{V}_1$ of the gait can be represented

TABLE I: Details of affective features used in our method. Each cell in the table represents a joint combination for calculating a feature. For posture-based affective features, each angle is computed among three joints with the second as the vertex, a distance is calculated between two joints, and an area represents the area of the triangle formed by three joints. For movement features, we compute velocity and acceleration for each joint. The position of each node can be referred to in Fig. 3(a). Besides, the up point in the table refers to any point vertically above the root node. In total, we can extract 31 posture-based features (*i.e.*, 14 angle, 9 distance, and 8 area features) for each frame and 8 movement features for each joint.

Features		Joint Combination	Joint Combination
Posture Features	Angle	left shoulder, neck, right shoulder	neck, right shoulder, left shoulder
		right shoulder, left shoulder, neck	neck, right shoulder, right elbow
		neck, left shoulder, left elbow	right shoulder, right elbow, right hand
		left shoulder, left elbow, left hand	neck, spine, root
		head, neck, spine	root, right hip, right knee
		root, left hip, left knee	right hip, right knee, right foot
		left hip, left knee, left foot	head, root, up point
	Distance	right hand, root	left hand, root
		right hand, right shoulder	left hand, left shoulder
		right elbow, root	left elbow, root
		right foot, root	left foot, root
		right foot, left foot	
	Area	left hand, neck, right hand	left shoulder, neck, right shoulder
		left hand, root, right hand	left elbow, neck, right elbow
		left foot, neck, right foot	left hip, neck, right hip
		left foot, root, right foot	left knee, neck, right knee
Movement Features	Velocity	the four-dimensional vector $(v_x, v_y, v_z, \|v\|)$ for each joint, where the first three are the velocity components in x, y, and z directions, and the last is its overall magnitude.	
	Acceleration	the four-dimensional vector $(a_x, a_y, a_z, \|a\|)$ for each joint, where the first three are the acceleration components in x, y, and z directions, and the last is its overall magnitude.	

by its 3D joint coordinates (c_x, c_y, c_z) that denote the spatial position of the joint. Let $\{a_1^1, a_1^2, ..., a_1^{C_n}\}$ be the output values of the last fully-connected layer of the first (classification) branch, which are normalized by a softmax function:

$$p_1^j = \frac{e^{a_1^j}}{\sum_{k=1}^{C_n} e^{a_1^k}}, j \in \{1, 2, ..., C_n\}. \tag{1}$$

Then the cross-entropy loss for the label prediction of the posture stream can be defined as:

$$\mathcal{L}_p^{cls} = -\sum_{j=1}^{C_n} y_1^j \ln p_1^j. \tag{2}$$

Let $\{b_1^1, b_1^2, ..., b_1^{C_f}\}$ be the output values of the last fully-connected layer of the second (regression) branch. Then the regression loss (mean squared error) of the second branch regarding the prediction of posture-based affective features can be depicted as:

$$\mathcal{L}_p^{reg} = \frac{1}{C_f} \sum_{k=1}^{C_f} (b_1^k - y_2^k)^2. \tag{3}$$

L_p^{reg} serves as the affective constraint. Furthermore, the total loss for the posture stream is formulated as:

$$\mathcal{L}_p = \mathcal{L}_p^{cls} + \mathcal{L}_p^{reg}. \tag{4}$$

• **Movement Stream.** The movement stream predicts the human emotions from the movement attributes of human joints. It is based on the fact that the movement attributes have closely relation with the intensity of emotion [74], which is a implicit cue for predicting human emotions. Note the movement stream contains only one classification branch and has no regression branch with it. This is because the posture-based affective features have been well-defined in the literature and can be easily computed based on joint positions, while it is a challenging task to build a similar constraint with velocity and acceleration as input. We are considering investigating such a constraint in the future.

For every joint, we calculate velocity in each direction by performing a simple subtraction between the coordinates of the current frame and the previous one. Hence the velocity of each joint is represented by a 4-dimensional vector $(v_x, v_y, v_z, |\mathbf{v}|)$ where the first three are the velocity components in x, y, and z directions, and the last is its overall magnitude. And we compute acceleration in the same way by performing a subtraction between the velocities of two neighboring frames (*i.e.*, as the second-order difference). Note that we do not use the higher-order information (*e.g.*, jerk) here, since it is far from the human perception.

Let $(G_2, \mathbf{y_1})$ be the training pair of a sample for the movement stream, where $G_2 = (\mathcal{V}_2, \mathcal{E})$ denotes the movement-feature-based gait. Each node $v_{2i}^t \in \mathcal{V}_2$ of the gait can be represented by an 8-dimensional movement feature $(m_1, m_2, ..., m_8)$, where the first four dimensions are the velocity attributes and the last four dimensions are the acceleration attributes. Let $\{a_2^1, a_2^2, ..., a_2^{C_n}\}$ denote the output values of the last fully-connected layer of the movement branch, which are then normalized by the softmax function:

$$p_2^j = \frac{e^{a_2^j}}{\sum_{k=1}^{C_n} e^{a_2^k}}, j \in \{1, 2, ..., C_n\}. \tag{5}$$

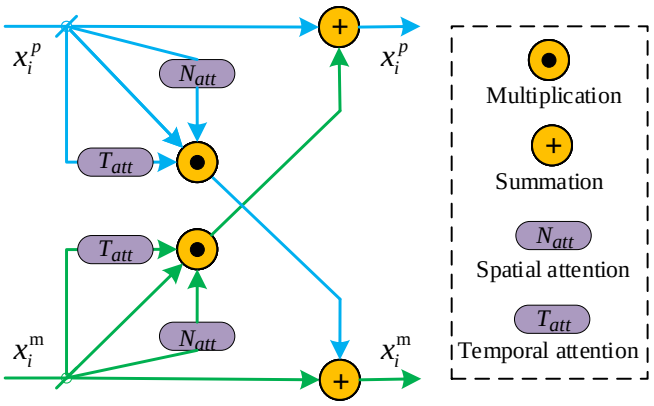

Fig. 5: Illustration of the PM-Interacted feature fusion mechanism. Each feature x_i^p (x_i^m) from the i^{th} stage of the posture (movement) stream is adaptatively fused with the feature x_i^m (x_i^p) from the i^{th} stage of the movement (posture) stream. Blue and green colors represent the posture feature flow and movement feature flow, respectively.

Again, C_n is the total number of classes. The cross-entropy loss function of the movement stream is defined as:

$$\mathcal{L}_m^{cls} = -\sum_{j=1}^{C_n} y_1^j \ln p_2^j. \tag{6}$$

Overall, the proposed BPM-GCN consists of the posture stream and the movement stream. The two streams can be optimized separately or together. When we train two streams together, the total loss is defined as:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_m^{cls}. \tag{7}$$

Empirical study shows that our model is robust to different weighting of loss, thus we directly add the two terms to form the $\mathcal{L}$.

C. BPM-GCN with PM-Interacted Feature Fusion

Although the naive implementation of *BPM-GCN* can achieve good results, the way of combining the two stream is relatively straightforward without considering any fusion at the feature level. To further improve the effectiveness and performance of the proposed two-stream architecture, we further propose a PM-Interacted feature fusion mechanism to make the two streams exchange complementary information. As shown in Fig. 5, the proposed PM-Interacted feature fusion mechanism consists of two main operations, *i.e.*, a temporal attention to make the model focus on the important frames and a spatial attention to make the model attend to the important nodes. Each selected middle-layer posture feature (movement feature) is updated by adding itself with the attended movement feature (posture feature), *i.e.*, making the model fuse the two view features adaptively. As shown in Fig. 2, the backbone of the *BPM-GCN* contains three convolution stages, each stage is with the same feature dimension. The output features of these stages in the two streams are fused using the proposed PM-Interacted feature fusion mechanism. We introduce the PM-Interacted feature fusion mechanism as follows.

The input gait is represented by $C \times T \times N$, where the first dimension C is the attribute axis, the second dimension T is the temporal axis, and the third dimension N denotes the node (joint) axis. Let x_i^p ($i \in \{1,2,3\}$) denote the output feature map of the i^{th} stage in the posture stream and x_i^m ($i \in \{1,2,3\}$) represents the output feature map of the i^{th} stage in the movement stream. The output feature map of the i^{th} stage in the posture stream is updated by being fused with the feature map of the i^{th} stage in the movement stream, which is formulated as:

$$x_i^p = x_i^p + x_i^m * T_{att}(x_i^m) * N_{att}(x_i^m), \tag{8}$$

where T_{att} is the temporal attention along the temporal axis, and N_{att} denotes the spatial attention along the node axis. Similarly, the output feature map of the i^{th} stage in the movement stream is updated by being fused with the feature map of the i^{th} stage in the movement stream, which is defined by:

$$x_i^m = x_i^m + x_i^p * T_{att}(x_i^p) * N_{att}(x_i^p). \tag{9}$$

• **Temporal Attention.** The temporal attention T_{att} aims to emphasize the emotion-relevant information along the temporal axis, which is defined as:

$$T_{att}(\cdot) = Sigmoid(MLP(P_{avg}^T(\cdot))), \tag{10}$$

where P_{avg}^T is the global average pooling along the temporal axis, $MLP(\cdot)$ represents a two-layer perception, and $Sigmoid(\cdot)$ is the sigmoid function that normalizes the value of the features to $[0,1]$.

• **Spatial Attention.** The spatial attention focuses on the important nodes that contribute to the PM-Interacted feature fusion. The operation is denoted by:

$$N_{att}(\cdot) = Sigmoid(MLP(P_{avg}^N(\cdot))), \tag{11}$$

where P_{avg}^N is the global average pooling along the node axis. Other symbols are the same as the temporal attention operation.

Such an attention-based PM-Interacted feature fusion mechanism has two aspects of advantages: 1) it enables the model to focus on the important frames in the sequence and important nodes in the human skeleton for perceiving emotions; 2) it can adaptatively reduce the modality difference of the two-stream features and make the fusion more effectively, which benefits the learning of both two streams.

IV. Experiments

In this section, we first introduce the implementation details in § IV-A. Then, we introduce a standard gait emotion perception benchmark and provide quantitative evaluation and qualitative analysis of the proposed *BPM-GCN* on this benchmark in § IV-B. Furthermore, in § IV-C, we conduct ablation studies to further investigate the effectiveness of our *BPM-GCN* model. Finally, we make a discussion about the overall performance improvement of the proposed model in § IV-D.

TABLE II: Performance comparisons of different models (ten emotion-specific methods and two action recognition methods (denoted by $*$)) on the Emotion-Gait dataset. Ours (Ours$^{\divideontimes}$) represents that we train the model without (with) the PM-Interacted feature fusion mechanism.

Methods	Venture *et al.* [31]	Karg *et al.* [32]	Daoudi *et al.* [33]	Wang *et al.* [34]	Crenn *et al.* [75]	LSTM [35]	ProxEmo [40]	STEP [30]	TNTC [76]	ST-GCN* [41]	2s-AGCN* [77]	BPM-GCN Ours	BPM-GCN$^{\divideontimes}$ Ours$^{\divideontimes}$
Accuracy (%)	30.83	39.58	42.52	53.73	66.22	75.10	82.40	83.15	85.97	65.62	84.40	88.99	**90.37**

TABLE III: To evaluate the robustness of our method, we compare performance of BPM-GCN with SOTA models in various dataset split settings. Except for the 5-fold cross validation, both the two settings are conducted 5 times with different sample separation.

Settings	ProxEmo	STEP	2s-AGCN	BPM-GCN	BPM-GCN$^{\divideontimes}$
8 -1-1	78.53±0.79	80.21±0.82	81.25±0.76	87.36±0.91	88.41±0.82
5-fold	80.01±0.59	80.38±0.65	82.57±0.60	87.39±0.54	88.94±0.50
9-1	82.53±0.63	83.53±0.78	84.53±0.81	88.84±0.63	89.93±0.57

TABLE IV: Performance comparison of state-of-the-art models on multi-class ELMD dataset.

Methods	Sub-class Accuracy (%)				MAP	Accuracy (%)
	Happiness	Sadness	Anger	Neutral		
ProxEmo	89.29	79.63	76.47	78.57	80.99	80.97
STEP	87.07	85.96	74.29	80.49	81.95	82.95
2s-AGCN	94.70	79.11	76.71	87.18	84.43	84.39
BPM-GCN	93.62	79.03	83.58	85.56	85.45	85.91
BPM-GCN$^{\divideontimes}$	93.71	79.03	83.82	85.56	**85.53**	**86.33**

A. Experiment Setup

• **Dataset.** In this paper, we conduct the experiments on the Emotion-Gait [30] and ELMB [42] datasets. The Emotion-Gait dataset contains $2,177$ real gaits and $1,000$ synthetic gaits with four emotion classes (*i.e.*, happiness, sadness, anger, and neutral). The real gaits contain $1,835$ samples collected from the Edinburgh Locomotion MOCAP Database [78] and 342 gaits collected by the authors. Each gait of $1,835$ samples has 240 frames, while the 342 samples have flexible numbers of frames (27-75). The 3D skeletal data of the gaits from videos are extracted by a representative pose estimation method [79]. All real gaits are labeled by domain experts. The synthetic gaits are generated from a trained auto-encoder that has been fed with emotion labels. Following [30], we only use the $2,177$ real gaits. Besides, we also compare the BPM-GCN with SOTA methods on the ELMB [42] dataset, which is annotated by multiple emotion categories for gaits. In total, the dataset contains 3,924 gaits of which 1,835 have emotion labels provided by 10 annotators. In detail, the proportion of happy, sad, angry, neutral are 58%, 32%, 23%, and 14% respectively. In this paper, we sample every fifth frame for each gait of the subset with $1,835$ samples and obtain the first 48 frames for the second subset with 342 samples. For those samples in the second subset whose total frame numbers are less than 48, we expand them to 48 frames by padding zeros. Finally, we obtain $2,177$ real gaits with 48 frames for our experiments. As suggested in [30], the dataset is split into 9-1 for training and testing.

• **Architecture of Two Streams.** The convolution network for spatial-temporal graph is following [41]. The architectures of the posture stream and movement stream are similar except for the input and the output layers. Both of them contain a stack of the STConv blocks (10 blocks), a global max-pooling layer, and the final fully-connected layers. For the posture stream, the input-output channels for each block are {(3, 64), (64, 64), (64, 64), (64, 64), (64, 128), (128, 128), (128, 128), (128, 256), (256, 256), (256, 256)}. The last fully-connected layer contains two branches, one for emotion prediction and the other for affective feature regression. For the movement stream, the input-output channels for each block are {(8, 64), (64, 64), (64, 64), (64, 64), (64, 128), (128, 128), (128, 128), (128, 256), (256, 256), (256, 256)}. The last fully-connected layer is for emotion prediction.

• **Implementation Details.** We conduct our experiments using the PyTorch [80] framework on a single GTX 1080Ti GPU. We use the mini-batch Stochastic Gradient Descent (SGD) to optimize the proposed model. We set Nesterov momentum to 0.9 to accelerate the training process. The learning rate is set to 0.01 initially and is divided by 10 every 30 epochs. The weight decay is set to 0.001. It takes about an hour with a batch size of 32 to train the model for 80 epochs. Additionally, to make data augmentation for the training gaits, we add some noise to the coordinates by randomly choosing small rotations and translations as proposed by [77]. For training, we optimize the posture stream and movement stream separately or together (*i.e.*, the loss is $\mathcal{L}_p + \mathcal{L}_m^{cls}$) as two variants. For testing, we make the final prediction by averaging the scores of the two streams.

• **Contenders.** We compare the proposed BPM-GCN with two kinds of methods. On the one hand, the methods designed for gait emotion recognition makes much progress in the past years [30], [42]. Specifically, we compare BPM-GCN with 5 traditional algorithms [31]–[34], [75], and 4 Deep Neutral Network (DNN) based methods [30], [35], [40], [76]. On the other hand, we provide the result of the mainstream action recognition methods [41], [77], which have powerful performance on the Emotion-Gait dataset as well.

B. Comparison with SOTAs

• **Comparison with Emotion-specific Methods.** As shown in Tab. II, the proposed BPM-GCN outperforms all methods by a large margin. Moreover, based on the robustness analysis in Tab. III, the BPM-GCN is robust on the various settings and stably achieves the best performance on the Emotion-Gait dataset. We observe that the traditional methods, *i.e.*, those based on hand-engineered features, achieve less than 70% accuracy due to the limited representation ability. For the deep methods, our BPM-GCN significantly outperforms

TABLE V: Performance comparison between our proposed BPM-GCN and BPM-GCN✳ with the PM-interacted feature fusion. ✳ represents that the proposed BPM-GCN employs the PM-interacted feature fusion mechanism. Besides, the Posture✳ and Movement✳ represent training with BPM-GCN✳ and testing with single stream (*i.e*. posture or movement) respectively.

Settings	Happiness	Sadness	Anger	Neutral	Accuracy (%)
Posture	92.06	75.61	90.09	61.11	86.24
Movement	96.03	73.17	81.82	72.22	87.61
BPM-GCN	95.24	78.05	87.89	72.22	88.99
Posture✳	96.83	80.49	81.82	72.22	89.45
Movement✳	96.83	80.49	75.76	77.78	88.99
BPM-GCN✳	96.83	82.93	81.82	77.78	**90.37**

the best emotion-specific method STEP (*i.e*., 88.99% versus 83.15%), and achieves at least 5.31% (*i.e*.,88.84% versus 83.53%) improvement on mean accuracy for various split. Based on the experimental results shown in the the Tab. IV, BPM-GCN outperforms ProxEmo and STEP by 4.94% and 2.97% on the accuracy of ELMD respectively. BPM-GCN improves the precision of Happy, Anger, and Neutral on the multiple label recognition dataset, which leads to the SOTA performance on the MAP. STEP employs ST-GCN on walking videos and its advantage lies in the use of STEP-Gen for generating annotated synthetic gait to make data augmentation. However, it does not effectively mine the information hidden in joints, *i.e*., the posture and movement cues. For the proposed BPM-GCN, the posture stream learns the posture features directly from gait, and the movement stream implicitly models emotional cues by mining the velocity and acceleration information. Both streams not only imitate human perception from two important views but also collaboratively contribute to the emotion recognition.

• **Comparison with SOTA Action Recognition Methods.** ST-GCN is the first work designing a spatial-temporal graph convolutional network to learn both the spatial and temporal patterns. Similarly, 2s-AGCN models position and direction information simultaneously. However, both ST-GCN and 2s-AGCN focus on the information useful for action recognition, which contains insufficient affective cues for emotion perceiving. In contrast, the proposed model fully exploits the relation between gait and human emotions from both posture and movement perspectives, and at the same time leverages prior affective knowledge to guide the training of the model. Thus, the accuracy of 88.99% and 86.33% has been achieved on the Emotion-Gait and ELMD datasets, which exceeds the current best method 2s-AGCN by 4.59% and 1.94% respectively. Moreover, with the help of the PM-interacted feature fusion strategy, our method steadily performs better on both Emotion-Gait and ELMD datasets. For various split experiment, BPM-GCN outperforms 2s-AGCN by 4.31% for mean accuracy, and has smaller standard deviation on most settings.

• **Improving the Performance with the PM-interacted Feature Fusion Mechanism.** We conduct experiments to demonstrate the effectiveness of the proposed PM-interacted feature fusion mechanism in Tab. V. As shown in the table, the model trained with the PM-interacted feature fusion mechanism (*i.e*., *BPM-GCN*✳, 90.37%) performs better than the model without this mechanism (*i.e*., *BPM-GCN*, 88.99%) by 1.38%. In addition, based on the result shown in the Tab. III, BPM-GCN✳ has smaller standard deviation for all the three settings. The improvement and robustness are attributed to the feature sharing between the two streams. The temporal attention and spatial attention used in the mechanism can make the model focus on the important temporal sequences and spatial nodes in the gait, reduce the modality discrepancy of the two-stream features, and fuse them adaptatively. Thus, each stream (one view) can get valuable features from the other stream (the other view), and then the performance of each stream can be improved. This is verified by the single-stream accuracies in Tab. V. As shown in the table, the accuracies of the posture stream and movement stream of the *BPM-GCN*✳ are higher than that of the *BPM-GCN*, which demonstrates the effectiveness of the proposed PM-Interacted feature fusion mechanism.

TABLE VI: Ablation study of the posture and movement streams. Note '✓' indicates the corresponding stream is utilized in the training process. '$\underline{\checkmark}$' represents the posture stream without the affective constraint. Top (bottom) groups: the two streams are optimized separately (together).

Stream		Sub-class Accuracy (%)				Accuracy (%)
Posture	Movement	Happiness	Sadness	Anger	Neutral	
✓	-	97.62	75.61	54.55	61.11	83.94
-	✓	96.83	73.17	75.76	66.67	86.70
$\underline{\checkmark}$	✓	96.83	80.49	63.64	61.11	85.78
✓	✓	97.62	80.49	72.73	66.67	**88.07**
✓	-	94.44	68.29	45.45	66.67	79.82
-	✓	92.06	75.61	90.09	61.11	86.24
$\underline{\checkmark}$	✓	96.03	73.17	81.82	72.22	87.61
✓	✓	95.24	78.05	87.89	72.22	**88.99**

TABLE VII: Performance comparison of SOTA models on Emotion-Gait dataset, the sub-class accuracy for each emotion is also presented. The results before / are acquired from the initial Emotion-Gait dataset, and the ones after / are obtained with the same number of samples among classes.

Methods	Sub-class Accuracy (%)				Accuracy (%)
	Happiness	Sadness	Anger	Neutral	
ProxEmo	95.24 / 70.00	79.68 / 75.00	60.40 / 45.00	44.44 / 70.00	82.40 / 65.00
STEP	96.03 / 70.00	79.09 / 65.00	63.64 / 55.00	44.44 / 60.00	83.15 / 63.75
2s-AGCN	96.03 / 70.00	82.93 / 75.00	63.64 / 55.00	44.44/ 70.00	84.40 / 67.50
BPM-GCN	95.24 / 75.00	78.05 / 75.00	87.89 / 55.00	72.22 / 70.00	88.99 / 68.75
BPM-GCN✳	96.83 / 75.00	82.93 / 80.00	81.82 / 55.00	77.78 / 70.00	90.37 / 70.00

TABLE VIII: Comparison of SOTA models in Emotion-Gait dataset adding the simulated samples.

Methods	Sub-class Accuracy (%)				Accuracy (%)
	Happiness	Sadness	Anger	Neutral	
ProxEmo	92.85	87.80	48.48	55.56	80.61
STEP	94.94	85.20	64.55	47.78	83.71
2s-AGCN	94.36	84.08	67.79	54.19	84.27
BPM-GCN	95.24	85.36	81.38	61.67	86.73
BPM-GCN✳	96.03	85.36	81.82	66.77	**88.34**

• **Difficulty of recognizing different emotions.** We con-

TABLE IX: Complexity comparison of different models. Specifically, we calculate the FLOPs and parameters to describe the computational complexity and storage cost of the different methods. Besides, we provide the training time and the inference time when the mini-batch is unified to 32.

	ProxEmo	STEP	2s-AGCN	BPM-GCN	BPM-GCN*
FLOPs (G)	1.58	0.31	0.95	1.91	1.91
Params (M)	0.08	0.71	3.44	7.26	7.27
Training Time (S)	22.88	3.76	11.01	14.44	15.11
Inference Time (S)	7.56	0.06	2.31	3.60	3.76
Accuracy (%)	82.40	83.15	84.40	88.99	90.37

TABLE X: Performance comparison between the averaging strategy and the feature concatenation strategy for the two streams. 'Final' represents the results when combining the two streams.

Settings	Posture (%)	Movement (%)	Final (%)
Feature concatenation	77.98	84.40	87.15
Averaging	79.82	86.24	**88.99**

duct experiments to probe the difficulty for recognizing the different emotions, which is shown in the Tab. VII First, we present the detailed sub-calss accuracy on the Emotion-Gait dataset. It can be observed that the sub-class accuracy of Happiness for all methods are over 95%, and the sub-class accuracy of Neutral for the methods are lower than 45% except for BPM-GCN. Actually, the phenomenon may be caused by two reasons, *i.e*. the imbalanced number of samples for the classes and the inherent difficulty of recognizing the emotions. To eliminate the effect of the imbalanced issue, we further conduct an experiment with the same number of samples among classes. In detail, the minimum category Neutral only has 198 samples, we split the training set, testing set by 9-1, and result in 178 samples per emotion in the training set and 20 samples per emotion in the testing set. Based on the results, Happiness and Neutral are no longer the emotions with the highest and lowest accuracy. It can be observed that the Sadness is easiest to be recognized, and the sub-class accuracy of Anger is lowest.

• **Effect of synthetic samples.** To probe the influence of using the 1000 synthetic samples from the Emotion-Gait dataset, we conduct the experiment by incorporating the samples into the training set. According to the result shown in the Tab. VIII, we have the following two observations. First, compared with the result without adding synthetic samples in Tab. VII, with the help of a higher proportion of samples, the sub-class accuracy of Sadness is improved for all 5 methods due to its relatively low difficulty. Moreover, recognizing the Anger and Neutral are relatively more challenging, thus the sub-class accuracy of these two emotions does not achieve consistent improvement for all the methods. Second, after leveraging synthetic samples, the sub-class accuracy of Happiness is decreased on four methods except the BPM-GCN, and the overall accuracy on the testing set is decreased for the 5 methods. Due to the proportion of samples with Happiness being diluted in the training set, the model pays less attention to Happiness and results in the sub-class accuracy decrease.

TABLE XI: Ablation study of multiple posture features (*i.e*., angles, distances, and areas between different nodes) for the posture stream. '✓' denotes the features used in the experiment.

Posture Features			Sub-class Accuracy (%)				Accuracy (%)
Angle	Distance	Area	Happiness	Sadness	Anger	Neutral	
-	-	-	94.44	68.29	39.39	55.56	77.98
✓	-	-	94.44	78.05	33.33	55.56	78.90
-	✓	-	94.44	73.17	48.48	44.44	79.36
-	-	✓	94.44	75.61	45.45	50.00	79.82
✓	✓	-	95.24	70.73	48.49	61.11	80.73
-	✓	✓	95.24	78.05	45.45	55.56	81.19
✓	-	✓	95.24	73.17	51.52	50.00	80.73
✓	✓	✓	97.62	75.61	54.55	61.11	**83.94**
-	-	-	96.03	82.93	66.67	63.67	87.61
✓	-	-	96.03	73.17	81.82	72.22	87.61
-	✓	-	97.62	80.49	81.82	55.56	88.53
-	-	✓	97.62	80.49	75.76	61.11	88.07
✓	✓	-	95.24	80.49	81.82	61.11	88.07
-	✓	✓	97.62	85.37	72.73	66.67	**88.99**
✓	-	✓	97.62	78.05	75.76	66.67	88.53
✓	✓	✓	95.24	78.05	87.89	72.22	**88.99**

• **Complexity of SOTA methods.** We present the complexity of five SOTA methods in two aspects. On the one hand, the FLOPs and parameters are calculated to reveal the computational complexity and the storage cost. On the other hand, we record the training time and inference time of each method for one epoch. The mini-batch is unified to 32. According to the result shown in Tab. IX, we observe that STEP has the lowest cost on both storage and computational time. More importantly, with 7.22% profit on accuracy, BPM-GCN* increases only a 7.27M numbers of parameters. The acceptable cost and significant performance demonstrate the practicality of our method.

C. Ablation Studies

In this section, we conduct extensive ablation experiments on the Emotion-Gait dataset to further investigate the effectiveness of the proposed method. Specifically, it includes six aspects to highlight the benefit of each module: (1) ablation analysis of the two streams; (2) effect of posture features in the posture stream; (3) effect of movement features in the movement stream; (4) contributions of multiple human body parts for the movement stream; and (5) comparison of multiple variants of the proposed PM-Interacted feature fusion mechanism.

• **Ablation Analysis of the Two Streams.** We show the effects of the posture and movement streams in Tab. VI. First, the experiment results show that the movement stream achieves the accuracy of 86.70% (86.24%), which is better than the posture stream (83.94% (79.82%)). It indicates the movement information is even more effective for predicting human emotions. Actually, this is caused by the characteristic of the emotions. The emotions may have different arousal, thus the extent is an important cue for identification. Second, when considering both the posture stream and movement stream, we can improve the performance further to 88.07%

TABLE XII: Performance comparison among the proposed affective constraint, the simple concatenation strategy, and affective mapping [42] for the posture stream.

Settings	Happiness	Sadness	Anger	Neutral	Accuracy (%)
Simple concatenation	90.48	70.73	54.55	55.56	78.44
Affective mapping [42]	95.24	75.61	48.49	61.11	81.65
Affective constraint	97.62	75.61	54.55	61.11	**83.94**
Simple concatenation	96.83	82.93	63.64	66.67	86.70
Affective mapping [42]	96.03	73.17	81.82	72.22	87.61
Affective constraint	95.24	78.05	87.89	72.22	**88.99**

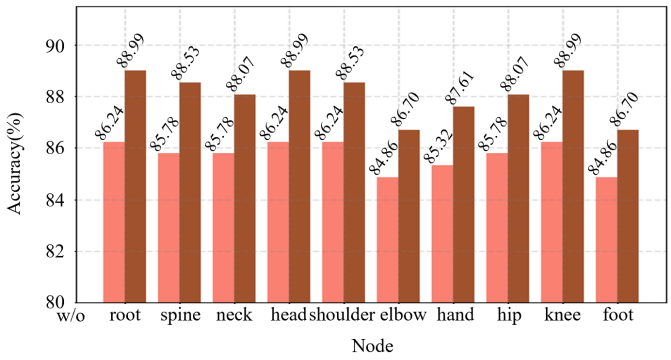

Fig. 6: Effect of body parts in the movement stream. The horizontal axis represents that we train and test the movement stream without using the information of the corresponding node.

(88.99%), which demonstrates that it is necessary to leverage both kinds of information to identify gait emotions. Furthermore, training the two streams together (88.99%) performs better than training them separately (88.07%) in terms of the overall performance. It demonstrates that combining the extent of emotion is more effective compared with only using explicit information. Note that, in the top group, the third row performs worse than the second row, this may be because the movement stream gets limited information from the posture stream (without the affective constraint) in a separate training manner. Besides, according to the sub-class accuracies, we can find that 'Happiness' is the easiest emotion to identify while 'Neutral' is the hardest one for our method (see the last row in Tab. VI). It may lie in two reasons. One is the unbalanced class distribution of the dataset, *e.g.*, 'Happiness' contains 1,160 samples while 'Neutral' includes only 198 samples. The other reason is that positive and negative emotions can be more easily revealed from posture and movement, while the neutral gait seems more difficult to perceive.

Note that the final output of the proposed model is the average score of the two streams. To demonstrate its effectiveness, we compare it with an alternative feature concatenation strategy. In detail, we add another branch by concatenating the last-layer features from the two branches and then use a fully-connected layer to predict the final result. As shown in Tab. X, the feature concatenation strategy (87.15%) performs worse than the averaging strategy (88.99%). This is because the posture and movement streams are two different views for perceiving emotions, the simple averaging strategy can make each stream pay full attention to a specific view and not be distracted by the other, thus achieving better results.

• Effect of Affective Features for the Posture Stream. To explore the effectiveness of different kinds of affective features (*i.e.*, angles, distances, and areas) for the posture stream, we perform an ablation experiment and show the results in Tab. XI. From the table, we can draw the following conclusions. First, comparing the first four rows in the table, each of the three kinds of affective features can improve the average performance of the basic model. Second, combining these affective features together can further boost the performance. Third, When we take all the three kinds of affective features into consideration and employ them in an affective constraint form, the accuracy reaches 83.94%, improving the performance by 5.96% compared to training only with classification branch (*i.e.*, the first row in the table). Such a significant gain demonstrates the effectiveness of the knowledge guidance from affective features in the posture stream. Fourth, with the help of movement stream, the accuracy is significantly improved. In addition, the affective constraint without angle achieves the same classification accuracy with just one feature, which demonstrate the effectiveness of the distance and area. However, with the help of the angle, the model pays more attention to the challenging emotions.

We also compare the proposed affective constraint with the commonly used concatenation strategy (*i.e.*, deep features and hand-crafted features in posture stream are concatenated) and the affective mapping [42] in Tab. XII. we implemented the affective mapping in our method based on [42]. Specifically, we set the channel dimension of the last block in the posture steam to 50×48. After a global max pooling layer, the output embedding feature $\mathcal{F}$ of the posture stream has the size of 50×48. Then we constrain the first 31×48 dimension of the feature (same as the dimension of the affective features) to the same values of the affective features using a Mean Square Error (MSE) loss. The whole feature $\mathcal{F}$ is also fed into another fully-connected layer followed by a softmax function to output the emotions. The posture stream is optimized by both the MSE and classification loss. As shown in Tab. XII, the proposed affective constraint can improve the performance by 5.5% and 2.29% compared with the simple concatenation strategy and affective mapping. When the model utilize both posture and movement streams, the affective constraint also brings 2.29% and 1.38%, respectively. We focus on the knowledge distillation, and such a constraint encourages the representations learned in the posture stream to carry more affective cues, so as to be more discriminative for predicting emotions. In contrast, [42] makes a constraint on part of the hidden features of the encoder-decoder network. When the network converges, the constrained hidden features fed into the classifier are nearly the same as the handcrafted affective features (*i.e.*, similar to the simple concatenation strategy), so the effectiveness is limited.

• Correspondence between handcrafted features and gait emotion recognition. Different from the general tasks *e.g.*

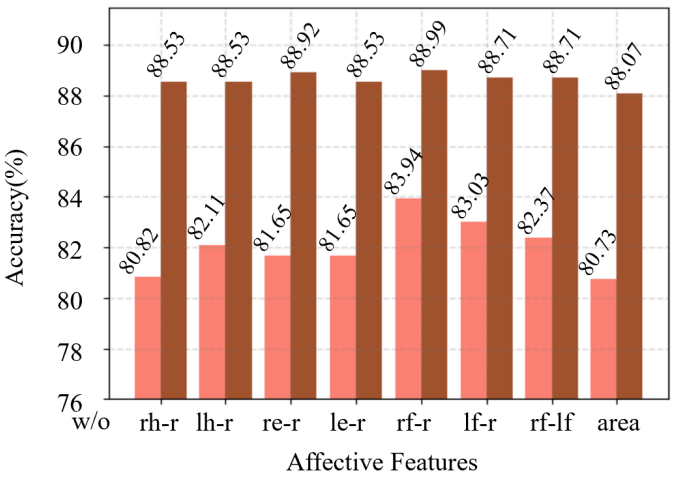

Fig. 7: Effect of affective features in the posture stream. The rh-r, lh-r, re-r, lr-r, rf-r, lf-r, rf-lf denote (right hand, root), (left hand, root), (right elbow, root), (left elbow, root), (right foot, root), (left foot, root), (left foot, right foot) respectively.

TABLE XIII: Ablation study of multiple movement features (*i.e.*, velocity and acceleration) for the movement stream. '✓' represents the features used in the experiment.

Movement features		Sub-class Accuracy (%)				Accuracy (%)
Velocity	Acceleration	Happiness	Sadness	Anger	Neutral	
✓	-	95.24	80.49	63.64	55.56	84.40
-	✓	95.24	65.85	24.24	38.89	74.31
✓	✓	96.83	73.17	75.76	66.67	**86.70**
✓	-	96.03	82.93	81.82	61.11	88.53
-	✓	95.24	73.17	78.79	66.67	86.24
✓	✓	95.24	78.05	87.89	72.22	**88.99**

action recognition focuses on the posture, gait emotion is closely related to the frequency of the gait, and the amplitude of the limbs. Accordingly, we calculate the movement in the parallel branch to perceive the frequency information, and design handcrafted features in the posture branch to inject the amplitude information into the model. Specifically, the distance is important for emotion recognition. For instance, when a person feels happiness, the amplitude is relatively greater, so the distance between the hand and root is larger. Therefore, we select 7 distances representing the amplitude of the joints. Moreover, considering the effect of symmetric joints such as the left hand and right hand, we compute the area to facilitate emotion recognition. To evaluate the influence of the 7 pairs of distance features and the area features, we conduct an ablation experiment in single posture stream and two stream architecture. The result is shown in Fig. 7. According to the performance in the figure, we can find the model without area features brings the most effect, which demonstrates the importance of the area features. For the distance features, all seven pairs bring profit to the model, specifically for the single stream architecture.

• **Effect of Different Movement Features.** We investigate the effectiveness of the movement features (*i.e.*, velocity and acceleration) in our model and summarize the results in Tab. XIII. As shown, when only velocity or acceleration feature is employed in the movement stream, the accuracies are 84.40% and 74.31%, respectively. For two stream case, the model only achieves 86.24% when only acceleration is employed. It shows that velocity carries more emotional cues than acceleration. Intuitively, we can roughly infer a person's emotion via the velocity of his walking and arm swing. Nevertheless, the acceleration cues can also contribute to this task. Acceleration can be seen as a supplement to the velocity (indicating the change of a person's movement state), which is also related to the emotion expression, and thus is auxiliary to enhancing the emotion recognition as shown in the last row of Tab. XIII.

TABLE XIV: Performance comparison of multiple feature fusion mechanisms.

Settings	Happiness	Sadness	Anger	Neutral	Accuracy (%)
1	95.24	78.05	87.89	72.22	88.99
2	97.62	78.05	75.76	66.67	88.07
3	97.62	82.93	75.76	66.67	88.99
4	97.62	85.37	78.79	61.11	89.45
5	96.83	78.05	75.76	61.11	87.16
6	96.83	90.24	69.70	61.11	88.53
7	96.83	80.49	81.82	66.67	88.99
8	96.83	82.93	81.82	77.78	**90.37**

[1] The proposed *BPM-GCN* without any feature fusion mechanism. [2] PM-Interacted feature fusion without the attention mechanism. [3] PM-Interacted feature fusion mechanism with only the spatial attention mechanism. [4] PM-Interacted feature fusion mechanism with only the temporal attention. [5] The spatial attention and temporal attention are integrated by summation. [6] The spatial attention and temporal attention are integrated in a sequential way, *i.e.*, the spatial attention is in front of the temporal attention. [7] The spatial attention and temporal attention are integrated in another sequential way, *i.e.*, the spatial attention is behind the temporal attention. [8] The proposed PM-Interacted feature fusion mechanism in this paper.

• **Effect of Body Parts for the Movement Stream.** To explore the effects of different human body parts (*e.g.*, hand, head, foot, *etc.*) for perceiving emotions, we experiment by masking out each part of the human body to train the model. As shown in Fig. 6, when masking out the elbow, foot, and hand, the performance is lowest. Particularly in two stream model, the accuracy are 86.70%, 86.70%, and 87.61% respectively, which drop over 1% compared with the complete method. It indicates the movement information of these nodes is more related to emotions, *i.e.*, the movement of arms and feet (*e.g.*, walking velocity and velocity of arm swing) is important to express one's emotions. In contrast, the movement of the root, shoulder, and knee contributes less to recognizing emotions, and this may be because these parts always move similarly with the whole body and lack independent movement.

• **Ablation Analysis of the PM-Interacted Feature Fusion mechanism.** We make an ablation analysis of the proposed PM-Interacted feature fusion mechanism in Tab. XIV. The second row of the table is a simple fusion of the movement stream feature and posture stream feature (*e.g.*, for the posture stream features, $x_i^p = x_i^p + x_i^m$). Compared with the results in the first row, we can find such a compulsory fusion can

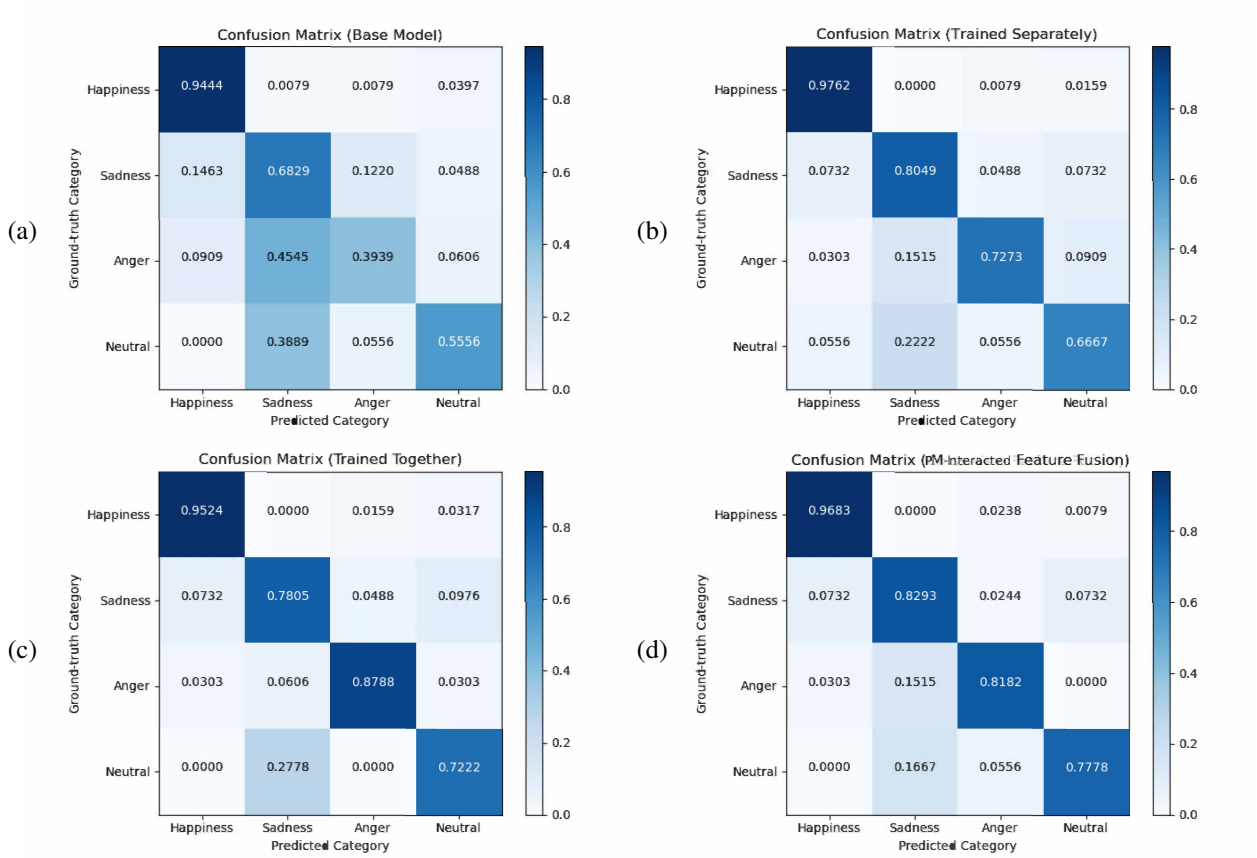

Fig. 8: Visualization of four confusion matrices. (a) The base model has a single posture stream without the affective constraint. (b) We train the proposed posture stream and movement stream separately. (c) We train the proposed posture stream and movement stream together (*BPM-GCN*). (d) The proposed *BPM-GCN** with the PM-Interacted feature fusion mechanism.

decrease the performance of the model. It may be attributed to the modality discrepancy between the posture feature and movement features. The third row of the table is the PM-Interacted feature fusion with only the spatial attention mechanism (*e.g.*, for the posture features, $x_i^p = x_i^p + x_i^m * N_{att}(x_i^m)$). The fourth row of the table is the PM-Interacted feature fusion with only the temporal attention mechanism (*e.g.*, $x_i^p = x_i^p + x_i^m * T_{att}(x_i^m)$ for the posture stream). Compared with the compulsory fusion (*i.e.*, the second row in the table), these two attention mechanisms can both improve the performance by focusing on discriminative features. The fifth-seventh rows are several variants of the proposed PM-Interacted feature fusion mechanism. The fifth row shows a fusion mechanism, in which the spatial attention and temporal attention are integrated by the summation operation (*e.g.*, $x_i^p = x_i^p + x_i^m * N_{att}(x_i^m) + x_i^m * T_{att}(x_i^m)$ for the posture stream). Compared with the third row and the fourth row, we can find that an inappropriate combination of the spatial attention and temporal attention mechanism can decrease the performance. The sixth row is a sequential spatial attention and temoporal attention mechanism, *i.e.*, the spatial attention is in front of the temopral attention (*e.g.*, $x_i^p = x_i^p + x_i^m * T_{att}(N_{att}(x_i^m))$ for the posture stream). The seventh row is another kind of sequential spatial attention and temporal attention mechanism, *i.e.*, the spatial attention is behind the temopral attention (*e.g.*, for the posture stream, $x_i^p = x_i^p + x_i^m * N_{att}(T_{att}(x_i^m)))$. The final row shows the proposed model with the PM-Interacted feature fusion mechanism (*i.e.*, *BPM-GCN**). The proposed PM-Interacted feature fusion mechanism (*i.e.*, *BPM-GCN**) can improve the model without any feature fusion mechanism (*i.e.*, *BPM-GCN*) by 1.38% and can also exceed the variants in fifth and seventh rows, which demonstrate its superiority.

D. Discussion about the Performance Improvement

To summarize, the proposed components are designed from three aspects to improve performance. The first is where we can find more emotional cues from gait. Different from most recent works that only use the 3D joint coordinates as input (posture view), we devise another movement stream that can extract more discriminative emotion cues from the body movement. The results in the first three rows of Tab. VI prove that adding the movement stream improves the performance of only using the posture stream by 4.13% (88.07% vs. 83.94%). The second aspect is how to reduce the gap between gait and emotion. To address this issue, we design an affective constraint to distill the prior affective knowledge into the model and guide the model to learn the emotion-related features better, which improves the accuracy by 1.38% than training without the constraint (88.99% vs. 87.61%). The

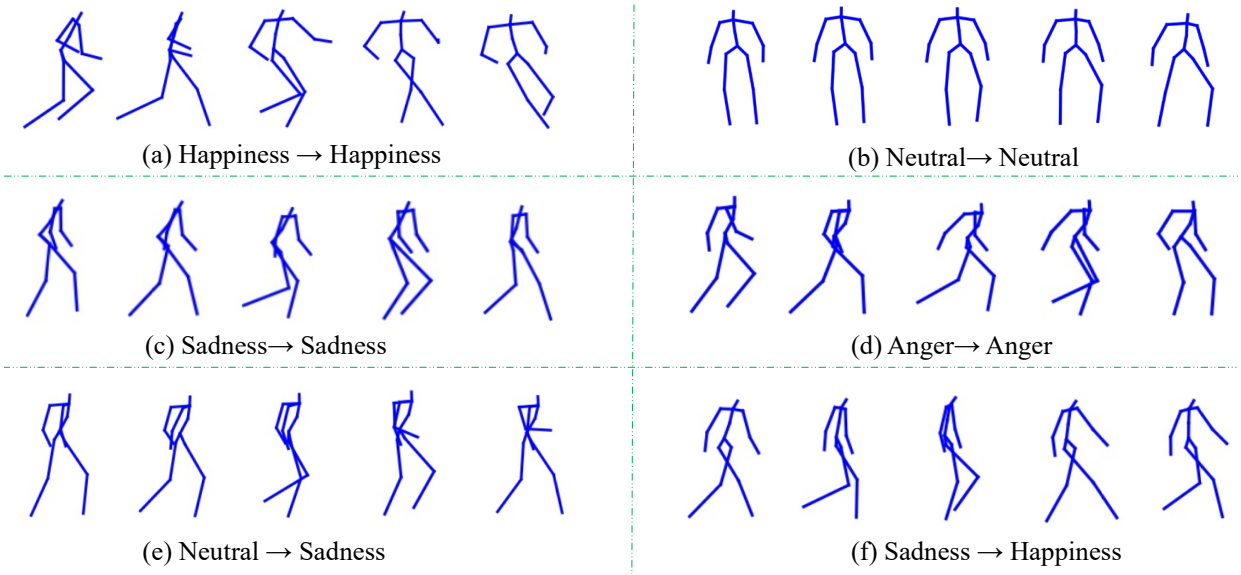

Fig. 9: Visualization of some samples in a sequence and their predictions by *BPM-GCN**. The texts before and after '$\longrightarrow$' represent the ground-truth category and the predicted one respectively. (a)-(d) are positive examples, while (e)-(f) are failure cases.

third aspect is how to exchange useful knowledge between the features from the posture stream and movement stream, and fuse them adaptatively to further improve the discrimination ability of the model. Thus, we propose a PM-Interacted feature fusion mechanism that can further improve the performance from 88.99% to 90.37%. As shown in Fig. 8, we plot the confusion matrices of different model variants, *i.e.*, (a) the base model that consists of only the posture stream without the affective constraint, (b) the proposed *BPM-GCN* using the separate training mechanism, (c) the proposed *BPM-GCN* with the two branches training together, and (d) the proposed *BPM-GCN** with the PM-Interacted feature fusion mechanism. From (a) to (d), the misclassification rate (shades of the color in the off-diagonal blocks) gradually decreases, validating the effectiveness of our proposed components, which are well motivated and thus clearly boost the performance.

As shown in Fig. 9, we plot some gait examples in a sequence. In most cases, *e.g.*, (a), (b), (c), and (d), the proposed model can accurately recognize the emotions. We can find that the happiness samples have more animated joint movements than others, while the neutral samples have fewer joint movements than others. The sadness samples are usually down in the dumps. The anger samples are frequently with large strides and cadence. In some cases, *e.g.*, (e) and (f), the predicted labels for a gait do not match the ground-truth. Most failure cases are like (e). It is reasonable that both neutral and sadness samples are always with low arousal scales. For (f), the sadness samples are with a relatively large stride, thus making it confused with the happiness. Increasing more training samples of the neutral and sadness categories may help to discover the subtle differences and learn discriminative features, therefore solving these failure cases.

V. Conclusions

In this paper, we address the problem of emotion recognition of individuals based on their walking styles (gait). We present a Bilateral Posture and Movement Graph Convolutional Network (BPM-GCN) to imitate the perception of emotions from two important views. The posture stream models emotions from 3D coordinates of gait and leverages the prior affective knowledge to reduce the gap between gait and emotions. The movement stream implicitly describe the extent of emotions by the informative velocity and acceleration pairs. We further design a PM-Interacted feature fusion mechanism, which can adaptatively fuse the features from the posture stream and movement stream. Therefore, in this way, BPM-GCN can imitate human emotion perception from both posture and movement views and the two views can benefit each other. Extensive experiments on the benchmark dataset demonstrate the superiority of BPM-GCN. We hope that this idea, which exploits posture and movement information in an end-to-end manner, can open a new perspective for improving the performance of gait emotion recognition. In the future, we plan to explore more robust emotion recognition from both gait and other modalities, *e.g.*, face, text, speech, *etc*.

Acknowledgements

This work was supported by the National Key Research and Development Program of China Grant (NO.2018AAA0100400), Natural Science Foundation of Tianjin, China (NO.20JCJQJC00020), Fundamental Research Funds for the Central Universities, and Supercomputing Center of Nankai University (NKSC).

References

[1] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *CVPR*, 2020.

[2] X. Yao, D. She, S. Zhao, J. Liang, Y.-K. Lai, and J. Yang, "Attention-aware polarity sensitive embedding for affective image retrieval," in *ICCV*, 2019.

[3] A. Bauer, K. Klasing, G. Lidoris, Q. Mühlbauer, F. Rohrmüller, S. Sosnowski, T. Xu, K. Kühnlenz, D. Wollherr, and M. Buss, "The autonomous city explorer: Towards natural human-robot interaction in urban environments," *International Journal of Social Robotics*, vol. 1, no. 2, pp. 127–140, 2009.

[4] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun, "A comprehensive and context-sensitive neonatal pain assessment using computer vision," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 28–45, 2022.

[5] L.-H. Kong, W. He, W.-S. Chen, H. Zhang, and Y.-N. Wang, "Dynamic movement primitives based robot skills learning," *Machine Intelligence Research*, vol. 20, no. 3, pp. 396–407, 2023.

[6] S. A. Denham, E. Workman, P. M. Cole, C. Weissbrod, K. T. Kendziora, and C. Zhan-Waxler, "Prediction of externalizing behavior problems from early to middle childhood: The role of parental socialization and emotion expression," *Development and Psychopathology*, vol. 12, no. 1, pp. 23–45, 2000.

[7] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, "Correction to: Yolop: You only look once for panoptic driving perception," *Machine Intelligence Research*, pp. 1–1, 2023.

[8] Q.-Y. Yin, J. Yang, K.-Q. Huang, M.-J. Zhao, W.-C. Ni, B. Liang, Y. Huang, S. Wu, and L. Wang, "Ai in human-computer gaming: Techniques, challenges and opportunities," *Machine Intelligence Research*, vol. 20, no. 3, pp. 299–317, 2023.

[9] M. Atcheson, V. Sethu, and J. Epps, "Gaussian process regression for continuous emotion recognition with global temporal invariance," in *IJCAI Workshops*, 2017.

[10] H. Yates, B. Chamberlain, G. Norman, and W. H. Hsu, "Arousal detection for biometric data in built environments using machine learning," in *IJCAI Workshops*, 2017.

[11] S. Al-Saqqa, H. Abdel-Nabi, and A. Awajan, "A survey of textual emotion detection," in *CSIT*, 2018.

[12] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text and speech: a survey," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–26, 2018.

[13] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 53–79, 2020.

[14] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: Two decades review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[15] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2021.

[16] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2012.

[17] J. Yang, J. Li, L. Li, X. Wang, Y. Ding, and X. Gao, "Seeking subjectivity in visual emotion distribution learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 5189–5202, 2022.

[18] J. C. S. Jacques Junior, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. J. van Gerven, R. van Lier, and S. Escalera, "First impressions: A survey on vision-based apparent personality trait analysis," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 75–95, 2022.

[19] Z. Chen, R. Ansari, and D. J. Wilkie, "Learning pain from action unit combinations: A weakly supervised approach via multiple instance learning," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 135–146, 2022.

[20] X. Jia, X. Zheng, W. Li, C. Zhang, and Z. Li, "Facial emotion distribution learning by exploiting low-rank label correlations locally," in *CVPR*, 2019.

[21] H. K. Meeren, C. C. van Heijnsbergen, and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," *Proceedings of the National Academy of Sciences*, vol. 102, no. 45, pp. 16 518–16 523, 2005.

[22] J. Michalak, N. F. Troje, J. Fischer, P. Vollmar, T. Heidenreich, and D. Schulte, "Embodiment of sadness and depression—gait patterns associated with dysphoric mood," *Psychosomatic Medicine*, vol. 71, no. 5, pp. 580–587, 2009.

[23] J. M. Montepare, S. B. Goldstein, and A. Clausen, "The identification of emotions from gait information," *Journal of Nonverbal Behavior*, vol. 11, no. 1, pp. 33–42, 1987.

[24] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 505–523, 2018.

[25] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *AAAI*, 2020.

[26] J. P. Singh, S. Jain, S. Arora, and U. P. Singh, "Vision-based gait recognition: A survey," *IEEE Access*, vol. 6, pp. 70 497–70 527, 2018.

[27] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On learning disentangled representations for gait recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 345–360, 2022.

[28] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "Gaitset: Cross-view gait recognition through utilizing gait as a deep set," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3467–3478, 2022.

[29] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition via semi-supervised disentangled representation learning to identity and covariate features," in *CVPR*, 2020.

[30] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha, "Step: Spatial temporal graph convolutional networks for emotion perception from gaits," in *AAAI*, 2020.

[31] G. Venture, H. Kadone, T. Zhang, J. Grèzes, A. Berthoz, and H. Hicheur, "Recognizing emotions conveyed by human gait," *International Journal of Social Robotics*, vol. 6, no. 4, pp. 621–632, 2014.

[32] M. Karg, K. Kuhnlenz, and M. Buss, "Recognition of affect based on gait patterns," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 40, no. 4, pp. 1050–1061, 2010.

[33] M. Daoudi, S. Berretti, P. Pala, Y. Delevoye, and A. Del Bimbo, "Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices," in *ICIAP*, 2017.

[34] W. Wang, V. Enescu, and H. Sahli, "Adaptive real-time emotion recognition from body movements," *ACM Transactions on Interactive Intelligent Systems*, vol. 5, no. 4, pp. 1–21, 2015.

[35] T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera, and D. Manocha, "Learning perceived emotion using affective and deep features for mental health applications," in *ISMAR-Adjunct*, 2019.

[36] J. Yang, J. Li, X. Wang, Y. Ding, and X. Gao, "Stimuli-aware visual emotion analysis," *IEEE Transactions on Image Processing*, vol. 30, pp. 7432–7445, 2021.

[37] Z. Shen, J. Cheng, X. Hu, and Q. Dong, "Emotion recognition based on multi-view body gestures," in *ICIP*, 2019.

[38] J. Yang, X. Gao, L. Li, X. Wang, and J. Ding, "Solver: Scene-object interrelated visual emotion reasoning network," *IEEE Transactions on Image Processing*, vol. 30, pp. 8686–8701, 2021.

[39] M. Han, Y. Zhan, B. Yu, Y. Luo, H. Hu, B. Du, Y. Wen, and D. Tao, "Region-adaptive concept aggregation for few-shot visual recognition," *Machine Intelligence Research*, pp. 1–15, 2023.

[40] V. Narayanan, B. M. Manoghar, V. S. Dorbala, D. Manocha, and A. Bera, "Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation," in *IROS*, 2020.

[41] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[42] U. Bhattacharya, C. Roncal, T. Mittal, R. Chandra, K. Kapsaskis, K. Gray, A. Bera, and D. Manocha, "Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping," in *ECCV*, 2020.

[43] C. L. Roether, L. Omlor, A. Christensen, and M. A. Giese, "Critical features for the perception of emotion from gait," *Journal of Vision*, vol. 9, no. 6, pp. 1–32, 2009.

[44] N. H. Frijda, "The laws of emotion." *American Psychologist*, vol. 43, no. 5, pp. 349–358, 1988.

[45] R. Sun, N. Wilson, and M. Lynch, "Emotion: a unified mechanistic interpretation from a cognitive architecture," *Cognitive Computation*, vol. 8, no. 1, pp. 1–14, 2016.

[46] D. N. Tam, "Computation in emotional processing: quantitative confirmation of proportionality hypothesis for angry unhappy emotional intensity to perceived loss," *Cognitive Computation*, vol. 3, no. 2, pp. 394–415, 2011.

[47] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-GCN: Correlation based graph convolutional network for audio-video emotion recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 3793–3804, 2021.

[48] A. Shirian, S. Tripathi, and T. Guha, "Dynamic emotion modeling with learnable graphs and graph inception network," *IEEE Transactions on Multimedia*, vol. 24, pp. 780–790, 2021.

[49] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014–4026, 2021.

[50] Z.-B. Yu and M.-L. Zhang, "Multi-label classification with label-specific feature generation: A wrapped approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5199–5210, 2022.

[51] S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, "Emotion recognition from multiple modalities: Fundamentals and methodologies," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 59–73, 2021.

[52] M. Argyle, *Bodily Communication*. Routledge, 2013.

[53] B. De Gelder, "Why bodies? twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3475–3484, 2009.

[54] B. de Gelder and N. Hadjikhani, "Non-conscious recognition of emotional body language," *Neuroreport*, vol. 17, no. 6, pp. 583–586, 2006.

[55] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement," *Cognition*, vol. 82, no. 2, pp. B51–B61, 2001.

[56] R. T. Boone and J. G. Cunningham, "Children's decoding of emotion in expressive body movement: The development of cue attunement." *Developmental Psychology*, vol. 34, no. 5, pp. 1007–1016, 1998.

[57] M. M. Gross, E. A. Crane, and B. L. Fredrickson, "Effort-shape and kinematic assessment of bodily expression of emotion during gait," *Human Movement Science*, vol. 31, no. 1, pp. 202–221, 2012.

[58] L. Omlor and M. A. Giese, "Extraction of spatio-temporal primitives of emotional body expressions," *Neurocomputing*, vol. 70, no. 10-12, pp. 1938–1942, 2007.

[59] B. Li, C. Zhu, S. Li, and T. Zhu, "Identifying emotions from non-contact gaits information based on microsoft kinects," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 585–591, 2016.

[60] M. Chiu, J. Shu, and P. Hui, "Emotion recognition through gait on mobile devices," in *PerCom Workshops*, 2018.

[61] X. Chen, X. Luo, J. Weng, W. Luo, H. Li, and Q. Tian, "Multi-view gait image generation for cross-view gait recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 3041–3055, 2021.

[62] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey," *Machine Intelligence Research*, pp. 1–36, 2023.

[63] Y. Gu, J. Wen, H. Sun, Y. Song, P. Ke, C. Zheng, Z. Zhang, J. Yao, L. Liu, X. Zhu *et al.*, "Eva2. 0: Investigating open-domain chinese dialogue systems with large-scale pre-training," *Machine Intelligence Research*, vol. 20, no. 2, pp. 207–219, 2023.

[64] M. Wang, B. Ni, and X. Yang, "Learning multi-view interactional skeleton graph for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[65] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *CVPR*, 2015.

[66] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *CVPR*, 2017.

[67] W. Guo, H.-L. Zhen, X. Li, W. Luo, M. Yuan, Y. Jin, and J. Yan, "Machine learning methods in solving the boolean satisfiability problem," *Machine Intelligence Research*, pp. 1–16, 2023.

[68] W. Xing, J. Chen, and Y. Guo, "Robust local light field synthesis via occlusion-aware sampling and deep visual feature fusion," *Machine Intelligence Research*, vol. 20, no. 3, pp. 408–420, 2023.

[69] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *ACPR*, 2015.

[70] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *CVPR Workshops*, 2017.

[71] Y.-F. Song, Z. Zhang, and L. Wang, "Richly activated graph convolutional network for action recognition with incomplete skeletons," in *ICIP*, 2019.

[72] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.

[73] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *CVPR*, 2019.

[74] S. M. S. Hasan, M. R. Siddiquee, J. S. Marquez, and O. Bai, "Enhancement of movement intention detection using eeg signals responsive to emotional music stimulus," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1637–1650, 2022.

[75] A. Crenn, R. A. Khan, A. Meyer, and S. Bouakaz, "Body expression recognition from animated 3D skeleton," in *3DV*, 2016.

[76] C. Hu, W. Sheng, B. Dong, and X. Li, "Tntc: Two-stream network with transformer-based complementarity for gait-based emotion recognition," in *ICASSP*, 2022.

[77] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.

[78] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NIPS*, 2017.

[79] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain, "Learning 3D human pose from structure and motion," in *ECCV*, 2018.

[80] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.