AST-GCN: Augmented Spatial Temporal Graph Convolutional Neural Network for Gait Emotion Recognition

Chuang Chen, *Graduate Student Member, IEEE*, Xiao Sun, *Member, IEEE*, Zhengzheng Tu, and Meng Wang, *Fellow, IEEE*

***Abstract*— Skeleton-based methods have recently achieved good performance in deep learning-based gait emotion recognition (DL-GER). However, the current methods have two drawbacks that limit the ability to learn discriminative emotional features from gait. First, these methods do not exclude the effect of the subject's walking orientation on emotion classification. Second, they do not sufficiently learn the implicit connections between the joints during human walking. In this paper, an augmented spatial-temporal graph convolutional neural network (AST-GCN) is introduced to solve these two problems. The interframe shift encoding (ISE) module acquires interframe shifts of joints to make the network sensitive to changes in emotion-related joint movements regardless of the subject's walking orientation. A multichannel implicit connection inference method learns more implicit connection relations related to emotions. Notably, we unify current skeleton-based methods into a common framework that validates the most powerful feature representation capability of our AST-GCN from a theoretical perspective. In addition, we extend the skeleton-based gait dataset using posture estimation software. Experiments demonstrate that our AST-GCN outperforms state-of-the-art methods on three datasets on two tasks.**

Manuscript received 3 July 2023; revised 15 September 2023; accepted 4 December 2023. Date of publication 12 December 2023; date of current version 6 June 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC3803202, in part by the Major Project of Anhui Province under Grant 202203a05020011, and in part by the General Programmer of the National Natural Science Foundation of China under Grant 62376084. This article was recommended by Associate Editor C. Yang. *(Corresponding author: Xiao Sun.)*

Chuang Chen is with the School of Artificial Intelligence, Anhui University, Hefei 230601, China, and also with the Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei 230088, China (e-mail: wa21301027@stu.ahu.edu.cn).

Xiao Sun is with the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machines, School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China, and also with the Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei 230088, China (e-mail: sunx@hfut.edu.cn).

Zhengzheng Tu is with the Information Materials and Intelligent Sensing Laboratory of Anhui Province and the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: zhengzhengahu@163.com).

Meng Wang is with the Key Laboratory of Knowledge Engineering With Big Data, Ministry of Education, and the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China, and also with the Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei 230088, China (e-mail: eric.mengwang@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3341728.

Digital Object Identifier 10.1109/TCSVT.2023.3341728

***Index Terms*— Gait emotion recognition, graph convolutional, affective computing.**

I. Introduction

AUTOMATIC emotion recognition [1], [2], [3], [4], [5], [6] is crucial for advancing artificial intelligence and has widespread applications across diverse domains, including human-computer interaction [7], intelligent health care [8], and anomaly detection [9]. Its primary purpose is to identify an individual's emotional state using visual cues (e.g., facial expressions [5], [6], [10], gait [11]) or nonvisual cues (e.g., conversation [12], physiological signals [13]), thereby supporting decision-making [14], [15] at the top. Traditional facial expression-based emotion recognition techniques [5], [6], [10] can be unreliable due to issues such as "mock" expressions [16], empirical error [17] and self-reported emotions in certain scenarios [18]. Additionally, conversation [12] and physiological signal-based [13] emotion recognition techniques may not be suitable for public scenarios, as they require subjects to wear special instruments for data acquisition. In contrast, gait-based emotion recognition techniques [11] mitigate these concerns through their characteristics, including difficulty of imitation [11], [19] and cost-effective data acquisition [20], [21].

Despite the many advantages of using gait for emotion recognition, obtaining discriminative emotional features from gait remains a research challenge. Psychologists [22] have empirically established the presence of various emotional states in humans during walking. Prior work has explored both dominant emotions and multiple emotions in gait, corresponding to multiclass classification tasks [23], [24], [25], [26], [27], [28] and multilabel classification tasks [25], [27], [29], respectively. Furthermore, previous research can be categorized into machine learning-based gait emotion recognition (ML-GER) methods [11], [30], [31] and deep learning-based gait emotion recognition (DL-GER) methods [23], [24], [25], [26], [27], [28], [29] according to the model type. ML-GER methods [11], [30], [31] put handcrafted affective features into traditional classifiers: SVM [32], random forest [33], etc., for classification. With the development of deep learning, DL-GER methods have achieved higher feature expression capability than ML-GER methods and have gradually replaced the latter. However, handcrafted affective features are often

cited as feature inputs [23], [24], [25], [26], [28] or pretrained features [27] for DL-GER methods.

DL-GER methods [23], [24], [25], [26], [27], [28], [29] are mainly classified into sequence-based methods [24], image-based methods [23], [26], [28], and skeleton-based methods [25], [27], [29] according to the encoding method. Sequence-based methods [24] rely on RNN architecture [34] and focus on exploring correlations among gait frames while neglecting the investigation of joint connections within the gait. Image-based methods [23], [26], [28] convert gait sequences into images and employ the CNN architecture to study relationships among gait frames and joints. The disadvantages of these methods [23], [26], [28] are that they are less efficient in mining the change pattern of nonadjacent joints or frames at longer geometric distances in the image space due to the local perception capability of the convolution kernel [35]. Finally, skeleton-based methods [25], [27], [29] based on the GCN architecture [36] view gait sequences as graph sequences to learn deep representations. Methods [25], [27], [29] currently achieve the best performance but have two drawbacks.

Inspired by natural observation, we observe other human beings walking, for example, pedestrians on the road. Regardless of the pedestrian's orientation, we can precisely identify the emotional state of the pedestrian. Essentially, we judge the emotional state from the relative shift of the pedestrian's joints between frames. As shown in Fig. 1d, if a pedestrian has a transient head down and bent during walking, the relative shift of the head joint and the waist joint point is fixed regardless of the observer's viewpoint and the pedestrian's orientation, based on which we think the pedestrian may be unhappy. However, skeleton-based methods [25], [27], [29] currently extract features from gait sequences and do not realize the importance of interframe shifts of joint points. Feature extraction directly from the gait sequence using a network is susceptible to the influence of subject orientation in the dataset when classifying emotions. In fact, the orientation of the subject is not related to his or her emotional state.

We can always identify the emotional state of a human based on certain actions. As shown in Fig. 1b, when a pedestrian holds his or her head up and takes a long stride, we usually assume that he or she is happy. The head being up widens the angle formed by the chest joint as the apex and the head-chest and hip-chest as the bilateral sides. A large stride is a widening of the angle formed by the crotch joint as the apex and the left foot-crotch and right foot-crotch as the bilateral sides. Therefore, we argue that there are many special connections between human joints, which we call implicit connections (as shown in Fig. 1d). A change in angle can be described by the strength of the connection between the joint points. Based on the above discussion, the implicit connection of numerous joint points during walking provides context-sensitive emotional information for gait emotion recognition. Prior works [27] mined only a limited number of kinds of implicit connections using a multihead self-attention mechanism [37]. Intuitively, human joints should have more kinds of implicit connections (as shown in Fig. 1c) that provide more context-sensitive emotional information for gait emotion classification.

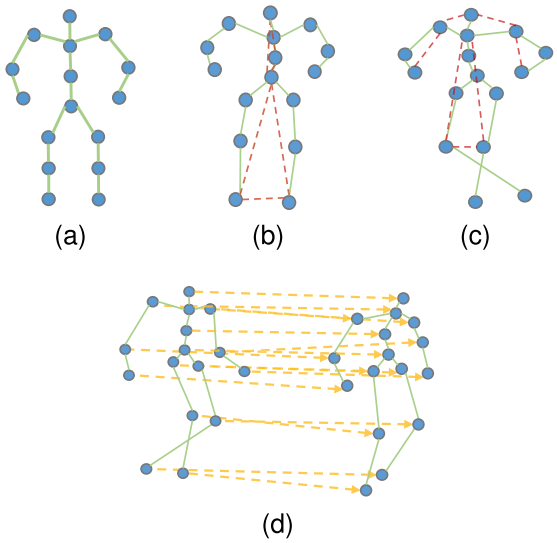


Fig. 1. The blue circles indicate joint points. Green solid lines indicate skeletons. The red dashed lines indicate the implicit connection relationships of the joint points. The yellow dashed lines indicate the interframe shifts of joint points. (a) Explicit connection. (b) Implicit connection. (c) More implicit connections. (d) Interframe shift of all joint points.

To address the above problem, we designed a novel general representation for skeleton-based gait emotion recognition by expanding the graph convolutional neural network [36] to an augmented spatial temporal graph convolutional neural network, which we call AST-GCN. AST-GCN has an interframe shift encoding (ISE) module to obtain the interframe shifts of joints, allowing the network to be more sensitive to changes in joints associated with emotions. It excludes the effect of the subject's walking orientation for emotion classification. In terms of spatial modeling, AST-GCN proposes a multichannel implicit connection inference method that can learn multiple emotion-related implicit connection relationships during human walking, solving the problem of insufficient learning of implicit connections in prior work [27]. The method mainly utilizes an attention-improved mechanism to inference the multichannel implicit connection of the mixture matrix after fusion of the embedded and accepted tensors. In terms of spatial modeling, [38] was introduced to allow the model to learn multiscale information about joint point motion. We unify all current skeleton-based methods [24], [25], [27] into a common framework, and we validate from a theoretical perspective that our AST-GCN has stronger feature representation capability than other skeleton-based methods [24], [25], [27]. In addition to obtaining more skeleton-based gait data to evaluate the performance of our AST-GCN, we first transformed the HEROES dataset [39] into 3D pose format using VideoPose3D [40], which we call HEROES3D. In summary, our main contributions are as follows:

- We propose a novel AST-GCN to address two problems of skeleton-based methods: we introduce an ISE module focusing on the intraframe shift of the joint points regardless of the effect of the subject's orientation, and we introduce a multichannel implicit connection inference method to learn more implicit connection relations related to emotion.

- We unify skeleton-based methods into a framework that theoretically demonstrates the most powerful feature representation of our AST-CGN.
- Extensive experiments were performed on EGait [25], EWalk [24] and HEROES3D. For the first time, we utilized VideoPose3D [40] to extract the 3D pose from HEROES [39], demonstrating that our AST-GCN achieves the best performance. Additional experiments validate the effectiveness of the proposed module, the robustness of our AST-GCN cross-datasets, the strengthen of skeleton-based methods and important joints in gait emotion expression.

The rest of the paper is organized as follows. First, Section II describes related works on DL-GRE method and ML-GRE methods. Next, Section III presents the module details of AST-GCN and the loss function for training. After that, Section IV provides comprehensive experiments and corresponding analysis. Finally, Section V concludes the paper.

II. Related Works

In this section, we focus on prior works [11], [23], [24], [25], [26], [27], [28], [29], [30], [31] in gait emotion recognition (GER), which is divided into machine learning-based gait emotion recognition (ML-GER) methods [11], [30], [31] and deep learning-based gait emotion recognition (DL-GER) methods [23], [24], [25], [26], [27], [28], [29] according to model type.

A. ML-GER

Karg et al. [30] extracted several key motion parameters from gait, including stride length, time of walking, speed of walking, and limb angle, and transformed these parameters mathematically, followed by classification using SVM [32]. Later, Li et al. [11] used a Microsoft Kinect device [41] to collect human no-contact gait information. Special attention was given to positional parameters such as the hand, shoulder, and spine. A time-frequency analysis method was proposed to obtain time and frequency domain features for classification with traditional classifiers: SVM [32] and naive Bayesian model [42]. Inspired by prior work and psychological studies [43], Crenn et al. [31] proposed the use of geometric features, such as the distance and area angle of joint points and motion features, including acceleration and velocity of joint point motion, and Fourier features such as the spectral size of different joints for emotional recognition of 3D gait skeleton sequences. Afterward, multiple features were fused and normalized into the SVM [32] classifier. In summary, the idea of ML-GER methods [11], [30], [31] is to extract hand-crafted affective features of gait that have been verified by psychologists or experiments and to classify them using different machine learning algorithms [32], [33], [42]. With the development of deep learning in the field of GER, neural network models have been used instead of traditional machine learning algorithms [32], [33], [42] for feature extraction and classification. Some handcrafted affective features that have been validated by psychological studies and experiments [43] for gait emotion classification are retained. Handcrafted affective features are often used to pretrain the network latent layer [27] or are fused with other features for classification [23], [24], [25], [26], [29].

B. DL-GER

DL-GER methods [23], [24], [25], [26], [27], [29] can be divided into three categories according to how the data are encoded: sequence-based [24], image-based [23], [26], [28], and skeleton-based [24], [25], [27]. The sequence-based methods [24], represented by Randhavane et al., use a long short-term memory network (LSTM) [44] to extract interframe association features from gait sequences consisting of 3D joint coordinates. The association features are fused with hand-crafted affective features to obtain fused features. The best result was achieved using fused features for classification in the EWalk [24] dataset. The disadvantage of this method is that it ignores the excavation of the association between joint points during human walking. Image-based methods [23], [26], [28] encode gait sequences consisting of 3D coordinates of joints as images according to different regulations and perform gait emotion feature extraction with the powerful feature representation capability of convolutional neural networks [36]. The pioneering work [23] by Narayanan et al. proposed a multiview skeleton graph convolution method by introducing group convolution to extract features from images encoded by multiview 3D joint coordinates and achieved accuracy superior to that of prior work. Later, Hu et al. [26] encoded handcrafted affective features as images and used ResNet-34 [45] to extract features from the affective feature images and joint point coordinate images. To capture the long dependencies, a transformer was used to complement the two types of extracted features. In summary, the core of image-based methods [23], [26], [28] is to encode joint coordinates into an image by adopting different types of encoding methods. The disadvantage of this approach is that due to the local perception ability of the convolutional kernel [35], the ability to perceive changes in the spatial relationship of an image at a long distance, i.e., nonadjacent joints or nonadjacent frames, is insufficient, which severely limits the accuracy improvement for emotion classification. Skeleton-based methods [25], [27], [29] consider joints as a graph structure in non-Euclidean space after being connected by human biological chain relations [46]. Therefore, a graph convolutional neural network [36] was used to model the graph structured data spatially and temporally. The method proposed by Bhattacharya et al. [25] extracts latent features in turn by aggregating the joint point features within human body parts and the features between body parts. This method does not adequately extract the joint associations between body parts, and is essentially a restricted implementation of the idea of graph neural network aggregation. Later, Bhattacharya [29] et al. introduced ST-GCN [47], which has shown superior performance in action recognition [48], to perform spatial and temporal modeling of joint point motion. However, this modeling method cannot represent the implicit connections of the joints in spatial modeling and does not extract sufficient multiscale temporal features for the joints in temporal modeling, which limits the ability of

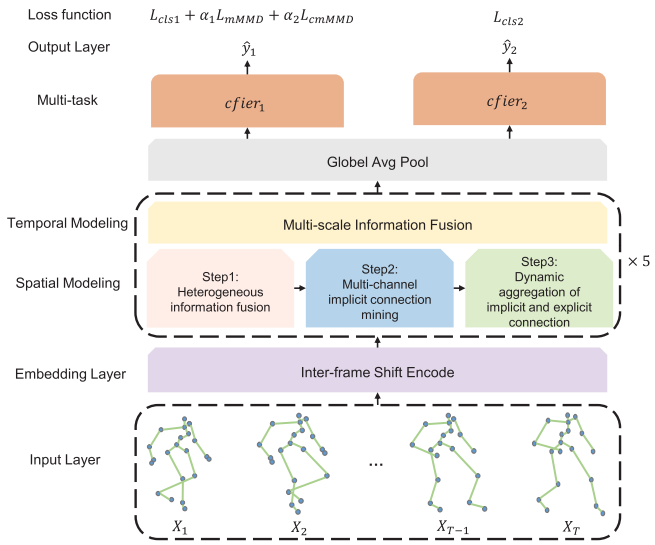


Fig. 2. The process of multi-channel Implicit connection mining.

the model to learn distinguishable features. Immediately after, Chen et al. [27] proposed STA-GCN, an implicit connection mining method for spatial modeling and a multiscale feature extraction and adaptive aggregation strategy for temporal modeling to address the above problem. The problem they have is that only limited kinds of implicit connections are mined using the multihead self-attention mechanism [37] in spatial modeling. Inspired by the observation of nature, there are multiple implicit connections related to emotions during human walking, so the method is inadequate for the excavation of implicit connections. The method, similar to most DL-GRE methods [23], [24], [25], [26], [27], [28], [29], did not exclude the effect of the subject's walking orientation on emotion classification.

III. Proposed Method

A. Pipeline

As shown in Fig. 2, a skeleton graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is composed of J joints according to the regularity of human biological chain connections [46], where $\mathcal{V}$ is the set of joints and $\mathcal{E}$ is the set of skeletons. We define $v_i \in \mathcal{V}$ as the i-th joint and the feature vector of v_i as $x_i \in \mathbb{R}^{d_1}$, where $1 \le i \le J$ and d_1 is the dimension of x_i. The skeleton graph sequence of T-frames is extracted from a segment of video of a human walking and is a representation of the human walking motion. We define v_i^t as the i-th joint of the t-th frame and the feature vector of v_i^t as $x_i^t \in \mathbb{R}^{d_1}$, where $1 \le i \le J, 1 \le t \le T$. $X_t = \left(x_1^t\ x_2^t \cdots x_J^t\right)^T$ is the feature matrix of the t-th frame of the skeleton graph sequence, where $X_t \in \mathbb{R}^{J \times d_1}$. $X = X_1 \oplus X_2 \cdots \oplus X_T$ is the feature tensor of the T frame skeleton graph sequence, where $X_t \in \mathbb{R}^{T \times J \times d_1}$, and $\oplus$ is the concatenate operation. A set of skeleton graphs $X_t \in \mathbb{R}^{T \times J \times d_1}$ containing T frames is fed into our AST-GCN in an ordered manner. Our AST-GCN is composed of an interframe shift encode (ISE) module, 4 encode blocks and a classification module. The purpose of the ISE module is to obtain the interframe shift $\overline{X} \in \mathbb{R}^{T-1 \times J \times d_1}$ of the joint point from the T frame skeleton graph $\overline{X} \in \mathbb{R}^{T \times J \times d_1}$, which is sensitive to the variation in the joint point shift. The role of the encode block is to model the gait skeleton sequence from a spatial perspective: 1) to obtain multiple implicit connections and 2) to mine context-sensitive emotional information in combination with explicit connections and to model the gait skeleton sequence from a temporal perspective, i.e., to focus on the motion pattern of joint points at multiple temporal scales. The interframe shift $\overline{X} \in \mathbb{R}^{T-1 \times J \times d_1}$ of the joint is fed into the first encode block. The output is $\overline{X}^{(1)} \in \mathbb{R}^{T-1 \times J \times d_2}$. Then, the output of the former layer is fed to the later encoding block in turn. The output of the l-th encode block is $\overline{X}^{(l)} \in \mathbb{R}^{T-1 \times J \times d_{l+1}}$. There are two types of classification modules, cfier_1 and cfier_2, which are classifiers for different emotion classification tasks: multiclass classification and multilabel classification, respectively. $\overline{X}^{(5)}$ is fed to cfier_1, cfier_2, and the output are $\hat{y}_1$, $\hat{y}_2$ respectively.

B. Interframe Shift Encode

1) Motivation: Inspired by natural observation, we observe human beings walking, for example, pedestrians on the road. Regardless of the pedestrian's orientation, we can correctly determine the pedestrian's emotional state. For example, if the pedestrian has their head down and is bent during walking, the relative shift of the head joint point and the waist joint point is stable regardless of the observer's viewpoint and the pedestrian's orientation. Therefore, we argue that he or she may be unhappy. Thus, the relative shift in the pedestrian's joints between frames is key part to judging the emotional state. **Advantages compared to previous methods:** Skeleton-based methods [25], [27], [29] extract features from gait sequences but fail to realize the importance of interframe shifts of joint points (as shown in Fig. 1d). Feature extraction directly from the gait sequence using the network is susceptible to the influence of subject orientation in the dataset when classifying emotions. In fact, the orientation of the subject is not related to his or her emotional state. The interframe shift encoding (ISE) module obtains the interframe shifts of joints, enabling the network to be more sensitive to changes in joints associated with emotions. It excludes the effect of the subject's walking orientation for emotion classification.

2) Operation: The feature tensor X of the T-frame skeleton graph sequence is fed to the ISE module. The output is $\overline{X} \in \mathbb{R}^{T-1 \times J \times d_1}$. This process can be represented as follows:

$$\overline{X} = (X_2 - X_1) \oplus (X_3 - X_2) \cdots \oplus (X_T - X_{T-1}), \quad (1)$$

where $\oplus$ is the concatenate operation, $\overline{X} \in \mathbb{R}^{T-1 \times J \times d_1}$.

C. In Spatial Modeling

1) Motivation: Inspired by natural observation, when we observe other human beings walking, for example, pedestrians on the road, we can always identify the emotional state based on specific actions. For example, when a pedestrian holds his or her head up and takes a long stride, we usually assume that he or she is happy. The head up is a widening of the angle formed by the chest joint as the apex and the head-chest and hip-chest as the bilateral sides. A large stride

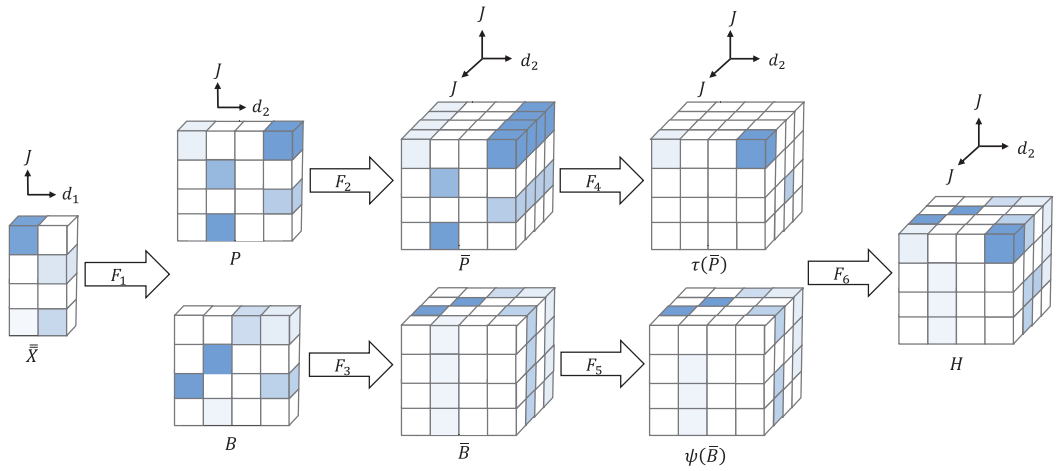


Fig. 3. The process of heterogeneous information fusion. F_1 represents the feature up-dimensioning of the joint points. F_2 and F_3 self-concatenate the tensor P and B in dimension 1 and dimension 0, respectively. F_4 and F_5 do $\tau(\cdot)$ and $\psi(\cdot)$ operations on $\overline{P}$, $\overline{B}$ respectively.F_6 is the summation of elements in the same position of the tensor.

is a widening of the angle formed by the crotch joint as the apex and the left foot-crotch and right foot-crotch as the bilateral sides. Therefore, we argue that there are a large number of special connections between human joint points that provide context-sensitive emotional information for gait emotion recognition; we refer to these as implicit connections. The variation in angle can be described by the strengths of the connections between joint points.

2) Advantages compared to previous methods: Prior work [27] mined a restricted variety of implicit connections (as shown in Fig. 1b). Inspired by the fact that humans can recognize the emotional state of pedestrians on the road through specific actions, human joints should have more implicit connections (as shown in Fig. 1c). Therefore, we propose a channel-level implicit connection inference method based on the self-attention mechanism [37], which infers a large number of implicit connections in spatial modeling, and the number of implicit connections is determined by the channel. In conclusion, our AST-GCN can extract more context-sensitive emotional information compared to other methods [27] in spatial modeling.

3) Operation: To facilitate the formulation of the following process, we predefine four symbolic operations.

1) We define $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_c}$ to represent a tensor, for $i = (i_1, \cdots, i_c)$ satisfying $1 \leq i_k \leq n_k$, $k = 1, \cdots, c$ and $\mathcal{A}(i) = \mathcal{A}(i_1, \cdots, i_c) \in \mathbb{R}^1$ to represent the value of a certain position in the tensor $\mathcal{A}$, which is also a tensor of rank 0. $\mathcal{A}(i_1, \cdots, i_{c-1}, \cdot) \in \mathbb{R}^{n_c}$ represents the rank 1 tensor of $\mathcal{A}$ in c dimensions, $\mathcal{A}(i_1, \cdots, i_{c-2}, \cdot, \cdot) \in \mathbb{R}^{n_{c-1} \times n_c}$ represents the rank 2 tensor of $\mathcal{A}$ in $[c-1, c]$ dimensions, and so on for any rank tensor of $\mathcal{A}$ in any dimension.
2) We define $\varsigma_l(\cdot)$ as the dimensionality ascension at the l-th dimensional position of the tensor. For example, suppose $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots n_l \times n_c}$; if $\mathcal{A}_1 = \varsigma_l(\mathcal{A})$, then $\mathcal{A}_1 \in \mathbb{R}^{n_1 \times \cdots \times 1 \times n_l \times n_c}$.
3) We define two operations $\psi(\cdot)$and $\tau(\cdot)$. For $\mathcal{A} \in \mathbb{R}^{n_1 \times n_1 \times n_2}$, $\psi(\mathcal{A})$ means that for $\forall 0 \leq m = n \leq n_1$ and $0 \leq o \leq n_2$ let $\mathcal{A}(m, n, o) = 0$. $\tau(\mathcal{A})$ means that for $\forall 0 \leq m \neq n \leq n_1$ and $0 \leq o \leq n_2$ let $\mathcal{A}(m, n, o) = 0$.
4) We define the operation $\oslash$; for $\mathcal{D}, \mathcal{E} \in \mathbb{R}^{d_1 \times d_1 \times d_2}$, let $\mathcal{D} \oslash \mathcal{E} = \mathcal{D}(1, \cdot, \cdot)\,\mathcal{E}(1, \cdot, \cdot) \oplus \mathcal{D}(2, \cdot, \cdot)\,\mathcal{E}(2, \cdot, \cdot) \oplus \cdots \oplus \mathcal{D}(d_1, \cdot, \cdot)\,\mathcal{E}(d_1, \cdot, \cdot)$.

To facilitate the description of spatial modeling, we divide the modeling process into three stages, namely, heterogeneous information fusion, multichannel implicit connection mining, and implicit connection and explicit connection dynamic aggregation.

4) Heterogeneous Information Fusion: First, we take the mean value of the feature $\overline{X} \in \mathbb{R}^{T-1 \times J \times d_1}$ obtained from the ISE module. The aim is to obtain the mean value of the interframe shift of the $T-1$ group of joint points in the T-frame skeleton sequence. This process can be represented as follows:

$$\overline{\overline{X}} = \frac{1}{T-1} \sum_{i=1}^{T-1} \overline{X_i}, \tag{2}$$

where $\overline{\overline{X}} \in \mathbb{R}^{J \times d_1}$.

Next, to learn more available information, the feature dimension of the joints is updimensioned from d_1 to d_2. We perform an updimensioning operation for the tensor in different dimensions to obtain the joint feature tensors P and B. This process can be represented as follows:

$$P = \varsigma_1\left(\overline{\overline{X}} W_p\right), \tag{3}$$

$$B = \varsigma_0\left(\overline{\overline{X}} W_b\right), \tag{4}$$

where $W_p, W_b \in \mathbb{R}^{d_1 \times d_2}$, $P \in \mathbb{R}^{J \times 1 \times d_2}$, $B \in \mathbb{R}^{1 \times J \times d_2}$.

We obtain the embedded tensor $\overline{P}$ and accepted tensor $\overline{B}$ by concatenating the joint feature tensors P and B in dimension 1 and dimension 0, respectively. We plan to assign the embedded tensor $\overline{P}$ to provide heterogeneous information for the accepted tensor $\overline{B}$. This process can be represented as follows:

$$\overline{P} = P \oplus \cdots \oplus P, \tag{5}$$

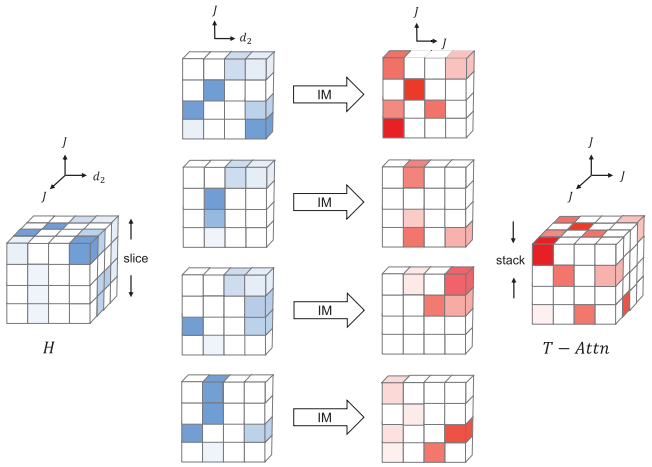


Fig. 4. The process of multi-channel Implicit connection mining.

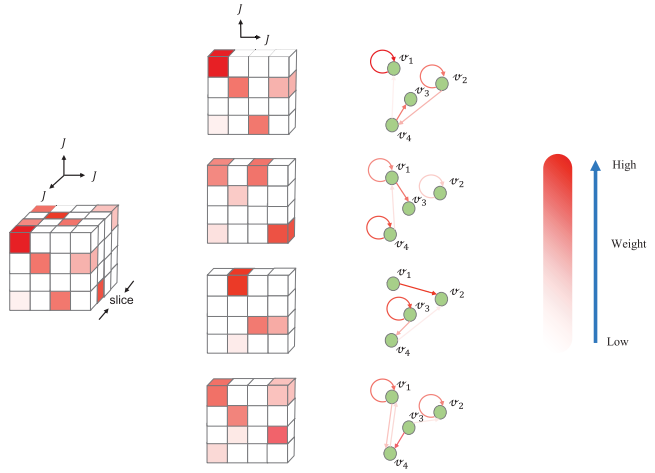


Fig. 5. The multi-channel tensor corresponds to multiple implicit connection relations of the joints.

$$\overline{B} = B \oplus \cdots \oplus B, \tag{6}$$

where $\overline{P} \in \mathbb{R}^{J\times J\times d_2}$, $\overline{B} \in \mathbb{R}^{J\times J\times d_2}$.

As shown in Fig. 3, our intention is to make embedded tensor $\overline{P}(m,n,\cdot) \in \mathbb{R}^{d_2}$ replace the tensor at the position of accepted tensor $\overline{B}(m,n,\cdot) \in \mathbb{R}^{d_2}$ for $\forall 0 \le m = n \le J$, that is, let embedded tensor $\overline{P}$ provide heterogeneous information to accepted tensor $\overline{B}$. In detail, the embedded tensor $\overline{P}$ provides J heterogeneous information to the accepted tensor $\overline{B}$, and the matrix we obtain after processing is called the mixture matrix H. This process can be represented as follows:

$$H = \tau\left(\overline{P}\right) + \psi\left(\overline{B}\right), \tag{7}$$

where $H \in \mathbb{R}^{J\times J\times d_2}$.

5) Multichannel Implicit Connection Mining: Before describing the next operations, let us review self-attention [37]; for $\mathfrak{C} \in \mathbb{R}^{d_1\times d_2}$, the attention matrix is:

$$attn = softmax\left(\frac{\left(\mathfrak{C}W_q\right)\left(\mathfrak{C}W_k\right)^T}{\sqrt{z}}\right), \tag{8}$$

where z is a scalar factor and $W_q, W_k \in \mathbb{R}^{d_2\times d_3}$.

As shown in Fig. 4, our intention is to mine J implicit connections using an attention-improved mechanism for a mixture matrix H that is rich in J heterogeneous information. Since J kinds of heterogeneous information are presented in the channel dimension, we call the implicit connection mined based on the mixture matrix H a channel-level implicit connection $Tattn$. As shown in Fig. 5, each channel of the tensor $Tattn$ represents a kind of implicit connection relation between the joint points. We now improve the conventional attention mechanism by expanding the input tensor to three dimensions; then, for $H \in \mathbb{R}^{J\times J\times d_2}$, the channel-level implicit connection $Tattn$ is:

$$Tattn = softmax\left(\frac{\left(H \oslash W_{q^*}\right)\left(H \oslash W_{k^*}\right)^T}{\sqrt{z}}\right), \tag{9}$$

where $W_{q^*}, W_{k^*} \in \mathbb{R}^{J\times d_2\times d_3}$ and $Tattn \in \mathbb{R}^{J\times J\times J}$.

6) Implicit Connection and Explicit Connection Dynamic Aggregation: We aggregate the channel-level implicit $Tattn$ and explicit connection A with the learnable parameter λ to obtain the final fusion connection F. This process can be represented as:

$$F = \lambda Tattn + \varsigma_0(A), \tag{10}$$

where $A \in \mathbb{R}^{J\times J}$, $F \in \mathbb{R}^{J\times J\times J}$.

Before describing the next operations, let us review vanilla GC [49], which is composed of two operations: 1) aggregating features of adjacent nodes and 2) linearly transforming aggregated features. This latent representation update rule for vanilla GC [49] can be formulated as:

$$H^{(l+1)} = \sigma\left(\hat{A}H^{(l)}W^{(l)}\right), \tag{11}$$

where $H^{(l)}$, $H^{(l+1)}$ denote the input and output data of the l-layer of the graph convolution, respectively, $W^{(l)}$ is a variable parameter, $\hat{A}$ is the matrix of the regularization of the adjacency matrix A, and $\sigma(\cdot)$ denotes the activation function.

In spatial modeling, feature learning is performed using graph convolution rules. Before feature learning, we must process the fusion connection F at the channel level to accommodate the dimensionality requirements of the T frame skeleton sequence $\overline{X}$. The process can be formulated as:

$$\overline{F} = max(F)_{dim=0} \in \mathbb{R}^{J\times J}, \tag{12}$$

$$\tilde{F} = \underbrace{\varsigma_0\left(\overline{F}\right) \oplus \cdots \oplus \varsigma_0\left(\overline{F}\right)}_{T-1}, \tag{13}$$

where $max(\cdot)$ represents the operation of obtaining the maximum value of a vector in a certain dimension.

To learn more information, $\overline{X} \in \mathbb{R}^{T-1\times J\times d_2}$ is updimensioned by a linear layer to obtain $\tilde{X} \in \mathbb{R}^{T-1\times J\times d_{l+1}}$. Notably, in the AST-GCN, each encoding block has a different upgrading strength for dimensionality; specifically, the l-th encoding block has an upgrade of d_{l+1}. Therefore, the output of the l-th encode block is $\overline{X}^{(l)} \in \mathbb{R}^{T-1\times J\times d_{l+1}}$. Finally, the graph convolutional latent layer update process of our

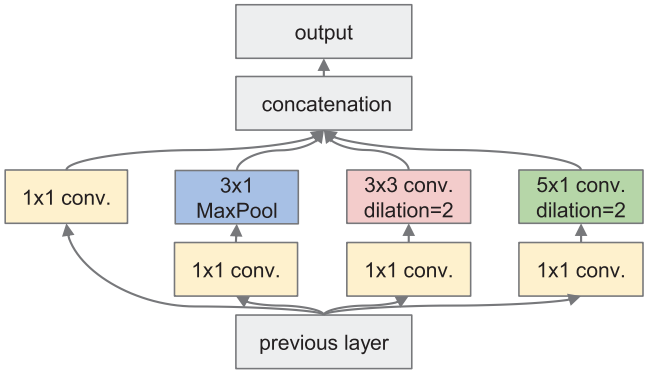


Fig. 6. Details of Temporal Modeling.

AST-GCN can be represented as:

$$\overline{X}^{(l+1)} = \sigma\left(\tilde{F} \oslash \overline{X}^{(l)}\right). \tag{14}$$

D. In Temporal Modeling

Regarding spatial modeling, [38] was introduced to allow the model to learn multiscale information about joint point motion. The module consists of three different convolutional branches, each focusing on different scales of joint point motion information. The outputs of the three convolutional branches are concatenated, and residual connections [50] are applied in this module. Details of the module are shown in Fig. 6.

E. Classification Modules and Loss Functions Based on Different Tasks

In this section, we focus on the design details of the classification modules and the loss functions for the two tasks.

1) Design Details of the Multiclass Classification Module: The tensor $\overline{X}^{(5)} \in \mathbb{R}^{T-1\times J\times d_6}$ generated by multiple encode blocks is fed into the classification module cfier_1. cfier_1 is composed of a global average pooling layer, a linear layer, a dropout layer, and three linear layers, whose details are shown in Table I.

2) Loss Function: Label smoothing [51] is a regularization method in machine learning that has been proven to be effective by many studies [52], [53], [54]; we apply it in classification loss. In supervised learning, when learning the mapping $X \longrightarrow Y$ between data and labels, the introduction of an information bottleneck (IB) [55] can restrict the encoding of the latent layer. This is equivalent to adding a constraint term to the learning, which can improve the generalization ability and robustness of the model. A learning objective-based IB is proposed in infoGCN [56] to handle skeleton-based tasks. We combine label smoothing [51] and learning objective-based IB [56] to form the complete loss function, which can be represented as:

$$L_{total} = L_{cls} + \alpha_1 L_{mMMD} + \alpha_2 L_{cmMMD}. \tag{15}$$

L_{cls} denotes the classification loss based on label smoothing, and L_{mMMD} and L_{cmMMD} denote the learning objective based on IB.

TABLE I

DESIGN DETAILS OF THE MULTICLASS CLASSIFICATION MODULE CFIER_1

Layer	Input Shape	Output Shape	Activation
Global Average Pooling	[8, 256, 19, 16]	[8, 256]	
Linear	[8, 256]	[8, 256]	Leaky ReLU [57]
Dropout [58]	Drop probability is 0.1		
Linear	[8, 256]	[8, 256]	Leaky ReLU [57]
Linear	[8, 256]	[8, 64]	Leaky ReLU [57]
Linear	[8, 256]	[8, 4]	Leaky ReLU [57]

TABLE II

DESIGN DETAILS OF THE MULTILABEL CLASSIFICATION MODULE CFIER_2

Layer	Input Shape	Output Shape	Activation
Global Average Pooling	[16, 64, 48, 21]	[16, 64]	
Linear	[16, 64]	[16, 32]	ELU [59]
Dropout [58]	Drop probability is 0.1		
Linear	[16, 32]	[16, 16]	ELU [59]
Dropout [58]	Drop probability is 0.1		
Linear	[16, 16]	[16, 4]	Sigmoid

3) Design Details of the Multilabel Classification Module: The tensor $\overline{X}^{(5)} \in \mathbb{R}^{T-1\times J\times d_1}$ generated from multiple encode layers is fed into the classification module cfier_2. cfier_2 is composed of a global average pooling layer, two sets of linear, activation and dropout layers, and a linear layer and activation layer, whose details are shown in Table II.

4) Loss Function: We assume that the multi-hot of the label of the gait sequence X is $y \in \{0,1\}^D$, denoted as $y = \left[p_1\ p_2\ \cdots\ p_D\right]^T$. $p_i = 1$ when there is the i-th emotion in X; otherwise, it is equal to 0. Since our classification problem is multilabel classification, we use weighted multilabel cross-entropy loss, which can be represented as:

$$Loss = -\sum_{d=1}^{D} w_d\, p_d \log \hat{p}_d. \tag{16}$$

F. Analysis of Graph Convolution

The main purpose of this section is to unify all graph convolution methods of skeleton-based methods [25], [27], [29] into a common graph convolution inference form. In this form, we analyze the model complexity of different graph convolutions and prove the maximum model complexity of our AST-GCN. According to [60], a higher model complexity will achieve better generalization performance in the case of a more complex data distribution.

In skeleton-based methods [25], [27], [29], the feature representation capability of graph convolutional neural networks [36] is used to model gait skeleton data. In the process of graph convolution inference, according to whether the connection relationship of the graph changes dynamically, it can be divided into static-based graph convolution [25], [29] and dynamic-based graph convolution. Static-based graph convolution [25], [29] is based on the graph structure of human biological chain connections [46], and the graph structure does not change during the inference process. The graph structure of dynamic-based graph convolution is generated based on the input gait data according to certain regulations, according to which dynamic-based graph convolution is divided into

single channel-based and multichannel-based dynamic graph convolution.

For a graph $G = (V, E)$, V is the set of nodes and E is the set of edges. For each node i, there is a feature x_i. The set of nodes can be represented by the matrix $X \in \mathbb{R}^{n\times d}$, where n denotes the number of nodes and d denotes the dimensionality of the feature vector of each node. The idea of graph convolution [36] is to let each node in the graph update according to the neighboring nodes, and the closer the relationship is, the stronger the influence of the neighboring nodes on that node. We define the connection relationship between nodes as $\mathbb{A}$ and represent the input and output data as H^l and H^{l+1}, respectively. The inference form of the graph convolution is represented as follows:

$$H^{l+1} = f\left(H^l, \mathbb{A}\right), \tag{17}$$

where $f(\cdot)$ is a mapping rule without parameters.

1) Static-Based Graph Convolution: In static-based graph convolution [25], [29], the connection relationship between nodes is the human biological chain connection relationship [46]. During the inference process, the connection relationship $\mathbb{A}$ between the nodes is constant and is described by the adjacency matrix A. The inference form is represented as:

$$H^{l+1} = f\left(H^l, A\right). \tag{18}$$

2) Single Channel-Based Dynamic Graph Convolution: It is represented by [27], who uses a modified multihead self-attention matrix to infer the connection relation (implicit connection) of the graph of the nodes based on gait data. The inference form is represented as follows:

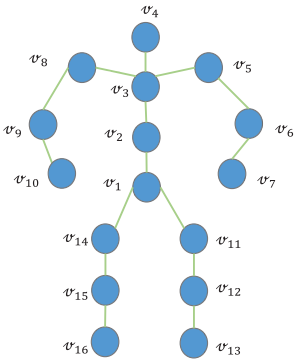
$$H^{l+1} = f\left(H^l, \sum_{i=1}^{N} softmax\left(\frac{X\theta_q^i \left(X\theta_k^i\right)^T}{\sqrt{d_k}}\right)\right). \tag{19}$$

Other things being equal, the connection relation $\mathbb{A}$ of the nodes dynamically changes in the single channel-based dynamic graph convolution compared to the static-based graph convolution. Two parameters $\theta_q^i \in \mathbb{R}^{d_1\times d_2}$ and $\theta_k^i \in \mathbb{R}^{d_1\times d_2}$ are added to the generation process of the graph structure, so the single channel-based graph convolution has greater model complexity. Single channel-based dynamic graph convolution has better generalization performance in the case of a more complex data distribution that of static-based graph convolution.

3) Multichannel-Based Dynamic Graph Convolution: In this paper, we propose a novel method for generating multichannel graph structures by mining J implicit connections for a mixture matrix H rich in J heterogeneity information using an improved attention mechanism. It is a multichannel dynamic graph convolution because the J kinds of heterogeneous information are expressed in the channel dimension. The inference form is expressed by Eq:

$$H^{l+1} = f\left(H^l, g\left(\overline{\overline{X}}, W_p, W_b, W_{q^*}, W_{k^*}\right)\right), \tag{20}$$

where $g(\cdot)$ denotes the mapping rules for graph structure generation, $W_p, W_b \in \mathbb{R}^{d_1\times d_2}$, $W_{q^*}, W_{k^*} \in \mathbb{R}^{J\times d_2\times d_3}$.



v_1	coccys	v_9	right elbow
v_2	spine	v_{10}	right hand
v_3	neck	v_{11}	left hip
v_4	head	v_{12}	left knee
v_5	left shoulder	v_{13}	left toe
v_6	left elbow	v_{14}	right hip
v_7	left hand	v_{15}	right knee
v_8	right shoulder	v_{16}	right toe

Fig. 7. The skeleton model with 16 joints.

We are not concerned with the exact size of d_i, since it is a hyperparameter and can be set at will. Our concern is that multichannel dynamic-based graph convolution requires approximately channel times more parameters than the single channel-based graph structure during the generation of the graph structure (connection relation of nodes $\mathbb{A}$). The generalization performance of multichannel-based dynamic graph convolution should be the best in the case of a complex data distribution.

IV. Experiments

A. Datasets

1) EGait: The EGait dataset is composed of 3D gait sequences with a sequence length of 240 frames. For each 3D gait sequence, a 16-joint skeleton model (as shown in Fig. 7) is used in the multiclass classification task and a 21-joint skeleton model is used in the multilabel classification task [27], [29]. Each gait is annotated by 10 domain experts. More details about EGait are available in [25] and [29].

2) EWalk: The EWalk dataset is composed of 3D gait sequences. For each 3D gait sequence, a 16-joint skeleton model is used (as shown in Fig. 7). Since the number of frames in a gait sequence is inconsistent, we loop each gait to make 240 frames, facilitating model input and retaining the temporal structure of the sequence. More details about EWalk are available in [25].

3) HEROES: For the HEROES dataset, 16 nonprofessional actors were recorded. This dataset contains 256 RGB gait sequences. The HEROES3D dataset contains 3D gait sequences extracted from HEROES dataset using the pose estimation software VideoPose3D [40]. Since the number of frames in a gait sequence is inconsistent, we loop each gait to make 240 frames, facilitating model input and retaining the temporal structure of the sequence. For each 3D gait sequence, a 16-joint skeleton model is used (as shown in Fig. 7). More details about HEROES are available in [39].

B. Task Introduction and Evaluation Metrics

Psychologists [22] have shown that there are multiple emotions expressed by humans during walking, each of which may occupy a certain percentage. If we take the dominant emotion as the human walking emotion and we design the agent to

recognize the dominant emotion, then this is a multiclass classification task [23], [24], [25], [26], [29]. If we design an agent to recognize all emotions, this is a multilabel classification task [25], [29]. For the multiclass classification task [23], [24], [25], [26], [29], we use $Accuracy = \frac{TP+TN}{TD}$ as the evaluation metric, where TP, TN and TD are the numbers of true positives, true negatives, and total data, respectively. For the multilabel classification task [25], [29], we use the average precision AP obtained for each class as an evaluation metric, which is the average of the precision values for all recall values between 0 and 1. We also use mAP as an evaluation metric, which is the average of the AP obtained in all classes.

C. Implementation Details

The platform for our experiments is an NVIDIA GeForce RTX 3090. The proposed method is implemented in the PyTorch learning framework. The hyperparameters of the experiments consist of two parts, the hyperparameters of the proposed method and the hyperparameters of the training framework.

1) The hyperparameters of the proposed method: The shape of the skeleton sequence that is fed into the model is $T \times J \times d_1$. $T = 240$ is the frame of the sequence, $J = 16$ is the number of joints of the skeleton, $d_1 = 3$ is the dimension of the 3D coordinates of the joints. The shape of the output tensor is output by the five encoder blocks $T \times J \times d_2$, $T \times J \times d_3$, $T \times J \times d_4$, $T \times J \times d_5$, and $T \times J \times d_6$. $d_2 = 64$, $d_3 = 64$, $d_4 = 128$, $d_5 = 256$, $d_6 = 256$.

2) The hyperparameters of the training framework: In the multiclass classification task, we divided the dataset into a training set, a validation set and a testing set with a ratio of 8:1:1. We set the batch size to 8, and the total number of training epochs is 1000. The

Adam optimizer [61] is used. The base learning rate is 0.001 and shrinks by $\frac{1}{10}$ at rounds 500, 750, 876, and 900. We set the momentum to 0.9 and weight decay to 9e-6. The hyperparameters of the loss function α_1 is 1e-4 and α_2 is 1e-1. In the multilabel classification task, we set the batch size to 16 and the weight decay to 5e-4. The other configurations are the same as those for the multiclass classification task.

We set four values for α_1 and α_2 and obtain 16 sets of data. The experimental results are shown in Table III. Based on the experimental results, we can draw two conclusions: 1) Both L_{mMMD} and L_{cmMMD} are effective. When using only the classification loss L_{cls} ($\alpha_1 = 0$, $\alpha_2 = 0$), the $Accuracy$ value is much smaller than the $Accuracy$ value when adding $L_{mMMD}(\alpha_1 > 0)$ or $L_{cmMMD}(\alpha_2 > 0)$; 2) The combination of L_{mMMD} and L_{cmMMD} is effective. When L_{mMMD} and L_{cmMMD} are used together ($\alpha_1 > 0$, $\alpha_2 > 0$), the $Accuracy$ is higher than the $Accuracy$ value when using L_{mMMD} or L_{cmMMD} alone ($\alpha_1 > 0$, $\alpha_1 = 0$) or ($\alpha_1 = 0$, $\alpha_2 > 0$).

D. Comparison With Other State-of-the-Art Methods

We compared our AST-GCN with the prior state-of-the-art methods [23], [24], [25], [26], [28], [29] on EGait [25], [29], EWalk [24], [25] and HEROES3D [39]. The experiments demonstrate that our AST-GCN achieves the best performance

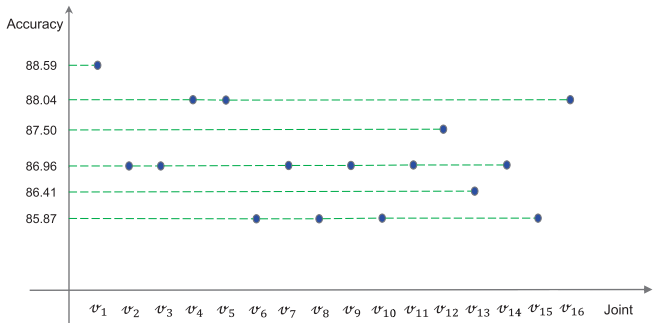


Fig. 8. Important joint points in the expression of gait emotions.

TABLE III
ABLATION EXPERIMENT WITH LOSS FUNCTION HYPERPARAMETERS

α_1 \ α_2	0	1.0×10^{-4}	1.0×10^{-3}	1.0×10^{-2}	1.0×10^{-1}
0	84.24	84.24	86.96	85.33	84.78
1.0×10^{-4}	85.87	**88.04**	86.41	86.41	86.96
1.0×10^{-3}	85.87	85.57	85.33	86.41	84.78
1.0×10^{-2}	84.78	85.57	86.96	85.33	**88.04**
1.0×10^{-1}	83.70	85.87	86.41	86.41	85.87

in two tasks, The multilabel classification task and multiclass classification task. The experimental results are shown in Tables IV and V. Based on the experimental results, we can draw two conclusions:

- Our AST-GCN outperforms prior state-of-the-art methods in two tasks on three datasets, achieving the best performance. On EGait, the largest public dataset, our AST-GCN outperforms the next best method [27] by approximately 2.17% in terms of $Accuracy$ in the multiclass classification task. On EGait, our AST-GCN outperforms the next best method [27] by approximately 1.3% in terms of the mAP in the multilabel classification task.
- It is worth mentioning that the interclass variation of our AST-GCN in terms of AP is small compared to that of other state-of-the-art methods [24], [25], [27], [29] in multilabel tasks. Because EGait [25], [29] has a long-tail distribution problem, it leads the model to prefer labels with a high percentage of the number of classes. Our AST-GCN uses weighted loss and weighted sampling to allow the model to learn different classes of samples in the same ratio.

E. Important Joint Points in the Expression of Gait Emotions

Gait is generated by the change in joint points according to certain rules during human walking. To explore which joint points change during motion, it is better for the agent to clearly identify the emotional state of the human. In other words, under the eyes of the agent, joint point movement changes have a strong correlation with the emotional state. We designed joint absence experiments to explore this issue. Our method uses a 16-joint-point model of the human skeleton, as shown in Fig. 7. We performed 17 experiments, and the first is to input the complete 16-joint-point data into the model. The next 16 experiments were performed with 15 joints as input,

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS IN THE MULTICLASS CLASSIFICATION TASK

Dataset	Method		*Accuracy*
EGait [25] [29]	Sequence-based method	LSTM (Vanilla) [24] (ISMAR-Adjunct'2019)	74.10
	Image-based method	TNTC [26] (ICASSP'2022)	79.52
		ProxEmo [23] (IROS'2020)	82.40
	Skeleton-based method	STEP [25] (AAAI'2020)	78.24
		G-GCSN [28] (ACCV'2021)	81.50
		STA-GCN [27] (ICME'2023)	85.87
		Our AST-GCN	**88.04** ↑ 2.17
EWalk [24] [25]	Skeleton-based method	STEP [25] (AAAI'2020)	77.14
		STA-GCN [27] (ICME'2023)	82.86
		Our AST-GCN	**85.71** ↑ 2.85
HEROES3D [39]	Skeleton-based method	STEP [25] (AAAI'2020)	69.23
		STA-GCN [27] (ICME'2023)	73.08
		Our AST-GCN	**76.92** ↑ 3.84

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS IN THE MULTILABEL CLASSIFICATION TASK

Method		*AP*				*mAP*
		Happy	Sad	Angry	Neutral	
Sequence-based method	LSTM (Vanilla) [24] (ISMAR-Adjunct'2019)	0.96	0.84	0.62	0.51	0.73
Skeleton-based method	STEP [25] (AAAI'2020)	0.97	0.88	0.72	0.52	0.77
	TAEW [29] (ECCV'2020)	0.98	0.89	0.81	0.71	0.84
	STA-GCN [27] (ICME'2023)	0.99	0.87	0.88	0.75	0.87
	Our AST-GCN	**0.911**	**0.844**	**0.907**	**0.870**	**0.883** ↑ 0.013

excluding a different joint for each experiment. The other training hyperparameters were kept constant. We observe that the experimental index decrease when joints are absent. The experimental results are shown in Fig. 8.

Based on the experimental results, we can draw two conclusions:

- From the experimental results in Fig. 8, the absence of joints v_6, v_8, v_{10} and v_{15} decreased the *Accuracy* value to the greatest extent. Therefore, the elbow, shoulder, hand and knee joints are important joint points that are more likely to allow the agent to clearly identify the emotional state of the humans. In fact, this is similar to what humans do when they observe the emotions of other humans. Psychologists [30], [62] have demonstrated that changes in limb movements at some joint points, such as the elbow, hand and shoulder, have more representational features and are easier for the observer to detect psychological states.
- Notably, in the v_1 joint absence experiment, the classification *Accuracy* of the model exceeds that of the 16-joint skeleton model. We consider that the motion of certain joint points may provide some redundant emotional information that interferes with the model's ability to identify the correct emotional state. With the absence of these joints, the model is able to focus on the joint movements associated with the emotional state.

TABLE VI
EFFECTIVENESS OF DIFFERENT COMPONENTS IN THE MULTICLASS CLASSIFICATION TASK

Component	SM		TM	ISE	*Accuracy*
	IC	EC			
Baseline		✓			76.08
S_1	✓	✓			84.24
S_2			✓		84.78
S_3				✓	82.07
S_4	✓	✓	✓		86.41
S_5	✓	✓		✓	87.50
S_6			✓	✓	85.33
S_7	✓	✓	✓	✓	**88.04**

F. Ablation Study

To validate the effectiveness of the three key components of our method in two tasks, we performed ablation experiments on the largest public dataset EGait [25], [29]. The experimental results are shown in Tables VI and VII. SM represents spatial modeling, TM represents temporal modeling,

TABLE VII
EFFECTIVENESS OF DIFFERENT COMPONENTS IN THE MULTILABEL CLASSIFICATION TASK

Component	SM		TM	ISE	AP				mAP
	IC	EC			Happy	Sad	Angry	Neutral	
Baseline		✓			0.837	0.727	0.711	0.709	0.746
S_1	✓	✓			0.806	0.846	0.792	0.887	0.832
S_2			✓		0.862	0.873	0.857	0.904	0.874
S_3				✓	0.816	0.779	0.857	0.848	0.825
S_4	✓	✓	✓		0.855	0.875	0.906	0.872	0.877
S_5	✓	✓		✓	0.878	0.806	0.914	0.838	0.859
S_6			✓	✓	0.861	0.891	0.872	0.888	0.878
S_7	✓	✓	✓	✓	**0.911**	**0.844**	**0.907**	**0.870**	**0.883**

and ISE represents the interframe shift encode module. The baseline's backbone network is composed of 4 layers of graph convolution, and the subsequent splicing classification modules cfier_2 and cfier_2 correspond to two tasks. For a fair comparison, the tensor shapes of the input and output of each layer of graph convolution of baseline are the same as those of each layer of graph convolution of our AST-GCN. S_i represents the combination of various components. Based on the experimental results, we can draw three conclusions:

- The experimental results demonstrate the effectiveness of our three key components. From S_1, S_2 and S_3, it can be concluded that temporal modeling is the most important from the perspective of enhancing the baseline, followed by spatial modeling and then the interframe shift encoding module.
- As shown by S_1, an implicit connection is important for gait emotion classification, which is able to mine context-sensitive emotion information.
- As shown by S_7, after combining the three key components, our AST-GCN achieves the best performance. As shown by S_4, S_5 and S_6, spatial modeling is the most important from the perspective of importance after the combination of components, followed by the interframe shift encode module and then temporal modeling. As shown by S_6, $Accuracy$ decreases the most in the absence of spatial modeling.

G. The Cross-Dataset Experiment

To illustrate the robustness of our AST-GCN, we performed cross-dataset experiments. The experimental results are shown in Table VIII. Based on the experimental results, we can draw two conclusions:

- The experimental results demonstrate that our model is robust. The multiclass classification task in our study involves categorizing data into four classes. From a probability perspective, for each sample, the probability of correct classification is 25%. In our cross-dataset experiments, we tested the model trained on the EGait/EWalk datasets. Although the $Accuracy$ does not exceed that of the model trained directly on the EGait/EWalk dataset, it far exceeds the $Accuracy$ from a probabilistic perspective. This means that our model learns common gait features that contribute to emotion classification. Empirically, our model will be more robust when the amount of data used for training is sufficiently large.
- We performed cross-dataset experiments with two open-source skeleton-based state-of-the-art methods. The experimental results demonstrate that our model has the best cross-dataset stability.
- The HEROES3D dataset [39] is not used for the cross-dataset experiment. This is because the emotion categories of this dataset [39] are different from those of the other two datasets, EGait [25], [29] and EWalk [24], [25]. Therefore, it is not scientific to use the HEROES3D dataset [39] for the cross-dataset experiment.

TABLE VIII
THE CROSS-DATASET EXPERIMENT

Train dataset	Method	Testing on EGait	Testing on EWalk
EGait	STA-GCN [27]	85.87	37.14
	STEP [25]	78.80	34.28
	Our AST-GCN	**88.04**	**42.85** ↑ 5.71
EWalk	STA-GCN [27]	57.60	82.86
	STEP [25]	55.98	77.14
	Our AST-GCN	**59.78** ↑ 2.18	**85.71**

H. Discussion on the Strength of Skeleton-Based Methods, Image-Based Methods and Sequence-Based Methods

Sequence-based, image-based and skeleton-based methods model gait data based on RNN, CNN and GCN architectures, respectively. We designed an experiment to discuss the strength of sequence-based methods, image-based methods and skeleton-based methods. We designed the simplest RNN, CNN and GCN models and allowed them to have approximate model sizes to test their classification accuracy. The experimental results are shown in Table IX. Based on the experimental results, we can draw two conclusions:

- The reason that skeleton-based methods are stronger than nonskeletal-based methods is because GCNs are more powerful than CNN and RNN architectures for feature representation of gait skeleton data.
- Gait skeleton data are essentially graph data. Joints can be viewed as vertices of the graph, and skeletons can be viewed as edges of the graph. GCN architectures are more powerful than non-GCN architectures in the representation of graph data. Therefore, the skeleton-based methods outperform the nonskeleton-based methods in terms of the basic model architecture.

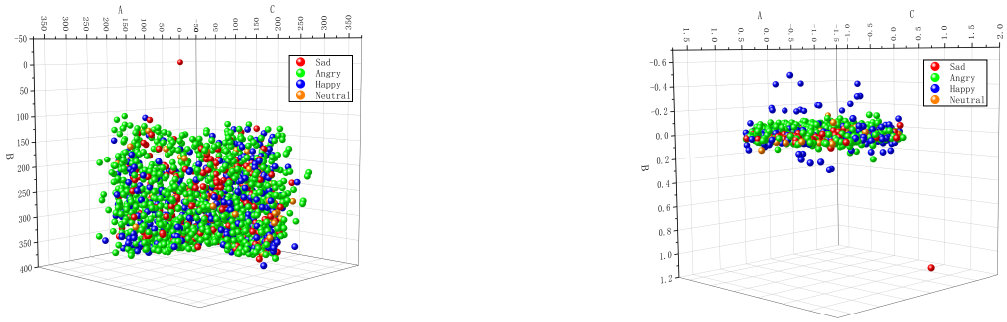


Fig. 9. Visual analysis of the importance of Interframe shift encode module. (a) The distribution of gait samples before input to the ISE module. (b) The distribution of gait samples after processing by the ISE module.

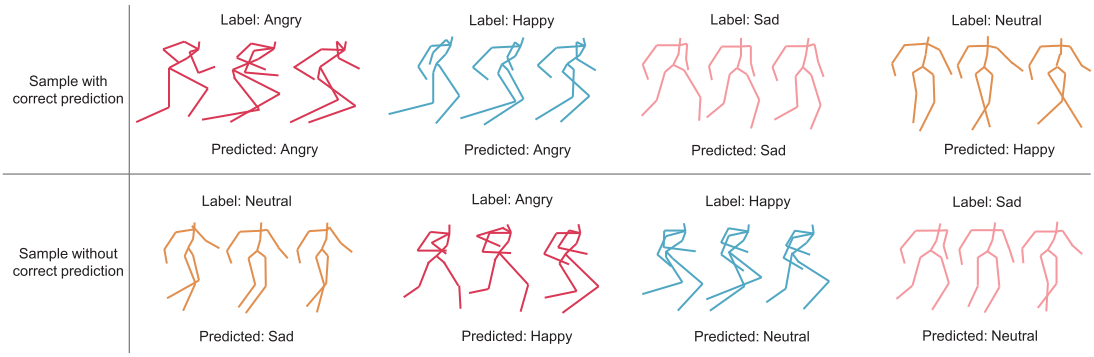


Fig. 10. Visual analysis of performance results in the multiclass classification task.

TABLE IX
DISCUSSES THE REPRESENTATIONAL ABILITY OF THE BASIS ARCHITECTURE OF SEQUENCE-BASED, SKELETON-BASED AND IMAGE-BASED METHODS FOR GAIT DATA

Method	Parameters (M)	*Accuracy*
LSTM	5.12	37.50
CNN	5.22	65.62
GCN	**4.85**	**75.00** ↑ 9.38

I. Comparison to the State-of-the-Art Methods in Terms of Parameters, Inference Time, and Runtime Speed

To demonstrate the superiority of our AST-GCN in terms of the parameters, inference time, and runtime speed, we calculated the above metrics for some state-of-the-art methods with publicly available code in the same environment, and the comparison experiments are shown in Table X. Based on the experimental results, we can draw two conclusions:

- In the multilabel classification task, although our AST-GCN performs worse than STEP [25] and LSTM (Vanilla) [24] in terms of inference time and runtime speed, our AST-GCN is far better than all previous state-of-the-art methods [24], [25], [27], [29] in terms of the parameters and mAP. Since our AST-GCN has a small number of parameters and high mAP, it is more suitable for deployment in real-world environments.
- In the multiclass classification task, our AST-GCN also shows superior performance. Although it is inferior to ProxEmo [23], STA-GCN [27] and LSTM [24] in terms of the parameters, it outperforms other previous state-of-the-art methods [23], [24], [25], [27] in terms of runtime speed, inference time and accuracy.
- We note that our AST-GCN has inferior parameters and inference time in the multiclass classification task compared to those of the multilabel classification task. This is because the encoder blocks of these two tasks use different hyperparameters. For example, ResNet50 [50] and ResNet18 [50] have the same design idea. However, the ResNet50 [50] convolutional kernel has a larger number of parameters compared to that of ResNet18 [50]. This phenomenon is also reflected in the two tasks using STA-GCN [27].

J. Visual Analysis of the Importance of the Interframe Shift Encode Module

To evaluate the performance of the ISE module more vividly, we visualized the distribution of gait samples before inputting them into the ISE module and after processing by the ISE module. Our approach is to perform $\frac{1}{T}\sum_{i=1}^{J}\sum_{t=1}^{T}x_i^t \in R^{d_1}$ for each joint feature $x_i^t \in R^{d_1}$ of gait X. The obtained result is used to represent the gait X. $d_1 = 3$ and $\frac{1}{T}\sum_{i=1}^{J}\sum_{t=1}^{T}x_i^t$ are points in the three-dimensional space. In this way, all the gaits in EGait can be represented as a scatter plot in three-dimensional space, as shown in Fig. 9a. Similarly, for the feature $\overline{X}$ output from the ISE module, $\frac{1}{T-1}\sum_{i=1}^{J}\sum_{t=2}^{T}\left(x_i^t - x_i^{t-1}\right) \in R^{d_1}$ is performed. The result is also a point in three-dimensional space. Therefore, the output data processed by the ISE module can be represented as a scatter plot in three-dimensional space, as shown in Fig. 9b.

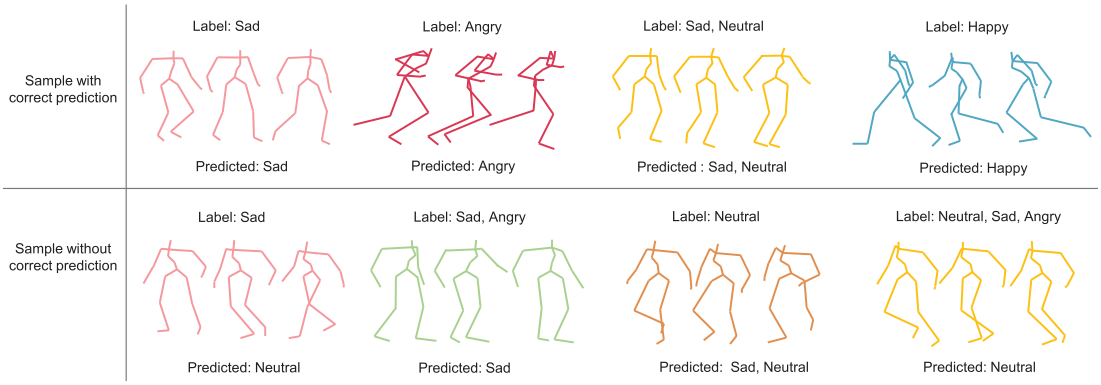


Fig. 11. Visual analysis of performance results in the multilabel classification task.

TABLE X
COMPARISON TO THE STATE-OF-THE-ART METHODS IN TERMS OF PARAMETERS, INFERENCE TIME, AND RUNTIME SPEED

Task	Methods	Parameters (M) ↓	Inference Time (ms) ↓	Speed (pcs/s) ↑
Multilabel classification	TAEW [29] (ECCV'2020)	40.43	23.294	42.93
	STEP [25] (AAAI'2020)	0.71	2.143	466.58
	STA-GCN [27] (ICME'2023)	0.07	9.299	107.53
	LSTM (Vanilla) [24] (ISMAR-Adjunct'2019)	0.31	3.337	299.67
	Our AST-GCN	**0.02**	**4.110**	**243.33**
Multiclass classification	STEP [25] (AAAI'2020)	0.71	2.918	342.65
	ProxEmo [23] (IROS'2020)	0.67	5.916	169.03
	STA-GCN [27] (ICME'2023)	0.51	24.59	40.67
	LSTM (Vanilla) [24] (ISMAR-Adjunct'2019)	0.31	4.162	240.24
	Our AST-GCN	**0.68**	**8.591**	**116.40**

By comparing the experimental results, we find that the ISE module makes the differences between classes more obvious, especially between the angry and happy samples, which have the largest amount of data. Therefore, the ISE module is effective for gait emotion classification.

K. Visual Analysis of the Performance Results in Two Tasks

As shown in Figs. 10 and 11, we visualized and analyzed the correctly classified and incorrectly classified samples using our AST-GCN in two tasks. The top row shows 4 gaits from the EGait, where the predicted emotions of our AST-GCN exactly matched the input labels. The bottom row shows 4 gaits where the predicted emotions did not match any of the input labels. Each gait is extracted from the three adjacent frames in the video. We carefully observed the samples with correct classification of each label and the samples with incorrect classification. According to the experimental results, two conclusions are drawn:

- In the two tasks, we observe more intense joint point movements for the happy and angry gaits compared to those for the neutral and sad gaits. This is also in line with psychological research [63] that happy and angry are high arousal emotions, while neutral and sad are low arousal emotions.
- In the multiclass classification task, by looking at the incorrectly classified samples, we find that sad gaits are often classified as neutral gaits, and angry gaits are often classified as happy gaits. This actually proves that our AST-GCN learns the relationship between the intensity of joint point movement and arousal strength. In addition, emotional states with different arousal levels may also be misclassified, e.g., a very small number of high arousal: happy gaits are misclassified as low arousal: neutral and sad gaits. We argue that during dataset collection, subjects showed different levels of happiness due to their different personalities. For example, when capturing happy gaits, shyer subjects may also have a relatively smaller range of motion. To solve the above problems, 1) Reasonable monitoring is carried out during data acquisition to prevent acquisition bias. 2) Multi-modal features are used for emotion recognition, combining gait, face, and other kinds of emotional information for emotion recognition. How to combine the advantages of features from different modalities for robust multimodal emotion recognition is an important research topic and our future research direction.
- In the multilabel task, by looking at incorrectly classified samples, we find that our AST-GCN often fails to identify multiple emotional states sufficiently. We believe that the

amount of data is relatively small and that the network cannot sufficiently learn discriminative features.

V. Conclusion

In this paper, an augmented spatial-temporal graph convolutional neural network is introduced to solve two problems respectively. The interframe shift encoding (ISE) module acquires interframe shifts of joints to make the network sensitive to changes in emotion-related joints regardless of the observer's viewpoint and the subject's walking orientation. A multi-channel implicit connection inference method learns more implicit connection relations related to emotions. It is worth mentioning that we unify current skeleton-based gait emotion recognition methods into a common framework that validates the most powerful feature representation capabilities of our AST-GCN from a theoretical perspective. In addition, we extended the skeleton-based gait dataset using posture estimation software. Experiments demonstrate that our AST-GCN outperforms the state-of-the-art methods on three datasets in two tasks.

References

[1] E. A. Veltmeijer, C. Gerritsen, and K. V. Hindriks, "Automatic emotion recognition for groups: A review," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 89–107, Jan. 2023.

[2] K. Feng and T. Chaspari, "A review of generalizable transfer learning in automatic emotion recognition," *Frontiers Comput. Sci.*, vol. 2, p. 9, Feb. 2020.

[3] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.

[4] M. Dubey and L. Singh, "Automatic emotion recognition using facial expression: A review," *Int. Res. J. Eng. Technol.*, vol. 3, no. 2, pp. 488–492, 2016.

[5] S. Xie, H. Hu, and Y. Chen, "Facial expression recognition with two-branch disentangled generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2359–2371, Jun. 2021.

[6] C. Wang, J. Xue, K. Lu, and Y. Yan, "Light attention embedding for facial expression recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1834–1847, Apr. 2022.

[7] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Netw.*, vol. 18, no. 4, pp. 389–405, 2015.

[8] M. G. Cherry, I. Fletcher, H. O'Sullivan, and T. Dornan, "Emotional intelligence in medical education: A critical review," *Med. Educ.*, vol. 48, no. 5, pp. 468–478, May 2014.

[9] X. Sun, C. Zhang, and L. Li, "Dynamic emotion modelling and anomaly detection in conversation based on emotional transition tensor," *Inf. Fusion*, vol. 46, pp. 11–22, Mar. 2019.

[10] Y. Gu, H. Yan, X. Zhang, Y. Wang, Y. Ji, and F. Ren, "Toward facial expression recognition in the wild via noise-tolerant network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2033–2047, May 2023.

[11] B. Li, C. Zhu, S. Li, and T. Zhu, "Identifying emotions from non-contact gaits information based on Microsoft kinects," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 585–591, Oct. 2018.

[12] M. Ren, X. Huang, J. Liu, M. Liu, X. Li, and A.-A. Liu, "MALN: Multimodal adversarial learning network for conversational emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6965–6980, Nov. 2023.

[13] V. Delvigne, H. Wannous, T. Dutoit, L. Ris, and J.-P. Vandeborre, "PhyDAA: Physiological dataset assessing attention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2612–2623, May 2022.

[14] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis. Support Syst.*, vol. 115, pp. 24–35, Nov. 2018.

[15] V. A. Petrushin, "Creating emotion recognition agents for speech signal," in *Socially Intelligent Agents*. Boston, MA, USA: Springer, 2002, pp. 77–84.

[16] P. Ekman and W. V. Friesen, "Head and body cues in the judgment of emotion: A reformulation," *Perceptual Motor Skills*, vol. 24, no. 3, pp. 711–724, Jun. 1967.

[17] J.-M. Fernández-Dols and M.-A. Ruiz-Belda, "Expression of emotion versus expressions of emotions: Everyday conceptions about spontaneous facial behavior," in *Everyday Conceptions of Emotion*. Dordrecht, The Netherlands: Springer, 1995, pp. 505–522.

[18] R. E. Nisbett and T. D. Wilson, "Telling more than we can know: Verbal reports on mental processes," *Psychol. Rev.*, vol. 84, no. 3, p. 231, 1977.

[19] S. Xu et al., "Emotion recognition from gait analyses: Current research and future directions," *IEEE Trans. Computat. Social Syst.*, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9969993

[20] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 264–284, Jan. 2023.

[21] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "GaitGraph: Graph convolutional network for skeleton-based gait recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2314–2318.

[22] S. Strike, C. Mohiyddini, and A. Carlisle, "Effect of emotional state on walking gait," *Gait Posture*, vol. 30, p. S123, Nov. 2009.

[23] 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 8200–8207.

[24] 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 *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct*, Oct. 2019, pp. 395–399.

[25] 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 *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 2, pp. 1342–1350.

[26] C. Hu, W. Sheng, B. Dong, and X. Li, "TNTC: Two-stream network with transformer-based complementarity for gait-based emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 3229–3233.

[27] C. Chen and X. Sun, "STA-GCN: Spatial temporal adaptive graph convolutional network for gait emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2023, pp. 1385–1390.

[28] Y. Zhuang, L. Lin, R. Tong, J. Liu, Y. Iwamoto, and Y.-W. Chen, "G-GCSN: Global graph convolution shrinkage network for emotion perception from gait," in *Computer Vision—ACCV 2020*. Cham, Switzerland: Springer, 2021, pp. 46–57.

[29] U. Bhattacharya et al., "Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 145–163.

[30] M. Karg, K. Kühnlenz, and M. Buss, "Recognition of affect based on gait patterns," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 40, no. 4, pp. 1050–1061, Aug. 2010.

[31] A. Crenn, R. A. Khan, A. Meyer, and S. Bouakaz, "Body expression recognition from animated 3D skeleton," in *Proc. Int. Conf. 3D Imag. (IC3D)*, Dec. 2016, pp. 1–7.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[33] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[34] L. R. Medsker and L. C. Jain, "Recurrent neural networks," *Des. Appl.*, vol. 5, pp. 64–67, Dec. 2001.

[35] D. Gupta, "Architecture of convolutional neural networks (CNNS) demystified," *Analytics Vidhya*, 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/06/architecture-of-convolutional-neural-networks-simplified-demystified/

[36] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: A comprehensive review," *Comput. Social Netw.*, vol. 6, no. 1, pp. 1–23, Dec. 2019.

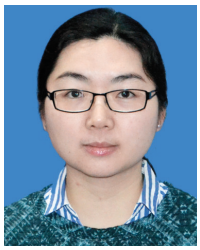
[37] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

[38] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 140–149.

[39] I. Mannocchi, K. Lamichhane, M. Carli, and F. Battisti, "HEROES: A video-based human emotion recognition database," in *Proc. 10th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Sep. 2022, pp. 1–6.
[40] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7745–7754.
[41] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013.
[42] K. M. Leung, "Naive Bayesian classifier," *Polytech. Univ. Dept. Comput. Sci./Finance Risk Eng.*, vol. 2007, pp. 123–156, Nov. 2007.
[43] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 15–33, Jan. 2013.
[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[45] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, Jun. 2019.
[46] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Comput. Vis. Image Understand.*, vol. 158, pp. 85–105, May 2017.
[47] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–15.
[48] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image Vis. Comput.*, vol. 60, pp. 4–21, Apr. 2017.
[49] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[52] Y. Xu, Y. Xu, Q. Qian, H. Li, and R. Jin, "Towards understanding label smoothing," 2020, *arXiv:2006.11653*.
[53] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
[54] C.-B. Zhang et al., "Delving deep into label smoothing," *IEEE Trans. Image Process.*, vol. 30, pp. 5984–5996, 2021.
[55] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, *arXiv:1703.00810*.
[56] H.-G. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, "InfoGCN: Representation learning for human skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20154–20164.
[57] A. L. Maas et al., "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1. Atlanta, GA, USA, 2013, p. 3.
[58] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
[59] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.
[60] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[62] D. Bernhardt and P. Robinson, "Detecting affect from non-stylised body motions," in *Affective Computing and Intelligent Interaction*. Cham, Switzerland: Springer, 2007, pp. 59–70.
[63] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. Cambridge, MA, USA: MIT Press, 1974.

Chuang Chen (Graduate Student Member, IEEE) was born in 1999. He received the B.E. degree in software engineering from the Hebei University of Economics and Business, Shijiazhuang, China, in 2021. He is currently pursuing the M.E. degree with the School of Artificial Intelligence, Anhui University, Hefei, China. His research interests include gait emotion recognition and affective computing.

Xiao Sun (Member, IEEE) was born in 1980. He received the M.E. degree from the Department of Computer Sciences and Engineering, Dalian University of Technology, Dalian, China, in 2004, and the dual Ph.D. degree from the University of Tokushima, Tokushima, Japan, in 2009, and the Dalian University of Technology, in 2010. His field of study was natural language processing. He is currently a Professor with the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, Hefei University of Technology, Hefei, China. He is also a Researcher with the Hefei Comprehensive National Science Center, Institute of Artificial Intelligence. His research interests include affective computing, natural language processing, machine learning, and human–machine interaction.



Zhengzheng Tu received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2007 and 2015, respectively. She is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. Her current research interests include computer vision and deep learning.

Meng Wang (Fellow, IEEE) received the B.E. and Ph.D. degrees from the Special Class for the Gifted Young, Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2008, respectively. He is a Professor with the Hefei University of Technology, Hefei. He has authored or coauthored more than 200 book chapters, journal articles, and conference papers in these areas. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He was a recipient of the ACM SIGMM Rising Star Award in 2014. He is an Associate Editor of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.