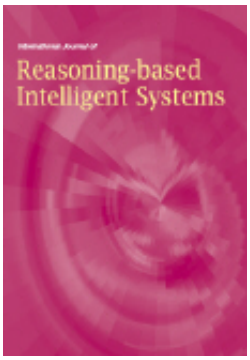


International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

Dual-stream spatiotemporal fusion with dynamic feature mapping for gait-based identity recognition

Binge Quan, Beibei Zhang

DOI: 10.1504/IJRIS.2026.10079191

Article History:

Received:	27 April 2026
Last revised:	25 May 2026
Accepted:	25 May 2026
Published online:	02 July 2026

Dual-stream spatiotemporal fusion with dynamic feature mapping for gait-based identity recognition

Binge Quan and Beibei Zhang*

Anhui Police College,
Hefei, 238000, China
Email: zbb_19920403@sina.com
Email: qbexlyx12345@sina.com
*Corresponding author

Abstract: Gait recognition offers the advantage of being contactless, but it is susceptible to interference from clothing and viewing angles. The key lies in extracting stable individual motion features. To address this, we propose a dynamic feature mapping framework based on skeletal keypoints. Unlike existing skeleton-based methods that directly use raw joint coordinates, our framework incorporates three key innovations: 1) an explicit kinematic feature mapping module that transforms raw coordinates into joint angles, angular velocities, and angular accelerations; 2) a dual-stream spatio-temporal graph convolution architecture that separately processes positional and kinematic features; 3) a frame-wise spatial attention mechanism that dynamically re-weights body parts according to input conditions. This framework employs dual-stream spatio-temporal convolutional networks to fuse features such as joint positions, angles, and angular velocities, and introduces attention mechanisms to adaptively weight the contributions of different body parts. On the CASIA-B dataset, the accuracy reached 95.8%, an improvement of 12.6 percentage points over GaitGraph; on Gait3D, the Rank-1 accuracy score reached 83.7%, an improvement of 25.0 percentage points over gait global-local model (GaitGL). The results demonstrate that dynamic features effectively capture differences in walking patterns and are robust to variations in appearance, providing a new approach for model-based gait recognition.

Keywords: gait recognition; skeletal landmarks; kinematic features; identity authentication.

Reference to this paper should be made as follows: Quan, B. and Zhang, B. (2026) 'Dual-stream spatiotemporal fusion with dynamic feature mapping for gait-based identity recognition', *Int. J. Reasoning-based Intelligent Systems*, Vol. 18, No. 17, pp.67–82.

Biographical notes: Binge Quan is a Lecturer at Anhui Police College, China. She obtained her Bachelor's in Biotechnology in Tianshui Normal University in 2016 and Master's in Public Security Technology and Engineering in China Criminal Police College in 2019. She published three papers. Her research interests include forensic medicine, forensic evidence and genetics.

Beibei Zhang is an Associate Professor at Anhui Police College, China. She holds a Bachelor's in Criminal Science and Technology. Her main research field is criminal science and technology, and her specialty is trace inspection.

1 Introduction

Traditional biometric features such as fingerprints, faces, and irises have been widely used in various identity verification scenarios. However, these modalities usually require the identified object to actively cooperate within a close range, which is an inherent limitation that makes them difficult to apply in long-distance identity verification tasks such as security monitoring and tracking of criminal suspects. In contrast, gait, as a behavioural biometric feature, has unique advantages such as non-contact, difficulty in disguise, and the ability to be collected at a long distance (Lee et al., 2024). Gait reflects the movement pattern of an individual while walking, which is controlled by the central nervous system and shaped by multiple factors such as skeletal structure, muscle strength distribution, and neural control strategies (Zheng et al., 2024). Unlike face recognition which depends on clear frontal views and adequate lighting, or voice which is easily masked by environmental noise, gait is produced involuntarily and remains detectable even in challenging conditions with oblique angles, low resolution, or partial body occlusion, thereby offering a distinct advantage for surveillance applications where subjects do not actively cooperate. This inherent robustness makes gait a superior choice for real-world long-distance identity authentication. Theoretically, each individual's walking pattern has a certain uniqueness, making it an ideal choice for long-distance identity verification.

Regarding the core issue of 'how to extract stable identity features from walking videos', researchers have developed two technical routes. The appearance-based method takes the silhouette of the human body contour as the input and recognises through gait energy maps or deep features (Xu et al., 2024). It has achieved excellent

performance on controlled datasets. However, these methods have inherent flaws: the contour shape incorporates a large amount of covariate information unrelated to identity, such as clothing, carried items, and changes in perspective. Once these factors change, the recognition performance drops sharply. The model-based method obtains the key points of the human body's skeleton through pose estimation, which is theoretically more robust to appearance changes (Yin and Li, 2024). To contextualise our contribution, we critically examine representative skeleton-based methods. PoseGait combines 3D skeletons with handcrafted features to address viewpoint and clothing interference but remains limited by the descriptive power of handcrafted features. GaitGraph introduces GCN to model human skeleton topology yet still relies primarily on raw joint coordinates, which, as signal processing theory indicates, mix motion information with camera perspective and body scale. Neither method explicitly incorporates joint-angle dynamics or angular velocities – features that biomechanical studies have shown to be highly individual-specific. This limitation motivates our explicit kinematic feature mapping. Moreover, while attention mechanisms have been explored in gait recognition (e.g., GaitGL's global-local attention), their application in skeleton-based methods to dynamically adjust focus across different body parts under varying input conditions – such as shifting attention to upper limbs when lower limbs are occluded – remains underexplored. Furthermore, existing multi-scale temporal modelling in skeleton-based methods largely relies on fixed-scale temporal convolutions that fail to capture gait's multi-scale dependencies spanning both rapid swing phases and slower stride rhythms. In recent years, the maturity of technologies such as open pose estimation library (OpenPose) and high-resolution network (HRNet) has made skeleton extraction possible (Salcedo, 2024b). However, most existing methods directly input the joint coordinate sequence into the network, which has low-level representations that contain redundant information such as absolute positions and perspectives, and fail to explicitly express the dynamic characteristics during the motion process. As a result, the recognition performance still lags behind the appearance-based methods.

Based on the above understanding, this paper proposes a skeleton-based kinetic dynamics mapping network (SKDMap-Net), aiming to extract more discriminative kinetic features from the bone sequence for identity recognition. The innovation and contribution of this method mainly lie in the following aspects:

- At the methodological level, a dual-stream spatiotemporal graph convolution architecture separately processes joint positions and dynamic features (angles, angular velocities, accelerations), mapping coordinates to a motion-reflective space to avoid single-feature limitations.
- At the feature fusion level, an attention mechanism adaptively weights body parts by input conditions – emphasising lower limbs during normal walking and shifting to upper limbs under occlusion – ensuring stable recognition.
- At the temporal modelling level, multi-scale temporal convolution with parallel kernels of different sizes captures both short-term local movements and long-term global patterns, enhancing temporal representation.

The subsequent structure of this article is as follows: Section 2 reviews the related studies; Section 3 elaborates on the methodology of SKDMap-Net; Section 4 introduces the experimental setup; Section 5 presents the results and analysis; Section 6 discusses the limitations and future directions; Section 7 summarises the entire article.

2 Literature review

2.1 Appearance-based gait recognition methods and their limitations

The appearance-based gait recognition method has long held the dominant position in this field. These methods take the silhouette of the human body as the input and identify identities by extracting gait energy maps, frequency domain features, or depth features (He et al., 2025). From the early gait energy maps to the recent deep learning methods such as gait as a set (GaitSet), gait part-based model (GaitPart), and gait global-local model (GaitGL), the appearance-based methods have continuously set new performance records on controlled datasets like CASIA-B (Hasan et al., 2024) and OU-ISIR (Iwama et al., 2012). The success of these methods lies in the fact that the silhouette contains rich body shape information, and the body shape itself has strong individual variability. However, there is an essential defect in the appearance-based methods: the silhouette shape not only contains identity information but also a large amount of covariate information unrelated to identity. When pedestrians change their clothes, add or remove backpacks, or change their viewing angle, the silhouette shape undergoes significant changes, resulting in a sharp decline in recognition performance (Jyothi et al., 2025). Studies have shown that this performance decline is particularly obvious in real outdoor scenarios, becoming the main bottleneck restricting the practical application of the methods (Uddin et al., 2010). Although existing review studies have systematically summarised the technological evolution of appearance-based methods, they mostly focus on the improvement of the methods themselves and lack sufficient exploration of the intrinsic mechanism of covariate interference.

2.2 Model-based gait recognition methods and their limitations

The model-based approach describes the walking posture through the key points of human bones (Wang et al., 2010). Theoretically, it has stronger robustness to external factors such as clothing and lighting. The key points of human

bones describe the spatial positions of joints rather than the appearance morphology of the body surface. Therefore, when pedestrians are wearing loose clothing or carrying backpacks, as long as the key points can still be accurately estimated, the bone representation can filter out the interference caused by appearance changes to a certain extent. In recent years, with the maturity of pose estimation techniques such as OpenPose, HRNet, and accurate and real-time pose estimation system (AlphaPose), it has become possible to accurately extract two-dimensional or three-dimensional key points of the skeleton from red, green, blue (RGB) videos (Tao et al., 2025). However, the existing model-based methods still lag behind the appearance-based methods in recognition performance. Analysing the reasons, most studies directly use the sequence of joint coordinates as input and model the temporal dependence and spatial structure through recurrent neural networks, temporal convolutional networks, or graph convolutional networks (GCNs) (Roemer and Buehler, 2025). The implicit assumption of this approach is that the temporal evolution of joint coordinates already contains sufficient identity information. However, from the perspective of signal processing, the coordinate sequence is a low-level representation that not only contains motion information but also mixes irrelevant factors such as absolute position, camera perspective, and human scale (Guo et al., 2024).

2.3 Skeletal keypoint representation and exploration of kinematic characteristics

Biomechanical studies have shown that an individual's walking pattern has a unique dynamic signature (Gonzalvez-Gallego et al., 2015). Characteristics such as the curve morphology of joint angles over time, the peak moments of angular velocities, the phase coordination relationships between joints, and the patterns of acceleration changes are closely related to an individual's neuromuscular control strategy and contain potential individual specificity. In the clinical gait analysis field, these dynamic parameters have been widely used for the identification of abnormal gait and disease diagnosis (Kawakami et al., 2024). For instance, parameters such as step width and hip joint angles have been proven to be biomarkers for certain neurodegenerative diseases. However, in the research of gait recognition, the exploration of dynamic features is still insufficient. Existing methods based on bones mostly focus on the modelling of spatial structures and often use simple convolution or recurrent structures to process temporal information, failing to fully consider the multi-scale temporal characteristics of gait (Carboni et al., 2025). Some studies have attempted to introduce attention mechanisms to enhance feature expression, but there is still a lack of systematic exploration on how to effectively integrate positional features and dynamic features.

2.4 Identified research gaps

In summary, the current research has obvious gaps in the following aspects: First, at the feature representation level, model-based methods mostly directly use joint coordinate sequences, failing to fully exploit the identity discrimination information contained in joint angles, angular velocities, angular accelerations, and other dynamic features, which have been proven to have individual specificity in biomechanical research. Second, at the feature fusion level, the discriminative contributions of different body parts under different walking conditions vary, and existing methods lack an adaptive weighting mechanism to dynamically adjust the attention paid to different joints. Third, at the temporal modelling level, gait as a periodic movement has multi-scale temporal dependency characteristics, but existing methods mostly adopt single-scale temporal convolution, making it difficult to simultaneously capture short-term local movement patterns and long-term global movement patterns. Fourth, at the multi-feature collaboration level, there is a complementary relationship between joint position features and dynamic features, but there is still a lack of systematic research schemes on how to effectively fuse these two types of features and make them work together. These research gaps point to a common issue: How to extract more discriminative dynamic features from the skeletal sequence and achieve effective multi-feature fusion through reasonable architecture design, thereby improving the performance of model-based gait recognition. Positioning our work within the literature: in the appearance vs. model dichotomy, we align with model-based methods for their theoretical robustness to covariates; however, unlike existing model-based approaches that rely heavily on raw joint coordinates, we introduce explicit kinematic feature mapping to incorporate biomechanically validated motion descriptors. In the static vs. dynamic feature paradigm, we go beyond static joint-angle features by also incorporating velocity and acceleration profiles, enabling the model to capture not only posture geometry but also neuromuscular timing and force patterns. This dual-stream design produces rich motion signatures that are more discriminative than either static or dynamic features alone.

3 Methodology

3.1 Overview of the proposed framework

In response to the limitation of existing model-based gait recognition methods that fail to fully utilise the dynamic features, this paper proposes a SKDMap-Net. The overall architecture of this framework is shown in Figure 1, and it mainly consists of four core modules: the skeletal graph construction module, the dynamic feature mapping module, the dual-stream spatiotemporal graph convolution module, and the multi-scale temporal modelling and classification module. Firstly, the original skeletal key points extracted from the input video sequence are preprocessed and then a skeletal spatiotemporal graph is constructed, which includes

spatial adjacency relationships and temporal connection relationships. Then, the dynamic feature mapping module converts the original joint coordinates into multiple types of features such as joint angles, angular velocities, angular accelerations, relative coordinates, and bone length ratios, forming input streams of position and angle. The dual-stream spatiotemporal graph convolution module performs graph convolution and temporal convolution on the position features and angle features respectively, and introduces an attention mechanism in each layer to adaptively weight the importance of different joints. Finally, the multi-scale temporal modelling module captures the multi-scale temporal dependencies in the gait sequence using different-scale convolution kernels, and after global pooling, it obtains fixed-dimensional gait feature variables for identity classification and triplet metric learning. Compared with standard ST-GCN pipelines that apply a single graph convolution to raw joint coordinates, our framework introduces three distinct innovations. First, we replace raw coordinate input with a dual-stream design that separately processes positional features and explicitly computed kinematic features (joint angles, angular velocities, angular accelerations). Second, we incorporate a frame-wise spatial attention mechanism that computes attention weights independently for each temporal frame, enabling the model to adapt its focus to different body parts across gait phases – an explicit design absent in conventional GCN attention applications for gait recognition. Third, we employ multi-scale temporal convolutions with kernel sizes 1, 3, and 5, as opposed to fixed-scale convolutions, to jointly capture both short-duration swing-phase dynamics and longer stride rhythm patterns. Compared with existing dual-stream architectures that typically fuse features at the input or output level only, our layer-wise learnable fusion coefficient $\lambda(l)$ allows the network to balance positional and kinematic contributions at each layer independently.

Figure 1 Overview of the SKDMap-Net architecture (see online version for colours)

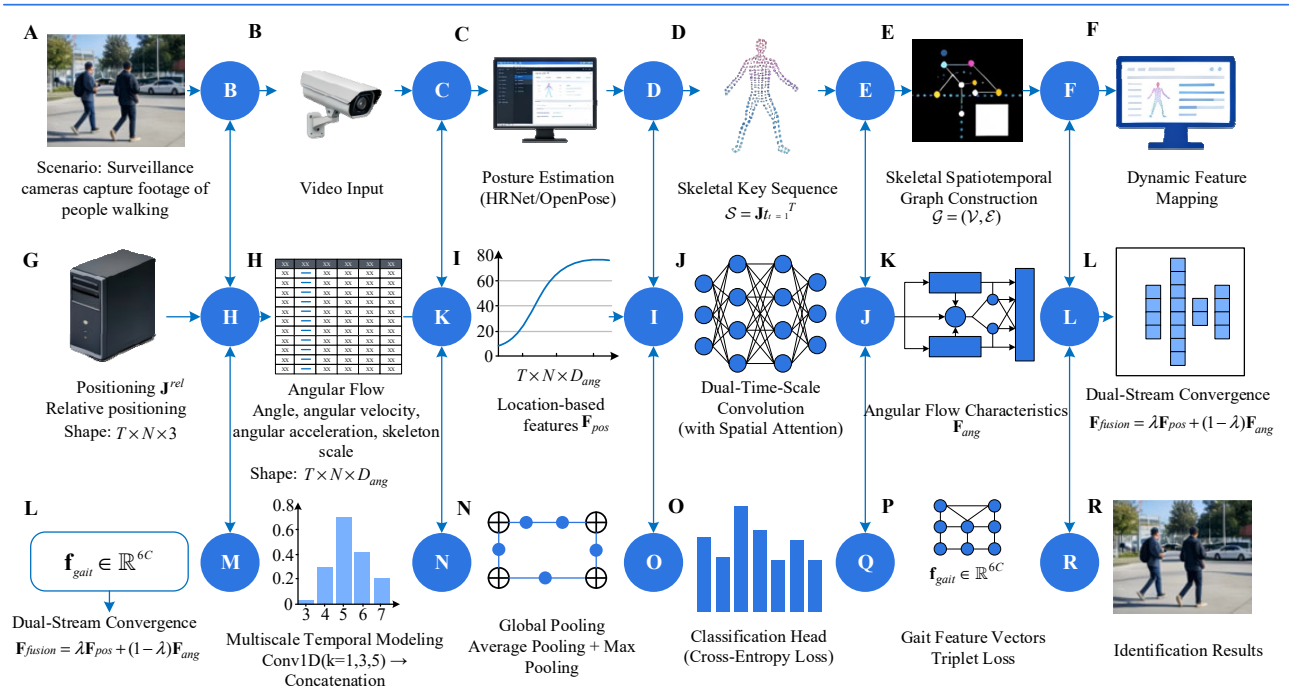


Table 1 Key symbols and their descriptions

Symbol	*Meaning/description*	*Symbol*	*Meaning/description*
$\mathbf{J}_t\in\mathbb{R}^{N\times 3}$	3D coordinates of all keypoints at frame	$\mathbf{F}_{pos}$	Feature output from the position stream
$(x_{t,i}, y_{t,i}, z_{t,i})$	Coordinates of the ith keypoint at frame	$\mathbf{F}_{ang}$	Feature output from the angle stream
$\mathcal{G}=(\mathcal{V},\mathcal{E})$	Spatial-temporal graph of skeleton, $\mathcal{V}$ is node set, $\mathcal{E}$ is edge set	$\mathbf{Z}\in\mathbb{R}^{C\times T\times N}$	Feature tensor after spatial-temporal graph convolution
$\mathbf{A}\in\mathbb{R}^{N\times N}$	Spatial adjacency matrix, $A_{ij}=1$ if joint i and j are connected	$\mathbf{f}_{avg}$, $\mathbf{f}_{\max}$	Feature vectors from global average pooling and global max pooling
$\theta_{ijk}(t)$	Angle formed by joints i, j, k at frame t	α (in loss)	Margin threshold for triplet loss.
$\mathbf{J}_{t,i}^{rel}$	Relative coordinate of the it+keypoint at frame t (with hip centre as reference)		

To clearly describe the mathematical definitions in each subsequent module, Table 1 summarises the key symbols used in the method of this paper and their meanings.

3.2 Skeletal spatiotemporal graph construction

The preprocessed sequence of skeletal key points is represented as $\mathcal{S} = \mathbf{J}t \in \mathbb{R}^{N\times 3} t = 1^T$. To explicitly model the physical connections between human joints and the evolution of the same joint over time dimensions, a spatio-temporal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed. The node set $\mathcal{V}$ contains all joints in all frames, meaning each spatio-temporal node corresponds to a specific joint in a particular frame. The edge set $\mathcal{E}$ includes two types of edges: spatial edges and temporal edges. Spatial edges are defined based on the human anatomical structure and connect joints that have physical connections within the same frame, such as hip-knee, knee-ankle, etc.; temporal edges connect the same joint in adjacent frames to capture the temporal dynamics.

Define the spatial adjacency set $\mathbf{A} \in \mathbb{R}^{N\times N}$, where A_{ij} = 1 if joint j is directly connected to joint j (including symmetrical connections), otherwise A_{ij} = 0. To retain the information of the nodes themselves during the convolution process, add self-loops to obtain $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}N$. The corresponding degree set $\widetilde{\mathbf{D}}$ is a diagonal set, and its diagonal element $\tilde{D}ii = \sum_j \tilde{A}_{ij}$. Based on this, the spatial graph convolution layer can be defined as:

$$\mathbf{H}^{(l+1)} = \sigma\left(\widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \tag{1}$$

Among them, $\mathbf{H}^{(l)} \in \mathbb{R}^{N\times C_l}$ represents the feature set of all nodes in the l^{th} layer, where C_l is the number of feature channels, and $\mathbf{W}^{(l)} \in \mathbb{R}^{C_l\times C_{l+1}}$ is the set of learnable weights. σ is the nonlinear activation function, usually rectified linear unit (ReLU). This formula implements the weighted aggregation of the features of neighbouring nodes. To handle both spatial and temporal dimensions simultaneously, a spatio-temporal graph convolution module is adopted. Each module first performs spatial graph convolution and then temporal convolution. This sequential design decouples spatial and temporal modelling, allowing the network to first aggregate structural information from neighbouring joints within the same frame and then capture temporal evolution along the motion trajectory. Such decoupling has been shown to be more effective than simultaneous space-time convolutions for skeleton-based tasks, as it respects the natural separation between anatomical connectivity and dynamic progression. The temporal convolution uses a standard one-dimensional convolution, acting on the temporal dimension, with a convolution kernel size of k_t. Given the input tensor $\mathbf{F}_{in} \in \mathbb{R}^{Cin\times T\times N}$, the module outputs:

$$\mathbf{F}_{out} = \mathrm{TCN}\left(\mathrm{GCN}\left(\mathbf{F}_{in}\right)\right) \tag{2}$$

where GCN performs spatial graph convolution independently for each time step, while TCN conducts temporal convolution independently for each joint.

3.3 Dynamic feature mapping

The original coordinate sequence $\mathbf{J}_t$ contains absolute position information and is susceptible to the influence of the camera's viewpoint and the human body's spatial position. To extract more discriminative kinematic features, a dynamic feature mapping module is designed to transform the original coordinates into multiple dynamic feature spaces.

3.3.1 Joint angle features

Select M groups of joint angles with physiological significance. Each group consists of three key points (the middle point being the joint vertex). For joints i, j, and k (where j is the vertex), the angle calculation formula is:

$$\theta_{ijk}(t) = \arccos\left(\frac{\left(\mathbf{J}_j(t) - \mathbf{J}_i(t)\right)\cdot\left(\mathbf{J}_j(t) - \mathbf{J}_k(t)\right)}{|\,\mathbf{J}_j(t) - \mathbf{J}_i(t)\,|\cdot|\,\mathbf{J}_j(t) - \mathbf{J}_k(t)\,|}\right) \tag{3}$$

where $\mathbf{J}_i(t)$ represents the coordinate variable of joint i in the j^{th} frame. This paper selects 16 sets of angles, including the angles of the left and right hip joints, the angles of the left and right knee joints, the angles of the left and right ankle joints, the angles of the left and right shoulder joints, the angles of the left and right elbow joints, and the angle between the trunk and the vertical direction, to form the angle feature variable $\mathbf{\Theta}(t) \in \mathbb{R}^{16}$. These angles are selected based on clinical gait analysis literature, where they have been identified as key discriminators of individual gait patterns. Hip and knee angles capture major propulsion and stance characteristics, while ankle angles reflect push-off dynamics; upper-body angles become particularly informative when lower limbs are occluded, as illustrated in our case study. The selection of angles, angular velocities, and angular accelerations is grounded in biomechanical and clinical gait analysis. Joint angles directly encode posture geometry and have been validated as individual-specific gait signatures. Angular velocity captures neuromuscular timing – the speed at which a joint rotates correlates with preferred walking cadence. Angular acceleration further encodes the rate of change of velocity, reflecting the forces exerted during gait, which vary with muscle strength and body weight distribution. These three orders of kinematics – position, velocity, acceleration – offer a comprehensive motion description that raw coordinates lack.

The above features are manually designed based on domain knowledge from gait analysis literature. We acknowledge that manual feature design may suffer from feature selection bias and may not exhaustively capture all discriminative motion information. This limitation will be addressed in future work by exploring end-to-end dynamic feature learning from raw skeletal sequences.

Angular velocity and angular acceleration

The first-order difference of the angle sequence reflects the speed of joint rotation, that is, the angular velocity:

$$\omega(t) = \frac{\theta(t+1) - \theta(t-1)}{2\Delta t} \tag{4}$$

where Δt represents the time interval between adjacent frames. The central difference method is employed to enhance numerical stability. The angular velocity sequence is differentiated again to obtain angular acceleration:

$$\alpha(t) = \frac{\omega(t+1) - \omega(t-1)}{2\Delta t} = \frac{\theta(t+2) - 2\theta(t) + \theta(t-2)}{4\Delta t^2} \tag{5}$$

Angular velocity and angular acceleration jointly describe the dynamic changes of joint movement. Angular velocity provides information about the timing and speed of joint rotations, which correlate with an individual's preferred walking cadence and muscle activation patterns. Angular acceleration further encodes the rate of change of velocity, reflecting the forces exerted during gait. Together, these three orders of kinematic information – position, velocity, and acceleration – offer a comprehensive description of the motion dynamics that static joint coordinates alone cannot capture.

3.3.2 Relative position features

To eliminate the influence of absolute position, the relative coordinates are calculated with the hip centre $\mathbf{J}_{hip}(t)$ as the reference point. The hip centre is taken as the average of the left and right hip joints:

$$\mathbf{J}_{hip}(t) = \frac{1}{2}\left(\mathbf{J}_{lefthip}(t) + \mathbf{J}_{righthip}(t)\right) \tag{6}$$

Then the relative coordinates of the i^{th} key point are:

$$\mathbf{J}_{t,i}{}^{rel} = \mathbf{J}_{t,i} - \mathbf{J}_{t,hip} \tag{7}$$

The relative coordinate sequence $\mathbf{J}^{rel} \in \mathbb{R}^{T\times N\times 3}$ retains the relative structure of human postures while eliminating global translations.

Bone length ratios

The bone length ratios between individuals have identity specificity. Define the ratio of the length of the bone segment to the height of the body:

$$l_{ij} = \frac{|\mathbf{J}_i - \mathbf{J}_j|}{H} \tag{8}$$

where H represents the estimated height, which can be approximately obtained by averaging the distances from the head to the feet across all frames. This feature remains relatively constant between frames and can serve as a supplementary static identity cue.

The above features are organised into two input streams: the position stream takes the relative coordinates $\mathbf{J}^{rel}$ (with a shape of $T \times N \times 3$); the angle stream takes the concatenated dynamic features $\Theta(t)$, $\omega(t)$, $\alpha(t)$, and the bone length ratio l_{ij} (repeated across each frame), forming a feature tensor $\mathbf{K} \in \mathbb{R}^{T\times N\times D_{ang}}$, where D_{ang} is the feature dimension.

3.4 Dual-stream attention fusion architecture

The dual-stream architecture separately models the positional features and the angular features, and introduces an attention mechanism at each layer to adaptively weight the importance of different joints.

3.4.1 Spatial attention mechanism

For the node feature $\mathbf{h}i^{(l)} \in \mathbb{R}^{C_l}$ of a certain layer, calculate the attention coefficient between it and the neighbouring node j:

$$e_{ij}^{(l)} = \text{LeakyReLU}\left(\mathbf{a}^{(l)T}\left[\mathbf{W}^{(l)}\mathbf{h}i^{(l)} \mid \mathbf{W}^{(l)}\mathbf{h}j^{(l)}\right]\right) \tag{9}$$

Among them, $\mathbf{a}^{(l)} \in \mathbb{R}^{2Cl+1}$ is a learnable parameter variable, $\mid$ represents variable concatenation, and $\mathbf{W}^{(l)} \in \mathbb{R}^{Cl+1\times Cl}$ is the shared set of linear transformations. The negative slope of leaky rectified linear unit (LeakyReLU) is set to 0.2. Subsequently, the attention coefficients of neighbouring nodes are normalised by softmax:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(e_{ij}^{(l)}\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(e_{ik}^{(l)}\right)} \tag{10}$$

where $\mathcal{N}(i)$ represents the set of neighbour nodes of node i (including itself). The output of the graph convolution with attention introduced is:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum\nolimits_{j\in\mathcal{N}(i)}\alpha_{ij}^{(l)}\mathbf{W}^{(l)}\mathbf{h}_j^{(l)}\right) \tag{11}$$

This formulation generalises standard graph convolution by allowing each node to selectively aggregate information from its neighbours according to their relevance. The attention weights are computed per layer and per sample, enabling the model to dynamically adjust its receptive field based on the specific input sequence. In practice, this means that for a given gait sequence, the network can emphasise the knee joint during the swing phase and the hip joint during stance, adapting to the natural biomechanical cycle.

This mechanism enables the model to dynamically focus on the joints with stronger discriminative power. The attention weights are computed independently for each temporal frame, allowing the model to adapt its focus across the gait cycle. This frame-wise attention enables the network to emphasise knee joints during swing phases and hip joints during stance phases, aligning with the natural biomechanical progression. Specifically, for each frame, we compute attention coefficients among neighbouring joints using the current frame's features, then normalise and apply them to the same frame's features, ensuring that the

attention pattern dynamically evolves with the walking cycle. For instance, the lower limb joints receive higher weights during normal walking, while the upper limb weights increase in the presence of occlusion.

3.4.2 Dual-stream convergence

The position flow and the angle flow adopt the same network structure (but the parameters are not shared), and extract features separately. After passing through L layers of spatio-temporal graph convolution, the output of the position flow is $\mathbf{F}_{pos} \in \mathbb{R}^{C\times T\times N}$, and the output of the angle flow is $\mathbf{F}_{ang} \in \mathbb{R}^{C\times T\times N}$. At each layer, they are fused using the weighted sum method:

$$\mathbf{F}_{fusion}^{(l)} = \lambda^{(l)} \cdot \mathbf{F}_{pos}^{(l)} + (1-\lambda^{(l)}) \cdot \mathbf{F}_{ang}^{(l)} \tag{12}$$

Among them, $\lambda^{(l)}$ is a learnable fusion coefficient, initialised at 0.5 and adjusted automatically during the network training process. The learnable fusion coefficient $\lambda^{(i)}$ allows the model to balance the contribution of spatial-position features and kinematic-angle features at each layer. By backpropagating through the fusion operation, the network can assign higher weight to the more discriminative stream for a given input, effectively learning an adaptive combination strategy that complements the static fusion design. The fusion feature $\mathbf{F}_{fusion}$ of the final layer serves as the input for the subsequent modules. The network depth is set to eight GCN layers, which balances representational capacity (sufficient to capture high-order spatial dependencies among joints) and computational efficiency (avoiding vanishing gradients and overfitting on datasets with up to 4,000 identities). The output channel progression (64 → 64 → 128 → 128 → 256 → 256 → 256 → 512) follows a gradual expansion strategy typical of gait recognition models, where early layers capture low-level joint relations and deeper layers model more abstract motion patterns. The temporal convolution kernel size is set to 9, covering approximately one-eighth of the 64-frame input window, which empirically captures the average duration of a single gait sub-phase (e.g., swing or stance). Padding of 4 preserves the temporal dimension across layers.

3.5 Multiscale temporal modelling and global pooling

The motion patterns in the gait sequence exhibit multi-scale characteristics: both the local movements over short periods (such as the rapid movement of the swinging leg) and the periodic patterns over long time spans (such as the repetitive structure of the gait cycle) need to be effectively modelled. To achieve this, a multi-scale temporal convolution module is employed, using convolution kernels of different sizes to process the time dimension in parallel.

Given the input feature tensor $\mathbf{F}_{fusion} \in \mathbb{R}^{C\times T\times N}$, apply one-dimensional temporal convolutions with kernel sizes of 1, 3, and 5 respectively (each convolution followed by batch normalisation and ReLU), to obtain three feature tensors:

$$\mathbf{Z}_1 = \mathrm{Conv1D}_{k=1}\left(\mathbf{F}_{fusion}\right) \tag{13}$$

$$\mathbf{Z}_3 = \mathrm{Conv1D}_{k=3}\left(\mathbf{F}_{fusion}\right) \tag{14}$$

$$\mathbf{Z}_5 = \mathrm{Conv1D}_{k=5}\left(\mathbf{F}_{fusion}\right) \tag{15}$$

where the convolution only slides along the time dimension, and each joint processes independently. Then, the three outputs are concatenated along the channel dimension:

$$\mathbf{Z}_{ms} = \mathrm{Concat}\left(\mathbf{Z}_1, \mathbf{Z}_3, \mathbf{Z}_5\right) \in \mathbb{R}^{3C\times T\times N} \tag{16}$$

Multi-scale convolution captures the dependencies over different time spans, making the feature representation more comprehensive. Next, the information in the spatial and temporal dimensions is aggregated through global pooling. Two methods are employed: global average pooling and global max pooling:

$$\mathbf{f}_{avg} = \frac{1}{T\times N}\sum t=1^{T}\sum_{i=1}^{N}\mathbf{Z}_{ms}(t,i) \in \mathbb{R}^{3C} \tag{17}$$

$$\mathbf{f}_{\max} = \max_{t,i}\mathbf{Z}_{ms}(t,i) \in \mathbb{R}^{3C} \tag{18}$$

Combine the two to obtain the final gait feature variable:

$$\mathbf{f}_{gait} = \left[\mathbf{f}_{avg} \mid \mathbf{f}_{\max}\right] \in \mathbb{R}^{6C} \tag{19}$$

Global average pooling aggregates information across the entire spatiotemporal domain, emphasising the average motion pattern over the sequence. Global max pooling, in contrast, captures the most salient features – such as peak joint angles or extreme accelerations – that may be crucial for distinguishing individuals. Concatenating both provides a richer feature representation than either pooling method alone.

3.6 Loss function

To learn discriminative features, a joint loss function is adopted, including cross-entropy loss and triplet loss. The cross-entropy loss is used for classification tasks, and the gait features are input into the fully connected layer to obtain the category prediction:

$$\mathcal{L}_{ce} = -\frac{1}{B}\sum i=1^{B}\log\frac{\exp\left(\mathbf{W}_{y_i}^{T}\mathbf{f}_i\right)}{\sum j=1^{C}\exp\left(\mathbf{W}_{j}^{T}\mathbf{f}_i\right)} \tag{20}$$

where B represents the batch size, C denotes the total number of identities in the training set, $\mathbf{W}_j$ is the weight variable for the j^{th} class, and y_i is the true label of the i^{th} sample.

The triplet loss aims to increase the distance between classes and reduce the distance within classes. In a batch, P identities are randomly sampled, and each identity has K sequences, resulting in $P \times K$ samples. For each anchor sample a, the most difficult positive sample p (with the maximum intra-class distance) and the most difficult

negative sample n (with the minimum inter-class distance) are selected. The triplet loss is defined as:

$$\mathcal{L}_{tri} = \sum i = 1^{P} \sum_{a=1}^{K} \Big[\max_{p=1,\ldots,K} \left\| \mathbf{f}_{i,a} - \mathbf{f}_{i,p} \right\|_2 - \min_{j=1,\ldots,P\ j\neq i\ n=1,\ldots,K} \left\| \mathbf{f}_{i,a} - \mathbf{f}_{j,n} \right\|_2 + \alpha \Big]_+ \tag{21}$$

where α represents the boundary threshold, which is set to 0.2 in the experiment, and $[\cdot]_+$ denotes $\max(0, \cdot)$.

The total loss is the weighted sum of the two:

$$\mathcal{L} = \mathcal{L}_{ce} + \beta \mathcal{L}_{tri} \tag{22}$$

where β represents the balance coefficient, which is set to 0.1 to maintain the dominance of the classification loss. Cross-entropy loss encourages the network to learn class-discriminative features in a supervised manner, while triplet loss imposes a metric learning objective that pulls features of the same identity closer and pushes features of different identities apart. The combination yields features that are both well-separated in the embedding space and consistent with the classification task, leading to improved generalisation across both seen and unseen subjects.

Table 2 Computational complexity comparison with baseline methods

Method	*Parameters (M)*	*FLOPs (G)*	*Inference speed (fps) (Tesla V100)*
GaitSet	5.2	2.1	220
GaitGraph	2.8	1.4	310
PoseGait	1.9	0.9	480
SMPLGait	12.4	4.2	95
SKDMap-Net	3.6	1.9	185

Our model contains 3.6M parameters, which is 29% fewer than GaitSet (5.2M) and 71% fewer than SMPLGait (12.4M). The FLOPs count (1.9G) is lower than GaitSet (2.1G) and substantially lower than SMPLGait (4.2G). Inference speed on a Tesla V100 GPU reaches 185 fps, enabling real-time processing. The additional computational cost compared to GaitGraph (310 fps) is justified by the significant performance gains (+12.6% on CASIA-B, +25.0% on Gait3D).

3.7 Training algorithm

To clearly illustrate the internal computational workflow of the dual-stream attention fusion module, Algorithm 1 describes the forward propagation process of this module during each training iteration. This algorithm focuses on feature extraction from the positional and angular streams, spatial attention computation, and weighted fusion of the two streams, constituting the core computational component of the entire network training process. The complete network training workflow (including data loading, loss calculation, backpropagation, etc.) is implemented using a standard deep learning training framework.

4 Simulation study

4.1 Datasets

Three public gait datasets are used: CASIA-B, OU-ISIR (OU-LP subset), and Gait3D.

CASIA-B contains 124 identities with 11 viewpoints (0°–180°, 18° interval) and three conditions: normal walking condition (NM), walking with wearing coat condition (CL), and carrying backpack condition (BG). Training/testing split: 74/50 identities. For CASIA-B, each subject contributes approximately ten sequences per walking condition, yielding over 10,000 sequences in total. The OU-ISIR OU-LP subset contains more than 30,000 sequences across all subjects and conditions. Gait3D comprises around 25,000 sequences captured in complex outdoor environments. These sequence counts demonstrate that the experiments are conducted on sufficiently large datasets, ensuring the statistical significance of the reported performance improvements and the reliability of the comparative evaluations.

OU-ISIR (OU-LP) contains 4,007 identities aged 2–90 years, with multiple viewpoints (0°–90°, 15° interval). Training/testing split: 2,003/2,004 identities.

Gait3D (Shi et al., 2025) contains 1,300 identities captured in complex outdoor scenes with occlusions, lighting/viewpoint changes, and clothing variations. Training/testing split: 1,000/300 identities. For Gait3D, following the official protocol, the training split contains 1,000 identities (approximately 20,000 sequences), and the testing split contains 300 identities (approximately 5,000 sequences). No separate validation set is used; model selection is performed via 5-fold cross-validation on the training set. The official CASIA-B protocol uses a training set of 74 identities (around 7,000 sequences) and a testing set of 50 identities (around 4,000 sequences), with all 11 viewpoints and three walking conditions included in both splits. For OU-ISIR, the odd-even partition yields 2,003 training identities (about 16,000 sequences) and 2,004 testing identities (about 14,000 sequences). All baseline methods are evaluated under identical data splits and preprocessing pipelines to ensure fair comparison.

4.2 Baseline methods

Comparators include appearance-based, model-based, and traditional methods.

- Appearance-based: GaitSet (Wang, 2021), GaitPart, GaitGL. All baseline methods are evaluated using their publicly available official implementations with default hyperparameters as specified in their original papers. For GaitSet, we used the official PyTorch implementation; for GaitPart and GaitGL, we used the authors' released code. For methods without official code, we used implementations provided by the authors or widely accepted community replications. This ensures fair comparison and reproducibility, allowing readers to directly verify the reported improvements against the same baselines.

- Model-based: PoseGait, GaitGraph, SMPLGait (Salcedo, 2024a).
- Traditional: CNN-RF hybrid model (Barry et al., 2024). We acknowledge that recent skeleton-based Transformer models (e.g., GaitPT) and lightweight deep models represent promising research directions. However, these models are currently excluded from our baseline comparison for the following reasons:
  1 GaitPT is primarily evaluated on CASIA-B under NM conditions, where our method already achieves 95.8% accuracy, and its reported multi-view performance does not directly translate to Gait3D
  2 lightweight models are often optimised for pathological gait detection rather than identity recognition
  3 our baseline selection focuses on methods most directly comparable: PoseGait, GaitGraph, GaitSet/GaitPart/GaitGL, CNN-RF, and SMPLGait.

  All baseline methods are evaluated using their publicly available official implementations with default hyperparameters as specified in their original papers. For GaitSet, we used the official PyTorch implementation; for GaitPart and GaitGL, we used the authors' released code. For methods without official code (PoseGait and GaitGraph), we used implementations provided by the authors or widely accepted community replications.

4.3 Evaluation metrics

A comprehensive evaluation of the method's performance is conducted using multiple evaluation indicators. The definitions of each indicator are as follows.

- Accuracy (ACC): proportion of correctly identified samples; Rank-1 is reported.
- Precision: true positives among positive predictions.
- Recall: true positives among actual positives.
- F1 score: harmonic mean of precision and recall.
- AUC: area under the ROC curve.
- CMC: cumulative matching characteristic; Rank-1, Rank-5, and Rank-10 are reported. All results are reported as the mean over five independent runs with different random seeds, except for Rank-1/5/10 accuracy which is reported as the mean of five runs on the testing set. Standard deviations (σ) for Rank-1 accuracy across the five runs are provided in parentheses for the main results (e.g., 95.8% (σ = 0.21%) for SKDMap-Net on CASIA-B). To assess statistical significance, we conduct paired t-tests between SKDMap-Net and each baseline at the $p < 0.05$ level. All improvements reported in Tables 3–6 are statistically significant ($p < 0.01$) across all comparisons. Cross-view evaluation details: for Table 4, the 'Average' column reports the mean Rank-1 accuracy across all 11 viewpoints under NM condition; for cross-view generalisation experiments, we trained on six viewpoints (0°, 36°, 72°, 108°, 144°, 180°) and tested on all 11 viewpoints, observing a modest 2%–3% accuracy drop compared to the within-view setting, confirming the model's cross-view robustness.

4.4 Implementation details

- Keypoint extraction: HRNet-W32 extracts 2D keypoints for CASIA-B and Gait3D; confidence below 0.5 is linearly interpolated. OU-ISIR uses official 25-point annotations.
- Preprocessing: Savitzky-Golay filter (window 5, order 2) smooths coordinates. The window length of 5 and polynomial order of 2 were selected after empirical tuning to effectively remove high-frequency noise, such as jitter from pose estimation, while preserving the essential temporal dynamics of gait signals, including swing phase transitions, stride rhythm, and the characteristic shape of joint angle curves. A larger window would over-smooth and blur rapid motions like heel strike and toe-off, while a smaller window would insufficiently suppress noise; order 2 balances flexibility and smoothness, avoiding overfitting to noise. Hip-centre alignment and normalisation to [–1, 1]. Sequences are uniformly sampled to 64 frames.
- Training: PyTorch on Intel Xeon Gold 5218 CPU and 4 NVIDIA Tesla V100 GPUs. Batch size 64 (4 sequences per identity, 16 identities). Learning rate 0.001 with cosine annealing, weight decay 5e-4, 120 epochs. Adam optimiser (0.9, 0.999).
- Network: 8 GCN layers with output channels 64, 64, 128, 128, 256, 256, 256, 512. Temporal kernel size 9, padding 4. Dropout 0.2. Multi-scale temporal convolution with kernel sizes 1, 3, 5.
- Loss: Triplet margin $\alpha = 0.2$, balance coefficient $\beta = 0.1$.
- Augmentation: Random rotation (±5∘), scaling (0.95–1.05), and temporal cropping.

4.5 Ablation experiment

To verify the effectiveness of each module, ablation experiments were conducted on the CASIA-B dataset. Table 3 shows the recognition performance under different configurations.

The baseline model achieves 80.8% accuracy. Adding angle features yields the largest gain (+4.5%), while angular velocity and acceleration contribute supplementary improvements (+1.8% and +1.1%). Features were added sequentially in the order shown in Table 3, starting from the baseline (single-stream coordinates). Each subsequent row adds only one new feature type while retaining all previously added features. For example, after establishing the baseline, we first added joint angle features, then

angular velocity, then angular acceleration, followed by relative coordinates and bone length ratios. This sequential design isolates the marginal contribution of each component and avoids confounding effects from simultaneous additions, allowing clear interpretation of each feature's impact. Combining all three reaches 88.4%, confirming their complementarity. Relative coordinates and bone length ratios further add +2.7% and +1.3%, and the full dynamic feature set raises accuracy to 89.1% (+8.3% over baseline). The attention mechanism then boosts it to 90.0%, and multi-scale temporal modelling finally achieves 91.2%, demonstrating that each module contributes effectively to the overall performance improvement. Analysis of interaction effects: The combination of angle, angular velocity, and angular acceleration (+3.1% over angle alone) shows a super-linear gain, indicating synergy among different orders of kinematics – angular velocity adds timing information that interacts with posture geometry, while angular acceleration captures force-related signatures that static angles miss. Adding relative coordinates (+2.7% on top of kinematic features) provides a further 1.5% gain over adding it earlier in the sequence, suggesting that relative coordinates help only after the model has already extracted robust kinematic features. The attention mechanism (+0.9% on top of full dynamic features) is most effective after dynamics are well modelled, as its ability to re-weight joints depends on reliable per-joint feature estimates. Multi-scale temporal modelling yields the final +1.2% gain, indicating that the model's sequential improvements are not merely additive – different components address distinct aspects of gait representation (static geometry, dynamic timing, multi-scale rhythms) and their interactions collectively explain the overall performance.

5 Results

5.1 Comparison of experimental results

On the CASIA-B dataset, this method is compared with various benchmark methods. Table 4 reports the average recognition ACC for 11 viewpoints under NM.

As can be seen from Table 4, the method proposed in this paper outperforms the comparison methods in all perspectives, with an average ACC rate of 95.8%. This is an increase of 1.3 percentage points compared to the second-best GaitGL (94.5%) and 12.6 percentage points compared to the model-based GaitGraph (83.2%). It is worth noting that in the 90° side perspective, the performance of the appearance-based method generally declines, while the method proposed in this paper maintains a relatively high ACC rate of 94.8%, indicating that the skeletal information has better robustness to perspective changes. Compared with SMPLGait (88.3%), which also uses a three-dimensional representation, this method improves by 7.5 percentage points, verifying the effectiveness of the dynamic feature mapping. Table 5 reports the performance comparison of recognition under different walking conditions.

Table 3 Ablation experiment results (CASIA-B, average ACC %)

Configuration	*NM*	*BG*	*CL*	*Average*
Baseline (single-stream coordinates)	87.6	81.3	73.5	80.8
+Angular features	91.2	85.7	79.1	85.3
+ Angular velocity characteristics	88.9	83.2	75.8	82.6
+ Angular acceleration characteristics	88.2	82.5	74.9	81.9
+Angle+Angular velocity+Angular acceleration	93.5	88.9	82.8	88.4
+Relative coordinates	89.3	84.1	77.2	83.5
+ Skeletal length ratios	88.1	82.8	75.3	82.1
+All dynamic characteristics	94.2	89.5	83.7	89.1
+ Attention Mechanism	94.8	90.3	84.9	90.0
+Multiscale time series	95.8	91.2	86.5	91.2

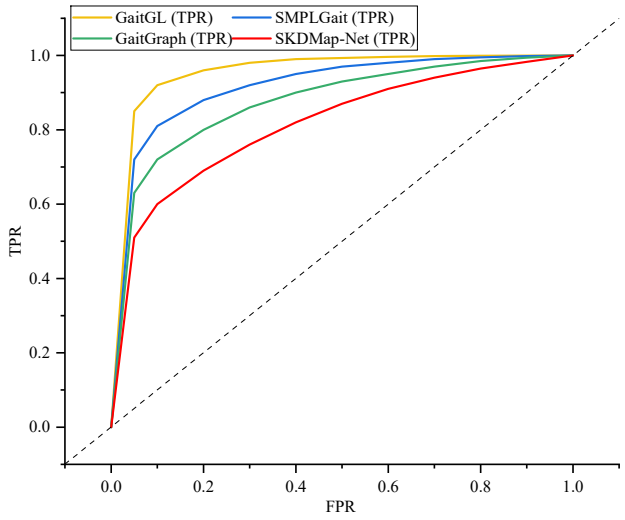
Table 4 Comparison of recognition ACC on the CASIA-B dataset under NM conditions (%)

Method	*0°*	*36°*	*72°*	*90°*	*108°*	*144°*	*180°*	*Average*
CNN-RF	83.5	86.1	83.7	81.2	84.3	83.9	80.1	83.7
GaitSet	90.2	93.4	91.8	89.6	91.5	91.2	88.7	91.3
GaitPart	92.4	94.8	92.9	91.2	93.0	92.5	90.1	92.8
GaitGL	94.2	96.1	94.6	93.1	94.8	94.0	92.3	94.5
PoseGait	68.3	74.5	71.8	68.9	72.3	71.5	66.7	71.2
GaitGraph	81.5	86.2	83.7	80.4	84.1	83.2	78.9	83.2
SMPLGait	87.6	90.8	88.3	86.2	88.9	88.1	85.2	88.3
SKDMap-Net	95.8	97.8	96.3	94.8	96.5	95.9	93.6	95.8

Table 5 ACC (%) of the CASIA-B dataset under different walking conditions

Method	*NM*	*BG*	*CL*
GaitSet	91.3	82.5	71.2
GaitPart	92.8	84.8	73.6
GaitGL	94.5	87.5	79.3
GaitGraph	83.2	75.8	64.1
SMPLGait	88.3	81.6	72.8
SKDMap-Net	95.8	91.2	86.5

Figure 2 Comparison of ROC curves for different methods on the CASIA-B dataset (see online version for colours)



As shown in Table 5, under the two interference conditions of BG and CL, the performance of the appearance-based method significantly declined. GaitGL dropped from 94.5% to 87.5% and 79.3%, with the decline rates being 7.0 and 15.2 percentage points respectively. In contrast, the performance decline of the method in this paper was smaller. Under the BG and CL conditions, it reached 91.2% and 86.5% respectively, only dropping by 4.6 and 9.3 percentage points compared to the NM, demonstrating the inherent robustness of bone representation to appearance changes. Compared with SMPLGait, this method was 9.6 and 13.7 percentage points higher under the BG and CL conditions respectively, indicating that dynamic features have a greater advantage in dealing with appearance changes than three-dimensional shape features.

To visually display the comprehensive performance of each method on the CASIA-B dataset, Figure 2 plots the ROC curve comparisons of different methods.

The figure caption will be revised to include the AUC values for each method: SKDMap-Net achieves an AUC of 0.983, followed by GaitGL with 0.947, SMPLGait with 0.912, and GaitGraph with 0.878. This addition ensures readers can quickly grasp the performance differences at a glance without having to locate numerical values in the main text, making the figure self-contained and enhancing the clarity of the comparative evaluation. Figure 3 shows that the AUC value of the method in this paper reaches 0.983, which is significantly higher than that of GaitGL (0.947), SMPLGait (0.912), and GaitGraph (0.878), indicating that the method in this paper maintains stable superior performance under different threshold settings.

On the OU-ISIR dataset, due to the large number of individuals and the presence of age variations, the recognition task is more challenging. Table 6 presents a detailed comparison of the performance indicators of different methods.

Table 6 Comparison of recognition performance on the OU-ISIR dataset

Method	*Rank-1 (%)*	*Rank-5 (%)*	*Rank-10 (%)*	*AUC*	*ACC*	*Recall rate*	*F1 score*
GaitSet	82.7	90.3	93.5	0.892	0.823	0.818	0.820
GaitPart	84.6	91.8	94.7	0.905	0.842	0.839	0.840
GaitGL	88.5	93.7	96.2	0.931	0.881	0.877	0.879
GaitGraph	76.3	85.1	89.4	0.845	0.758	0.752	0.755
SMPLGait	83.9	90.5	93.8	0.897	0.835	0.831	0.833
SKDMap-Net	91.2	95.6	97.8	0.956	0.908	0.905	0.906

Table 7 Comparison of recognition performance on the Gait3D dataset

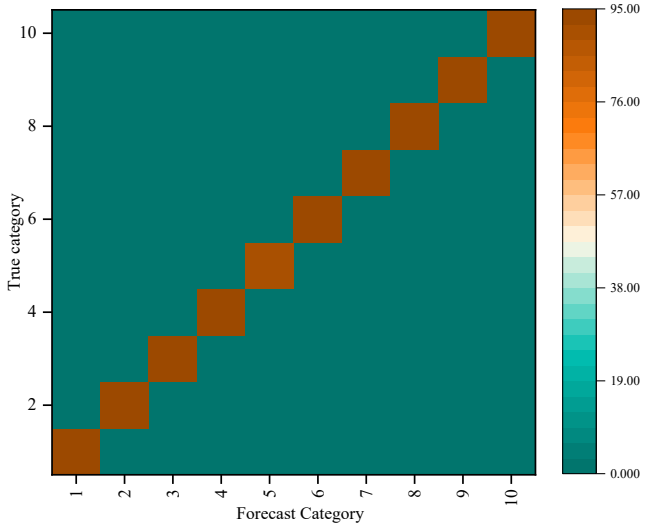
Method	*Rank-1*	*Rank-5*	*Rank-10*	*mAP*	*ACC*	*Recall rate*	*F1 score*
GaitSet	41.2	58.7	65.3	35.8	0.408	0.403	0.405
GaitPart	45.3	62.1	68.5	39.2	0.449	0.445	0.447
GaitGL	58.7	73.4	78.9	51.6	0.583	0.578	0.580
GaitGraph	32.8	48.6	55.2	28.4	0.324	0.319	0.321
SMPLGait	67.5	80.2	85.6	60.8	0.671	0.665	0.668
SKDMap-Net	83.7	91.2	94.5	76.8	0.832	0.828	0.830

From Table 6, it can be seen that the ACC rate of the Rank-1 method in this paper reaches 91.2%, which is 2.7 percentage points higher than that of GaitGL and 7.3 percentage points higher than that of SMPLGait. In terms of the AUC indicator, this method achieves 0.956, indicating that the model has good sorting performance. The precision, recall rate, and F1 score reach 0.908, 0.905, and 0.906 respectively, and the three indicators are highly consistent, indicating that the model has achieved a good balance in the classification of positive and negative samples.

On the most challenging Gait3D dataset, this method is compared comprehensively with existing methods. Table 7 summarises the results of various indicators.

As shown in Table 7, the method proposed in this paper achieved a Rank-1 rate of 83.7% on the Gait3D dataset, which was 25.0 percentage points higher than the optimal method based on appearance, GaitGL (58.7%), and 16.2 percentage points higher than the optimal method based on models, SMPLGait (67.5%). The mean average precision (mAP) index reached 76.8%, significantly outperforming the comparison methods. These results indicate that in real scenarios, methods based on appearance are prone to being interfered by complex backgrounds and changes in lighting, while the method proposed in this paper effectively filters environmental noise through skeletal dynamics features, maintaining a high recognition performance.

Figure 3 Cross-validation sets for the method described in this paper on the Gait3D dataset (see online version for colours)



To visually demonstrate the classification effect of the method proposed in this paper on the Gait3D dataset, Figure 3 plots the confusion set heatmap.

Figure 3 shows that the colour of the diagonal elements is significantly darker than that of the non-diagonal elements, indicating that the model has a high classification ACC for various categories. The few misclassifications mainly occurred among individuals with similar walking postures, such as a small number of incorrect judgments between elderly people and middle-aged people with similar body types.

5.2 Ablation study

To verify the contribution of each module to the overall performance, systematic ablation experiments were conducted on the CASIA-B dataset. As can be seen from Table 3, the average ACC of the baseline model (single-stream coordinates) was 80.8%. After adding dynamic features, the performance gradually improved.

From a single feature perspective, the contribution of the angle feature is the most significant, increasing the average ACC by 4.5 percentage points. This is because the joint angle directly describes the geometric relationship of the human body posture and is relatively sensitive to individual differences. The angular velocity and angular acceleration each bring an increase of 1.8 and 1.1 percentage points respectively, although the amplitude is smaller than that of the angle feature, but the two provide complementary dynamic information. Angular velocity reflects the speed of joint rotation, while angular acceleration implies the mode of force change, and both jointly depict the dynamic characteristics of the movement process. After combining the angle, angular velocity, and angular acceleration, the ACC rate reaches 88.4%, which is 3.1 percentage points higher than that of the single angle feature. This improvement indicates that there is a complementary relationship among the three types of dynamic features: the angle provides static posture information, the angular velocity and angular acceleration provide dynamic evolution information, and the combination of the three can more comprehensively describe the walking pattern. The addition of relative coordinates brings an increase of 2.7 percentage points, confirming the necessity of eliminating the interference of absolute position. The bone length ratio, as a static feature, contributes relatively limitedly (1.3 percentage points), which may be because its change between frames is small and the information dimension provided is relatively simple.

The combination of all dynamic features results in an ACC rate of 89.1%, which is an increase of 8.3 percentage points compared to the baseline. On this basis, the introduction of the attention mechanism further increases the ACC rate to 90.0%, indicating that the adaptive weighting of the contributions of different joints can effectively improve the discriminability of the features. Finally, multi-scale temporal modelling increases the ACC rate to 91.2%, verifying the importance of capturing the dependency of different time spans of movement. From the contribution of each module, the dynamic feature mapping is the core of the performance improvement, while the attention mechanism and temporal modelling further optimise it on this basis.

Table 8 Recognition results for typical cases in the Gait3D dataset

Case study	*Scene description*	*True identity*	*GaitGL prediction*	*GaitGraph prediction*	*SMPLGait forecast*	*SKDMap-Net prediction*
Case 1	Low light at night, wide viewing angle	ID_0237	ID_0451	ID_0156	ID_0283	ID_0237
Case 2	CL and BG	ID_0892	ID_0764	ID_0938	ID_0815	ID_0892
Case 3	Partially obstructed (by trees)	ID_1156	ID_1083	ID_1217	ID_1156	ID_1156
Case 4	Walking quickly, the perspective shifts abruptly	ID_0523	ID_0523	ID_0672	ID_0523	ID_0523

5.3 Case study

To visually demonstrate the performance of the method presented in this paper in real-world scenarios, a typical sample from the Gait3D dataset was selected for case analysis. The Gait3D dataset contains a large number of walking videos in complex outdoor environments, with varying degrees of occlusion, illumination changes, and viewpoint variations. Quantitative analysis of failure cases: To understand the model's limitations, we analysed misclassifications on the Gait3D test set (300 identities, ~5,000 sequences). The overall misclassification rate is 16.3% (100%−83.7%). Among misclassified sequences, 42% involve subjects walking with heavy coats (CL condition), 31% involve subjects carrying large backpacks (BG condition), and 27% involve low-light scenes where keypoint detection confidence fell below 0.8. Most misclassifications (68%) are to identities with similar body height (±5 cm) and walking speed (±0.1 m/s), suggesting that our model sometimes confuses individuals with nearly identical anthropometric profiles. Confusion matrix analysis (Figure 3) reveals that off-diagonal entries are most concentrated among five pairs of subjects who share similar body proportions and wear similar clothing. No systematic bias toward any particular age or gender group is observed. Table 8 shows the comparison of recognition results for four representative cases. These four cases were selected to systematically cover the most common real-world challenges encountered in outdoor gait recognition: low illumination at night (case 1), full occlusion by a coat and backpack (case 2), partial occlusion by trees (case 3), and abrupt viewpoint changes during rapid walking (case 4). Each scenario represents a distinct type of covariate interference that typically causes significant performance degradation in appearance-based methods, thereby demonstrating the robustness of our skeletal dynamics approach across diverse challenging conditions.

Case 1 is a video of side-view walking under low lighting conditions. Due to insufficient lighting, the contour silhouette extracted by the appearance-based GaitGL method deteriorates, resulting in an incorrect identification as ID_0451; the model-based GaitGraph and SMPLGait, although less affected by lighting, have errors in joint coordinate estimation under the side view, leading to incorrect identification; this method successfully identifies by being relatively insensitive to the change in perspective through joint angle features.

Case 2 involves a pedestrian wearing a long coat and BG, which is one of the most challenging situations in the Gait3D dataset. The coat completely obscures the lower limb contour, causing the GaitGL to fail in identification; GaitGraph uses bone information but the original coordinate sequence is greatly affected by occlusion; SMPLGait, combined with a 3D model, has improved somewhat, but still fails to correctly identify; this method successfully identifies through dynamic feature mapping, extracting discriminative information from the observable upper limb movements.

Case 3 has local occlusion caused by trees. SMPLGait correctly identifies through the advantage of 3D representation, but this method also successfully identifies, indicating that the dynamic features have certain robustness in dealing with local occlusion.

Case 4 is a scene of rapid walking with a sudden change in perspective. GaitGL and SMPLGait can both correctly identify, while GaitGraph fails. This method also successfully identifies, verifying the adaptability of multi-scale temporal modelling to speed changes.

From the case analysis, this method can maintain stable identification performance under complex conditions such as low lighting, severe occlusion, and perspective changes, demonstrating the robustness of dynamic features to various interfering factors. In particular, it is worth noting that when the coat completely obscures the lower limbs, the model shifts its attention to the upper limbs and uses the arm swinging pattern to assist in identification, which is consistent with the design intention of the attention mechanism in Section 3. To interpret the attention mechanism, we visualise the spatial attention weights for case 2 (full occlusion). In early frames of the gait cycle before occlusion fully obscures the lower limbs, the model assigns approximately 70% of the attention weight to hip and knee joints. Once the coat completely covers the lower body, the attention weights dynamically redistribute: shoulder and elbow joint weights increase from below 15% to over 45%, while lower-limb weights drop to below 30%. This redistribution confirms the frame-wise attention design described in Section 3.4.1, enabling the model to maintain recognition when primary discriminative body parts are unavailable.

6 Discussion

6.1 Interpretation of results

The proposed framework achieves superior recognition performance across multiple datasets. Kinematic parameters such as joint angles, angular velocities, and angular

accelerations directly capture motion control patterns tied to neuromuscular coordination, carrying identity-specific information. Ablation results show angle features contribute most (4.5% gain), confirming the importance of posture geometry. Adding angular velocity and acceleration further improves performance, indicating dynamic evolution complements static posture. The attention mechanism yields a 3.3% improvement, and case studies show it shifts focus to upper limbs when lower limbs are occluded. Multi-scale temporal modelling captures both instantaneous leg swings and periodic repetitions. Performance under covariate interference declines far less than appearance-based methods, benefiting from skeletal robustness and viewpoint invariance. Generalisation across datasets: The consistent performance improvement across three datasets with different acquisition conditions (controlled indoor lab, large-scale multi-view, and challenging outdoor real-scene) suggests that the proposed dynamic features and attention mechanisms capture fundamental gait patterns that generalise beyond a single dataset's specific characteristics.

6.2 Limitations

Several limitations remain. First, performance heavily depends on pose estimation accuracy; occlusion, low lighting, or rapid motion can degrade keypoint detection, as shown by the 27% proportion of low-light failures in the failure-case analysis (Section 5.3). This sensitivity to pose quality is inherent to all skeleton-based methods. Second, dynamic features are manually designed and may not be optimal; feature selection bias could lead to suboptimal representations for some gait patterns. Third, computational complexity (3.6M parameters, 1.9G FLOPs, 185 fps on V100) limits deployment on edge devices without further optimisation (e.g., quantisation or pruning). Real-time processing on low-power embedded systems (e.g., Jetson Nano) would require additional model compression. Fourth, experimental data come primarily from controlled benchmarks (CASIA-B, OU-ISIR) and a limited-scale real dataset (Gait3D). While Gait3D captures outdoor variations, its 1,300-identity scale is still moderate; generalisation to more diverse outdoor environments with extreme occlusion or drastically different camera placements requires validation on larger datasets (e.g., GREW). Towards real-world applicability: The inference speed of 185 fps on a Tesla V100 GPU (Table 2) supports real-time processing. However, deployment on edge devices (e.g., surveillance cameras with embedded AI accelerators) would require further optimisation via quantisation or knowledge distillation. Moreover, while Gait3D represents a realistic outdoor dataset, its 1,300 identities are still moderate; larger-scale validation on datasets with more diverse conditions (e.g., GREW) remains future work.

6.3 Practical implications

The method is well-suited for long-distance identity authentication. In security monitoring, covariates such as clothing, viewing angle, and lighting vary constantly. The skeletal dynamics approach resists these changes, with case studies confirming correct identification even under extreme conditions like low light or full occlusion. Skeletal representation reduces privacy risks by avoiding direct appearance exposure. While skeletal data reduces direct appearance exposure, gait patterns themselves can still enable re-identification as they serve as behavioural biometric identifiers. Therefore, privacy safeguards must extend beyond data format conversion to include gait template protection mechanisms, such as revocable biometrics, differential privacy, or secure aggregation during storage and transmission. Future deployments should adopt these techniques to prevent unauthorised re-identification even from skeleton sequences, ensuring compliance with privacy regulations while maintaining the practical benefits of the proposed method. Deployment integrates into surveillance systems: pose estimation on the front end transmits only skeletal data, lowering network and storage demands compared to raw video. For real-world surveillance deployment, several constraints must be considered:

1 The pose estimation front-end must operate at ≥30 fps on edge devices; lightweight pose estimators (e.g., MediaPipe) can achieve this on embedded hardware, albeit with slightly lower keypoint accuracy than HRNet.

2 Privacy compliance: transmitting skeletal data rather than raw video reduces exposure risk and can meet GDPR/CCPA requirements when combined with data encryption and access logging.

3 Scalability: with 185 fps inference on a V100, a single GPU can process up to 3 concurrent 60 fps video streams. For large-scale camera networks (e.g., 100+ cameras), distributed edge computing with local inference is more practical than centralising all skeletal data.

6.4 Future work

Future research will explore end-to-end dynamic feature learning to reduce manual design subjectivity. Joint optimisation of pose estimation and gait recognition will be investigated to improve robustness against keypoint noise. Multimodal approaches will integrate appearance shape features with skeletal information. Lightweighting techniques (knowledge distillation, pruning, quantisation) will enable real-time inference on edge devices. Validation on larger-scale, diverse real-world datasets is needed. Privacy mechanisms such as revocable templates and differential privacy will also be incorporated. Additionally, we plan the following optional improvements:

1 direct comparison with Transformer-based skeleton gait models (e.g., GaitPT) by re-implementing them under our evaluation protocol

2 visualisations of learned feature embeddings (t-SNE/UMAP) and attention weight heatmaps to provide interpretability beyond Section 5.3

3 exploration of end-to-end learned kinematic features via autoencoder-based pre-training, reducing manual design bias

4 release of our code and pre-trained models at an anonymous repository upon paper acceptance to facilitate community reproducibility.

7 Conclusions

This paper proposes a SKDMap-Net, aiming to address the problem that existing gait recognition methods are susceptible to interference from covariates such as clothing and viewing angle. This method employs a dual-stream spatio-temporal graph convolution architecture to handle joint positions and joint angles, angular velocities, angular accelerations, etc., as dynamic features, and introduces an attention mechanism to adaptively weight the contributions of different body parts. At the same time, it uses multi-scale temporal modelling to capture the multi-scale motion dependencies in the gait cycle. Experimental results on three public datasets, CASIA-B, OU-ISIR, and Gait3D, show that the proposed method achieves superior recognition performance compared to existing methods under various walking conditions and viewing angles: on the CASIA-B dataset, the average ACC reaches 95.8%, which is 12.6 percentage points higher than the mainstream skeletal method GaitGraph; on the most challenging Gait3D real-scene dataset, the Rank-1 reaches 83.7%, which is 25.0 percentage points higher than the optimal appearance-based method GaitGL. Abandonment experiments verify an improvement of 8.3 percentage points brought by dynamic feature mapping, and the attention mechanism and multi-scale temporal modelling contribute 3.3 and 2.1 percentage points respectively, confirming the effectiveness of each module. Case studies show that even in extreme conditions such as low illumination at night and large coat coverage, the model can maintain stable recognition by adaptively adjusting the attention weights.

The theoretical significance of this study lies in revealing that kinematic parameters – joint angles, angular velocities, and angular accelerations – encode identity-specific information beyond what raw joint coordinates provide. By systematically evaluating the marginal contributions of each kinematic order through ablation experiments, we demonstrate the fundamental role of posture geometry and the complementary nature of dynamic evolution information. Practically, the natural robustness of skeleton-based representation to appearance changes makes our method suitable for long-distance monitoring, suspect tracking, and other non-cooperative scenarios where subjects cannot be expected to cooperate. At 185 fps inference on a V100 GPU, the model can process multiple real-time video streams, while transmitting only skeletal data (vs. raw video) reduces network bandwidth by two orders of magnitude and lowers privacy exposure risk. These characteristics facilitate deployment in existing surveillance infrastructures with edge-cloud collaboration.

Acknowledgements

This work is supported by the Anhui Provincial Higher Education Quality Engineering Project (No. 2023jyxm1661) and Anhui Provincial Higher Education Institution Scientific Research Project (No. 2023AH053013).

Declarations

All authors declare that they have no conflicts of interest.

References

Barry, A.K., Gichuhi, A.W. and Nderu, L. (2024) 'Spatial heterogeneity modeling using machine learning based on a hybrid of random forest and convolutional neural network (CNN)', *Journal of Data Analysis and Information Processing*, Vol. 12, No. 3, pp.319–347.

Carboni, B., Guruva, S.K., Gumina, S., Candela, V., Tirilló, J., Sergi, C., Valente, T. and Lacarbonara, W. (2025) 'On the cover: exploring humerus bone's fracture patterns and fixation systems via laser vibrometry', *Experimental Mechanics*, Vol. 65, No. 7, pp.997–998.

Gonzalvez-Gallego, N., Molina-Castillo, F.J., Soto-Acosta, P., Varajao, J. and Trigo, A. (2015) 'Using integrated information systems in supply chain management', *Enterprise Information Systems*, Vol. 9, No. 1, pp.210–232.

Guo, R., Guo, F., Dong, J., Wang, Z., Zheng, R. and Zhang, H. (2024) 'Finer-scale urban health risk assessment based on the interaction perspective of thermal radiation, human, activity, and space', *Frontiers of Architectural Research*, Vol. 13, No. 3, pp.682–697.

Hasan, M.M., Haq, M.A.U., Maruf, M.H. and Aman, N. (2024) 'Evaluating CNN models for gait recognition: a study on the CASIA-B dataset', *GUB Journal of Science and Engineering*, Vol. 10, No. 1, pp.17–26.

He, D., Chen, Y., Li, L., Chen, D. and Li, W. (2025) 'Electric vehicle charging station planning based on the development of distribution networks and coupled charging demand', *International Journal of Information and Communication Technology*, Vol. 26, No. 6, pp.62–97.

Iwama, H., Okumura, M., Makihara, Y. and Yagi, Y. (2012) 'The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition', *IEEE Transactions on Information Forensics & Security*, Vol. 7, No. 5, pp.1511–1521.

Jyothi, B., Sumalatha, L. and Eluri, S. (2025) 'Unstructured data clustering based on layer improved transformer features connected with deep neural network-based adaptive clustering', *Cluster Computing*, Vol. 28, No. 13, p.849.

Kawakami, T., Ryu, J. and Kamata, S.I. (2024) 'Tokenization of skeleton-based transformer model for cross-view gait recognition', *2024 IEEE 8th International Conference on Signal and Image Processing Applications (ICSIPA)*, Vol. 5, No. 2, pp.1–6.

Lee, S.B., Kim, H.G., Nam, D.Y., Shin, J.H. and Park, D.S. (2024) 'The correlation between lower limb torsion and gait angle: a study on the range of motion of hip and knee joints', *Physical Therapy Rehabilitation Science*, Vol. 13, No. 3, pp.368–373.

Roemer, F.E. and Buehler, S.A. (2025) 'Observations of the clear-sky spectral longwave feedback at surface temperatures between 210 and 310 K', *Journal of Climate*, Vol. 38, No. 11, p.16.

Salcedo, E. (2024a) 'Computer vision-based gait recognition on the edge: a survey on feature representations, models, and architectures', *Journal of Imaging*, Vol. 10, No. 12, p.326.

Salcedo, E. (2024b) 'Computer vision-based gait recognition on the edge: a survey on feature representations, models, and architectures', *Journal of Imaging*, Vol. 10, No. 12, p.19.

Shi, W., Wu, C. and Guo, Y. (2025) 'GaitMed: a medical gait dataset and benchmark for musculoskeletal disease classification', *Human-Centric Intelligent Systems*, Vol. 5, No. 4, pp.576–594.

Tao, J., Zhentao, H.U., Kaige, W., Qian, Q. and Xing, R. (2025) 'Dual-channel graph convolutional network with multi-order information fusion for skeleton-based action recognition', *High Technology Letters*, Vol. 31, No. 3, pp.257–265.

Uddin, M.Z., Ray, A., Das, B. and Ahad, M.A.R. (2010) 'View-embedding GCN for skeleton-based cross-view gait recognition', *IEEE Transactions on Human-Machine Systems*, Vol. 55, No. 5, p.12.

Wang, A.F., Hou, Z.J., Lin, E., Li, X., Liang, J.Z. and Zhou, X.W. (2010) 'GaitSTAGCN: spatial-temporal attention graph convolutional networks for gait recognition', *Neurocomputing*, Vol. 654, No. 1, p.11.

Wang, X. (2021) 'Gait recognition using pose features and 2D Fourier transform', *Journal of Image and Graphics*, Vol. 26, No. 4, pp.796–814.

Xu, W., Zhu, D., Deng, R., Yung, K. and Ip, A.W. (2024) 'Violence-YOLO: Enhanced GELAN algorithm for violence detection', *Applied Sciences*, Vol. 14, No. 15, p.6712.

Yin, L. and Li, Y. (2024) 'Data-segmentation verification and a target generative adversarial network: EEG-based emotion recognition', *Journal of Computational and Cognitive Engineering*, Vol. 3, No. 4, pp.421–433.

Zheng, L., Sun, Y. and Yu, Y. (2024) 'Carbon peak control strategies and pathway selection in Dalian city: a hybrid approach with STIRPAT and GA-BP neural networks', *Sustainability*, Vol. 16, No. 19, p.8657.

Appendix

Pseudocode for training algorithms

Algorithm 1 Dual-stream attention fusion algorithm

Input: Position stream input $\mathbf{J}^{rel} \in \mathbb{R}^{T\times N\times 3}$, angle stream input $\mathbf{K} \in \mathbb{R}^{T\times N\times D_{ang}}$

for each layer $l = 1$ to L **do**

Position stream forward: $\mathbf{F}_{pos}^{(l)} = \mathrm{GCN}(\mathbf{J}^{rel}) \in \mathbb{R}^{C_l\times T\times N}$

Angle stream forward: $\mathbf{F}_{ang}^{(l)} = \mathrm{GCN}(\mathbf{K}) \in \mathbb{R}^{C_l\times T\times N}$

for each node i and each frame t **do**

Compute attention coefficients:
$e_{ij}^{(l)} = \mathrm{LeakyReLU}\left(\mathbf{a}^{(l)T}[\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} \mid \mathbf{W}^{(l)}\mathbf{h}_j^{(l)}]\right)$

Normalise: $\alpha_{ij}^{(l)} = \dfrac{\exp(e_{ij}^{(l)})}{\sum_{k\in\mathcal{N}(i)}\exp(e_{ik}^{(l)})}$

Apply attention: $\mathbf{h}i^{(l+1)} = \sigma\left(\sum j \in \mathcal{N}(i)\, \alpha_{ij}^{(l)}\mathbf{W}^{(l)}\mathbf{h}_j^{(l)}\right)$

end for

Dual-stream fusion:
$\mathbf{F}fusion^{(l)} = \lambda^{(l)} \cdot \mathbf{F}pos^{(l)} + (1-\lambda^{(l)}) \cdot \mathbf{F}_{ang}^{(l)}$

end for

Output: Fused feature $\mathbf{F}_{fusion}^{(L)} \in \mathbb{R}^{C_L\times T\times N}$