ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Gait Style Modeling Based on Posture Feature Learning

Ruoqi Shi[1,*]

1 School of Design, Jianghan University, Wuhan 430000, Hubei, China

SUMMARY: *Model gait has the characteristics of stable posture rhythm, limb coordination and trajectory control, which can be expressed by structured posture sequence. This paper proposes a framework of posture feature learning and spatio-temporal relation encoding for modeling gait style of models. The dataset contains 4680 gait sequences of six categories of styles, of which 3276 samples are used for training, 468 samples for validation, and 936 samples for testing. After normalization and segment segmentation, the key point flow is mapped into a time-aligned pose map, and input into a hierarchical network that integrates local joint interaction, spatial cooperative propagation and stage gating mechanism for encoding. Experimental results show that the accuracy of the method reaches 94.6%, Macro-F1 reaches 93.4%, feature clustering purity reaches 91.7%, and the average inference time is 1.21s. The framework can support gait analysis and interpretable style representation of models, and serve fashion action understanding and retrieval in computer vision.*

KEYWORDS: *Posture feature learning; Model gait; Style modeling; Spatio-temporal relation encoding*

1 Introduction

Gait style modeling belongs to the fine-grained representation task in the intersection scene of human motion analysis and visual understanding. Its goal is not to complete identity discrimination in the general sense, but to extract dynamic information with style attributes such as posture rhythm, limb coordination, center of gravity transfer and trajectory control from continuous walking process, and convert it into feature expressions that are computable, comparable and retrivable. Compared with ordinary gait recognition, model gait has stronger structural features in stride organization, arm swing amplitude, pelvic mobilization, shoulder and neck posture and turning control. At the same time, it is accompanied by camera viewpoint change, clothing swing, heel height difference and action rhythm fluctuation, which puts forward higher requirements for posture sequence modeling, spatio-temporal relationship coding and style boundary characterization. In the context of computer vision, stable modeling of model gait style not only provides the underlying support for fashion video retrieval, digital human motion driving, virtual runway generation and stylized motion analysis, but also extends the application range of skeleton sequence learning in fine-grained behavior understanding.

Existing researches have provided a solid technical foundation for skeleton-based gait modeling and temporal feature extraction. Liu X et al. studied the symmetrical structure modeling in skeleton gait recognition, and proposed a symmetic-driven super-feature graph convolutional network to enhance the expression of the topological relationship between the

*shiruoqi2024@163.com
https://doi.org/10.65102/is20261083 Published: 10 June, 2026

left and right limbs of the human body [1]. Li H et al. studied the representation of spatio-temporal slice features in gait recognition, and proposed GaitSlice model to describe the local dynamic changes in gait sequence with slicing structure [2]. Li N and Zhao X studied the effect of gait cycle prior on the stability of skeleton recognition, and proposed a robust skeleton gait recognition method fused with periodicity constraints [3]. Rashmi M and Guddeti studied the human recognition mechanism of 3D skeleton gait features, and proposed an identity recognition system combined with LSTM to improve the expression ability of time dependence [4]. Zhang C et al. studied the influence of spatial transformation on skeleton-based gait alignment, and proposed a spatial transformation network for skeleton-based gait recognition to weaken the offset caused by view perturbation [5]. These studies have driven the development of skeleton-based gait analysis from graph convolution, temporal slicing, periodic priors, cyclic modeling, and spatial alignment.

Under more complex modeling conditions, Jun K et al. studied the joint expression of skeleton features and gait features, and proposed a hybrid deep neural network framework for multi-modal fusion in pathological gait recognition [6]. Xu C et al. studied human grid-based gait modeling under occlusion conditions, and proposed a human grid-based gait recognition method for occlusion perception to improve the identificability in complex scenes [7]. Yan J et al studied the organization of SMPL parameters in graph structure and proposed GaitSG model to realize human gait recognition based on graph structure [8]. Model gait styles are not exactly equivalent to general gait categories. Style differences are often reflected in the coordinated relationship between multiple joint groups, such as the coupling between torso stability and lower limb propulsion rhythm, the synergy between swing arm radian and foot Angle, and the linkage between stride length and longitudinal body stretch. It is difficult to completely retain such hierarchical style information by only relying on single-layer keypoint input or simple time accumulation. At the same time, the gait video of models usually has factors such as camera position change, background disturbance and clothing deformation, and the local posture may fluctuate significantly in a short time. Without joint modeling of joint connection, temporal stage division and cross-scale feature flow, the style representation is easy to stay at the surface action difference, and it is difficult to form a stable discrimination boundary. Based on this understanding, this paper designs a technical path that combines posture sequence construction, multi-layer feature learning and spatio-temporal relationship coding. On the basis of keypoint graph representation, local joint interaction, global gait context aggregation and phasic timing coding are introduced to establish a computational model for model gait style. This method aims to enhance the separability and interpretability of style features, and provide a reliable representation basis for subsequent style recognition, style retrieval and action generation tasks. On this basis, the style discrimination ability under different module configurations is compared, and the stability, adaptability and computational efficiency, configuration differences and results changes in fine-grained gait representation are evaluated through performance analysis and ablation validation.

2 Related work

The gait style modeling driven by posture feature learning relies on both the spatio-temporal representation ability in the skeleton sequence and the stable description of style differences at the fine-grained level. Related research can be roughly divided into four directions: key point extraction and continuous gait representation, multi-feature fusion and cross-view modeling, hierarchical refinement and global local joint learning, style transfer and action retargeting.

In the aspect of keypoint-driven gait representation, Han K and Li X studied intermittent

gait image recognition supported by human skeleton keypoint extraction, and proposed a recognition method based on continuous recovery of keypoints to improve gait expression in discontinuous sequences [9]. Yousef R N et al. studied the joint representation of model-driven features and model-free deep features, and proposed a dual-path deep feature fusion method to enhance discriminant stability in gait recognition [10]. Chen J et al. studied the cross-view gait representation aggregation mechanism and proposed GaitAMR method to complete cross-view recognition through multi-feature aggregation [11]. Pan H et al. studied the joint representation of complete view and high-level pose, and proposed a high-level pose gait recognition method for complete view to enhance the robustness under the condition of view coverage [12].

In terms of skeleton refinement and spatio-temporal modeling, Wang R et al. studied the multi-level skeleton guided gait refinement process and proposed a multi-level skeleton guided optimization method to improve the contribution of key parts to recognition results [13]. Deng M et al. studied the expression of gait dynamics under positive view sequence, and proposed a gait dynamics recognition method with deep learning to improve the feature utilization efficiency under positive view condition [14]. Li Z et al. studied multi-feature expression and periodic segment time modeling, and proposed a gait recognition method combined with periodic part modeling to strengthen the utilization of repetitive motion rhythm [15]. Wei S et al. studied the global and local fusion of skeleton gait in the field scene, and proposed the GaitDLF method to improve the recognition ability in complex environments through the collaboration of global and local features [16]. Zhang Z et al. studied the fusion of multi-scale time dimension and global local feature, and proposed GaitMGL method to enhance the effect of information integration on different time scales [17]. Li R et al. studied the automatic coding representation of gait under the constraints of cognitive model, and proposed GaitAE method to generate more compact potential gait features [18].

In order to more clearly compare the technical path and application characteristics of related research, Table 1 summarizes the research content and modeling focus of representative methods.

Table 1: Comparison of representative related studies

Study	Method	Main Content	Applicable Characteristics	Reference
Han K et al.	Keypoint extraction method	Intermittent gait image recognition	Suitable for sequence recovery	[9]
Chen J et al.	Multi-feature aggregation method	Cross-view gait recognition	Emphasizes view adaptation	[11]
Pan H et al.	Full-view posture modeling method	High-level posture-based gait recognition	Emphasizes view completeness	[12]
Wang R et al.	Multi-layer skeleton refinement method	Key-part-guided recognition	Emphasizes hierarchical optimization	[13]
Li Z et al.	Periodic temporal modeling method	Multi-feature gait recognition	Emphasizes rhythmic expression	[15]
Zhang Z et al.	Multi-scale fusion method	Joint global-local recognition	Emphasizes scale collaboration	[17]

In terms of action style calculation, Hu L et al. studied semantic-guided human action

style transfer, and proposed a style transfer method based on diffusion model to improve the continuity and semantic consistency of action style generation [19]. Zhao Q et al. studied the Pose prior constraints in cross-domain action retargeting and proposed a pose-to-motion method to improve the structure preservation ability in the pose-to-action mapping process [20]. This kind of research shows that action style does not rely solely on static skeleton morphology, but is closely related to posture stage organization, temporal relationship transmission and latent style space construction.

From the perspective of technology evolution path, the existing research has gradually moved from single skeleton recognition to multi-source representation, relational modeling and latent space constraints. The dual-path deep feature fusion method emphasizes the complementarity of model-driven features and data-driven features, which provides a reference for the joint expression of explicit motion information and implicit style cues in model gait. The modeling of positive view gait dynamics also shows that the relationship between gait rhythm and view condition is not simple, which is especially critical for the sequence expression of catwalk scenes with large changes in camera position. The global local fusion strategy and automatic coding representation method further illustrate that high-quality gait modeling needs to consider local motion details, global structure organization and latent feature compression at the same time. In contrast, model gait style does not only serve for recognition tasks, but also involves style discrimination, style retrieval and action generation. Therefore, it is difficult to fully describe the coupling relationship between stride control, body stretch and posture temperament by only relying on the traditional identity discrimination framework. On the other hand, the research of action style transfer and action redirection indicates that style computation must deal with semantic continuity, structure preservation and dynamic mapping consistency. Therefore, the continued extension of related work should not stop at the coarse-grained recognition level of gait appearance differences, but should turn to the stable representation of style states. This means that the model should not only maintain the continuity of the stage order in the time dimension, but also preserve the linkage strength between the torso, upper limbs and lower limbs in the spatial dimension, but also form a compact aggregation of the same class and a clear boundary between different classes in the embedding space. Only when the modeling of these three levels is done simultaneously, the style features in the gait of the model will not be covered by the general motion information.

It can be seen that there are obvious differences in task objectives between modeling gait style modeling and general gait recognition. The former focuses on the continuous expression of style states and emphasizes the fine-grained relationship between stride control, body stretch, trunk stability and limb coordination. The latter focuses more on category discrimination or identity discrimination, and focuses on separability and recognition accuracy. For this reason, when the relevant methods are transferred to the model gait scene, they cannot only retain the original recognition framework, but also need to supplement the ability of pose sequence organization, stage relationship expression and spatio-temporal coupling modeling. Such a technical shift enables the subsequent methods to focus on sequence construction, multi-layer feature extraction and relation encoding, and also makes the main research line of this paper more focused. At the same time, this distinction also indicates that style cues in model gait cannot be simply compressed into a single action label, but should be regarded as a composite representation composed of multiple stages, multiple regions, and multiple dynamic relationships.

3 Methods

3.1 Modeling gait pose sequence construction and multi-layer feature learning model

The model gait pose sequence construction and multi-layer feature learning model takes video-level key point trajectories as input, and the goal is to preserve the rhythm organization, body extension amplitude, torso control state and lower limb propulsion relationship in consecutive gait frames, so that the style differences in the model gait can enter the subsequent network in a structured form. The original video is first processed by human detection and pose estimation to obtain the two-dimensional joint coordinates and confidence scores of each frame, and then the initial pose sequence is composed in time order. Due to the obvious scale fluctuations caused by lens distance, model height and camera position switching, local action amplitude and overall pose proportion are easy to be mixed in the same space if the original coordinates are directly used. Therefore, at the beginning of sequence construction, center normalization and length normalization are performed on the joint coordinates to make the pose points of different samples enter a unified metric space. The normalization expression is given in Equation (1).

$$Q_t^{(i)} = \frac{P_t^{(i)} - C_t}{\sqrt{\frac{1}{J}\sum_{j=1}^{J}\left\| P_t^{(j)} - C_t \right\|_2^2} + \varepsilon} \tag{1}$$

Here, $P_t^{(i)}$ represents the original coordinates of the i joint in the t frame, C_t represents the human body center position in the current frame, J represents the total number of joints, ε represents the stability term, and $Q_t^{(i)}$ represents the normalized joint coordinates. The function of Equation (1) is to compress the individual body size and shooting scale differences, so that the subsequent model can perceive the style-related dynamic structure more centrally, rather than being dominated by external imaging conditions.

As shown in Fig. 1, the gait pose sequence of the model goes through four steps of posture cleaning, temporal alignment, segment segmentation and structure encoding in turn before entering the deep learning, and then it is sent to the multi-layer feature learning branch. In the figure, the left is the original keypoint sequence, the middle is the pose segment segmented by a fixed window, and the right is the fusion output after shallow, middle and deep parallel learning. Instead of simply compressing the long sequence into a whole vector, this method first preserves the local gait phases, and then builds a continuous representation link from joint motion to regional coordination to global rhythm layer by layer.

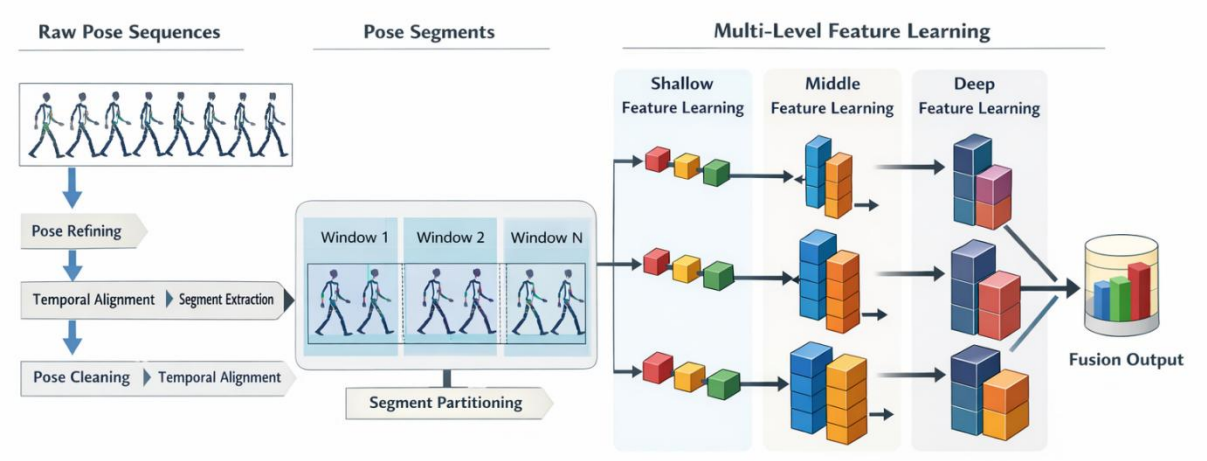


Figure 1: Modeling gait pose sequence construction and multi-layer feature learning process

In order to maintain the continuity of gait phase, the model uses a fixed window and sliding step to segment the time series jointly, and splits the long video into parallel gait segments. The purpose of this is to separate out the local action phases such as the step lift, transition, foot drop, and arm swing from the long sequence, so that each segment corresponds to a relatively stable action unit. The fragment tensor is constructed in the same way as in Equation (2).

$$S_m = \left[Q_{(m-1)\Delta+1}, Q_{(m-1)\Delta+2}, \cdots, Q_{(m-1)\Delta+L}\right] \tag{2}$$

Here, S_m represents the m pose segment, Δ represents the sliding step, and L represents the segment length. Equation (2) transforms the complete sequence into multiple local timing units, so that local action changes are no longer submerged by the global average, and also leave a clear boundary for subsequent cross-segment rhythm aggregation.

After segment segmentation, the model no longer treats the joints as isolated points, but constructs a weighted pose graph according to the connection relationship of the human skeleton and the consistency of joint motion. The reason for this is that the style difference in the gait of the model does not only appear in a single joint, but is more reflected in the structural synergy between shoulder and hip interaction, torso stretch and arm swing and stride. Therefore, both topological connectivity and dynamic similarity need to be preserved in the structural encoding stage. The construction of edge weights is given in Equation (3).

$$A_t(i,j) = \lambda G(i,j) + (1-\lambda)\exp\left(-\frac{\left\|V_t^{(i)} - V_t^{(j)}\right\|_2^2}{\sigma^2}\right) \tag{3}$$

Here,$A_t(i,j)$ represents the connection weight between joint i and joint j in frame t, $G(i,j)$ represents the skeleton topological adjacency, $V_t^{(i)}$ represents the joint velocity, λ represents the balance coefficient between topological prior and dynamic similarity, and σ represents the scale parameter. Equation (3) makes the pose graph reflect not only "which joints are connected", but also "which joints are moving in a similar way at the current time", which is more suitable for gait style modeling rather than common posture classification.

In the multi-layer feature learning stage, the model sets up a shallow pose branch, a mid-level regional coordination branch and a deep global rhythm branch. The shallow branch mainly responds to high-frequency motion nodes such as ankle, knee and wrist, the middle

branch focuses on the shoulder and hip linkage, center of gravity transfer and arm swing amplitude, and the deep branch aggregates the overall style trend of the whole gait. Each layer completes the feature mapping through convolution, normalization and nonlinear activation, and the calculation process of the r-th layer is shown in Equation (4).

$$F_r = \phi\left(BN(W_r * X_r + b_r)\right) \tag{4}$$

Here, X_r represents the input feature of the r layer, W_r and b_r represent the convolution kernel parameter and bias term, $BN(\cdot)$ represents the normalization operation, $\phi(\cdot)$ represents the nonlinear activation function, and F_r represents the output feature of the layer. Equation (4) is not a simple convolution calculation description, but a basic mapping unit common to the branches of the three layers. Through repeated mapping under different receptive fields, local joint motion and regional pose patterns are pushed layer-by-layer into higher-level representations.

As shown in Fig. 2, the three branches are not concatenated directly after output, but double reweighting of local attention and temporal attention is performed first. The middle and lower paths of the graph are responsible for the screening of joint dimension response, the middle path is responsible for the estimation of segment rhythm intensity, and the upper path completes cross-layer fusion and output mapping. The reason for this arrangement is that the style discrimination of the model gait depends on both the key action segments and the local pose regions. If we only do channel splicing, it is easy to bring redundant responses unrelated to style into the final representation, weakening the effect of fine-grained modeling.

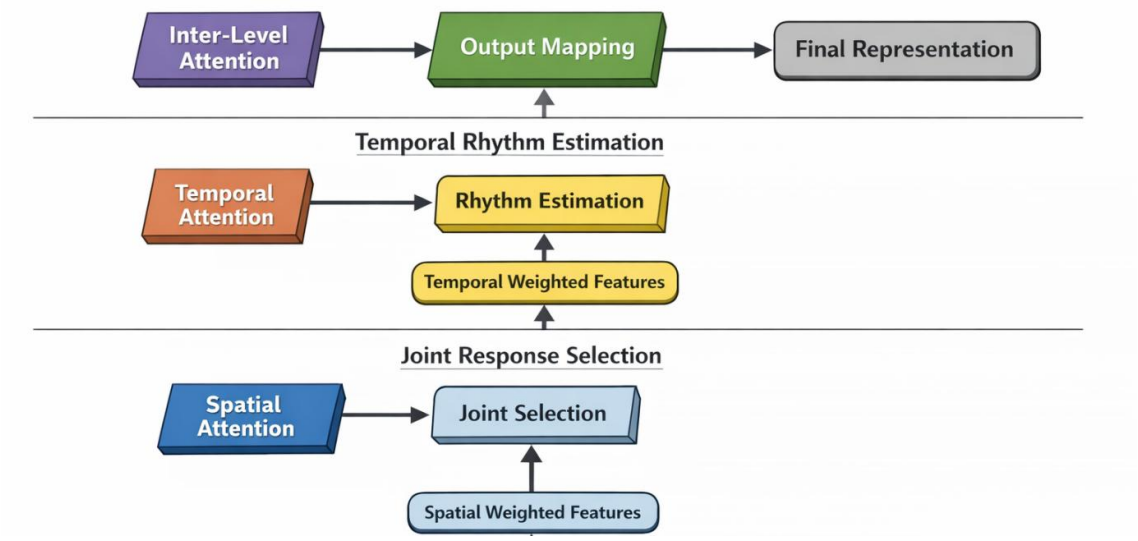


Figure 2: Multi-layer feature learning and cross-layer fusion architecture

In the attention modeling stage, the model first generates local attention based on the joint response strength, and then generates temporal attention based on the rhythm stability between segments. The joint weighted segment features are shown in Equation (5).

$$\alpha_m = \mathrm{Softmax}(W_a \cdot \mathrm{Pool}(F_m)), \quad U_m = \alpha_m \odot F_m \tag{5}$$

Here, F_m represents the input feature of the m segment, ool$(\cdot)$ represents the sink operator, W_a represents the attention mapping parameter, α_m represents the attention weight, $\odot$ represents the element-wise multiplication, and U_m represents the reweighted segment feature. Equation (5) introduces "which part of actions is more important" and

"which time segment is more critical" into the segment expression at the same time, so that style-related local actions are strengthened, while short-term noise and accidental disturbance are depressed.

After obtaining the enhanced features of all segments, the model continues to perform cross-layer gated fusion to compress the multi-segment and multi-level representations into a unified style vector. The fusion here is not a simple sum, but different weights are assigned according to the contribution of the fragments in the overall style expression. The gated fusion expression is given in Equation (6).

$$\beta_m = \frac{\exp(\psi(U_m))}{\sum_{k=1}^{M} \exp(\psi(U_k))}, \quad Z = \sum_{m=1}^{M} \beta_m U_m \tag{6}$$

Here, β_m represents the contribution weight of the m segment in the final representation, $\psi(\cdot)$ represents the gating scoring function, M represents the total number of segments, and Z represents the final gait style representation vector. Equation (6) integrates local pose differences, regional synergistic relationships and overall rhythm patterns into the same representation space, so that the output vector not only retains the hierarchical structure of gait actions, but also has good style discrimination ability.

After the above construction process, the pose points in the original video are converted into the gait style features of the model with scale consistency, structural constraints and hierarchical expression ability. The local joint dynamics, regional coordination patterns and global rhythm trends are organized and aggregated in the same computing link, which provides a unified input for the spatio-temporal relationship coding framework in the next section, and also enables the style differences in the model gait to enter the subsequent modeling in a more stable way.

3.2 Gait style modeling framework integrating spatio-temporal relationship encoding

After completing the pose sequence construction and multi-layer feature learning, the model enters the stage of spatio-temporal relationship encoding. Instead of viewing segment features as separate local responses, this stage organizes the rhythmic advancement, torso control, arm swing amplitude, and lower limb extension of the model's gait as a continuous stream of style states. For model gait, style differences are not only limited to single frame posture, but also reflected in the transition mode between adjacent segments, the linkage strength between body regions, and the order of multiple action stages in the overall rhythm. Therefore, the encoding of spatio-temporal relations here is no longer just "weighting the fragments once", but to establish a unified modeling framework that can simultaneously express phase continuity, regional synergy and global style stability.

As shown in Fig. 3, the segment features obtained in Section 3.1 first enter the temporal relation coding unit, then enter the spatial collaborative propagation unit, and then complete the reorganization through the stage gating and cross-stage aggregation module, and finally output the style embedding and category probability. The left part of the figure shows the segment input, phase label and region division information, the middle part shows the temporal relationship matrix, spatial relationship matrix and relationship propagation path, and the right part corresponds to the style embedding generation and classification discrimination results. The core function of Fig. 3 is to put the calculation process of "how fragments are connected", "how regions are linked" and "how styles are aggregated" into the same structure, so as to ensure that each step corresponding to the subsequent formula has a

clear module destination.

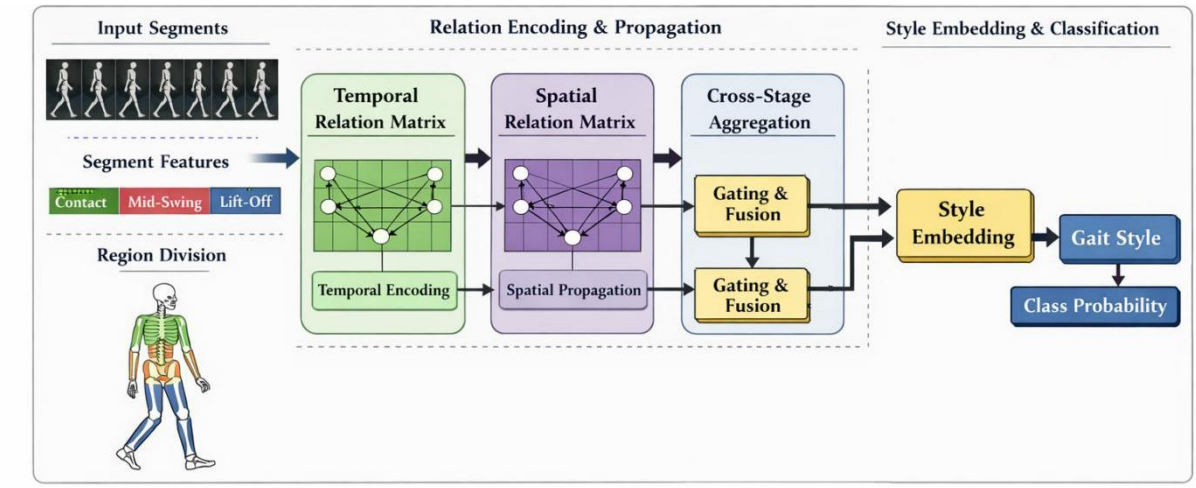


Figure 3: Gait style modeling framework for models incorporating spatio-temporal relation encoding

In order to describe the rhythm continuity between adjacent segments, the model first explicitly models the temporal relationship between segments. Instead of directly using the simple distance as the similarity criterion, both the segment feature difference and the phase distance are incorporated into the temporal relationship matrix. The reason for this treatment is that style continuation in the gait of the model depends both on whether the actions are similar and on the relative position of the two segments in the complete gait cycle. The temporal relation is defined in Equation (7).

$$R_{mn}^{t} = \exp\left(-\frac{\|U_m - U_n\|_2^2}{\tau_1}\right)\exp\left(-\frac{|p_m - p_n|}{\tau_2}\right) \tag{7}$$

Here, R_{mn}^{t} represents the weight of the temporal relationship between segment m and segment n, U_m represents the input feature of the m segment, p_m represents the phase position of the segment in the complete gait cycle, τ_1 and τ_2 represent the feature difference attenuation parameter and phase distance attenuation parameter, respectively. The function of Equation (7) is not to simply screen adjacent segments, but to establish the time relationship under the double constraints of "action simila-phase proximity", so that the model gives priority to transmitting the information of segments with continuous rhythm and consistent style.

Temporal relations alone are still not sufficient to express the style formation mechanism in the gait of models. The gait style of the model often relies on the collaborative changes between torso extension, shoulder-hip linkage, arm swing radian and stride control, so the model continues to establish a regional collaborative matrix in the spatial dimension. Instead of using the fixed adjacency method, the response consistency between regional features is used to adaptively calculate the spatial relationship. The spatial synergy relationship is shown in Equation (8).

$$R_{ab}^{s} = \frac{H_a^{\top} H_b}{\|H_a\|_2 \|H_b\|_2 + \epsilon} \tag{8}$$

Here, R_{ab}^{s} represents the synergy strength between region a and region b, H_a

represents the aggregated feature vector of the a functional region, and ϵ represents the stability term. Equation (8) is used to measure the degree of synergy of different body regions in the current sample. Synergy values increase significantly when trunk aligns with lower limb propulsion and upper limb swing echoes the overall rhythm, which enables shoulder-hip linkage, swing arm control, and lower limb propulsion to enter the model with explicit structures rather than being implicit in coarse-grained features.

After obtaining the temporal relationship matrix and spatial collaboration matrix, the model starts to perform joint propagation update. Here, the update is not simply adding the two relationship matrices, but let the fragment state propagate synchronously in the time neighborhood and the region neighborhood, and then complete the nonlinear reorganization through the activation function. Such a design can avoid the bias brought by a single relationship pathway, so that each segment can see both "continuous changes in front and back stages" and "linkage patterns of body regions" when updated. The joint propagation expression is given in Equation (9).

$$Y_m^{(l+1)} = \rho\left(\sum_{n=1}^{M} R_{mn}^{t} W_t^{(l)} U_n + \sum_{b=1}^{B} R_{ab}^{s} W_s^{(l)} H_b\right) \tag{9}$$

Here, $Y_m^{(l+1)}$ represents the segment state after propagation at layer l+1, $W_t^{(l)}$ represents the temporal propagation weight matrix, $W_s^{(l)}$ represents the spatial propagation weight matrix, $\rho(\cdot)$ represents the activation function, and B represents the number of functional regions. The core role of Equation (9) is to implement the temporal proximity constraint and spatial collaboration constraint in the same step of propagation, so that the single segment representation is promoted from the original local action level to the relationship expression level with context support.

If the propagation process only emphasizes relational integration, all segments may gradually tend to be similar, and style boundaries will be significantly weakened. To preserve the original rhythm difference and local action intensity in the model gait, the model adds a stage gating unit after propagation to adaptively blend the updated state and the original segment features. The goal of the gating mechanism is not to simply preserve old features, but to establish a balance between "structural propagation gain" and "original style personality". The gated recombination expression is given in Equation (10).

$$g_m = \sigma\left(W_g[Y_m^{(l+1)}; U_m] + b_g\right), \quad \hat{Y}_m = g_m \odot Y_m^{(l+1)} + (1 - g_m) \odot U_m \tag{10}$$

where g_m represents the gating weight of the m fragment, $\sigma(\cdot)$ represents the Sigmoid function, W_g and bg represent the gating mapping parameters, $[\,;\,]$ represents the splicing operation, and $\hat{Y}_m$ represents the fragment representation after gated recombination. Equation (10) avoids excessive smoothness caused by relationship propagation through differential fusion of propagation state and original state, and also preserves the rhythm tension and movement style in the gait of the model.

As shown in Fig. 4, the sequence of segments after gated recombination is fed into the cross-stage aggregation and style discrimination module. The structure unfolds along the computational link of "stage-keeping-weight assignment-global aggregation-category output". The bottom path preserves the original temporal order of the segments, which is used to maintain the continuity between gait phases. The middle path calculates the stage weight according to the contribution of the segment in the overall style expression, which is used to distinguish the key action segment from the weak response segment. The upper path generates

the global style embedding after completing the weighted aggregation and outputs the corresponding style category scores. Firstly, the contribution of segment features is redistributed according to the spatio-temporal relationship, and then a unified style state flow is formed by cross-stage aggregation, so that the continuous action stages such as lift preparation, body stretching, stride expansion, and foot bandana are differentiated in the final representation.

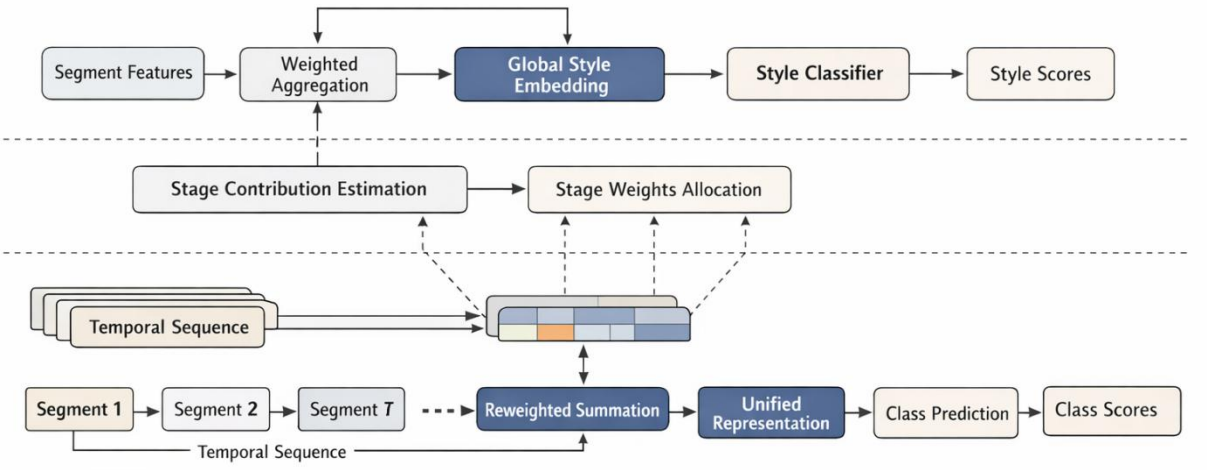


Figure 4: Cross-stage aggregation and style discrimination structure after encoding spatio-temporal relations

In the cross-stage aggregation stage, the model generates stage weights according to the contribution of each segment to the overall style expression, and forms a global style embedding on this basis. Here the weights are no longer given by the fixed rules, but are determined by the discriminative power of the segments themselves. The weights are calculated and aggregated as shown in Equation (11).

$$\omega_m = \frac{\exp(\eta^\top \hat{Y}_m)}{\sum_{k=1}^{M} \exp(\eta^\top \hat{Y}_k)}, \quad E = \sum_{m=1}^{M} \omega_m \hat{Y}_m \tag{11}$$

Here, ω_m represents the contribution weight of the m segment in the global representation, η represents the learnable projection vector, and E represents the final style embedding. Equation (11) makes the stage with high recognition gain a larger proportion in the overall representation, so as to highlight the key style segments in the gait of the model and reduce the interference of the unstable segments of the rhythm on the global results.

After obtaining the global embedding, the model outputs the style probability distribution through the classification head. The classification mapping is given in Equation (12).

$$\hat{y} = \mathrm{Softmax}(W_c E + b_c) \tag{12}$$

where $\hat{y}$ represents the predicted probability distribution of style classes, W_c and b_c represent the classification head weight and bias, respectively. Equation (12) directly maps the style embedding encoded by relation to the category space to form the final style discrimination result. The classification result here is not just a label output, but a centralized reflection of the whole spatio-temporal relationship encoding link at the category level.

In order to make the samples of the same class more compact in the embedding space and open the boundary of the samples of different classes, the model added structural constraints

outside the classification objective. The reason for this is that modeling of model gait style requires not only correct classification, but also interpretable organizational form of style distribution in the embedded space. The joint loss is given in Equation (13).

$$\mathcal{L} = \mathcal{L}_{ce} + \mu_1 \sum_{i,j \in \mathcal{S}} \| E_i - E_j \|_2^2 + \mu_2 \sum_{i,k \in \mathcal{D}} \max(0, \delta - \| E_i - E_k \|_2) \tag{13}$$

Here, $\mathcal{L}_{ce}$ represents the cross-entropy loss, $\mathcal{S}$ represents the set of samples from the same class, $\mathcal{D}$ represents the set of samples from different classes, μ_1 and μ_2 represent the balance coefficient, and δ represents the inter-class interval. Equation (13) integrates label supervision and structural constraints into the same objective function, so that the style discrimination result depends not only on the category label, but also on the boundary organization in the embedding space.

In order to further constrain the continuity of style state flow on adjacent stages, the model also adds a smoothing term to the state changes before and after aggregation, which is used to weaken sudden jumps. This term mainly controls the transition amplitude between phases, so that local action changes and the overall rhythm can be consistent. The stage smoothness constraint is given in Equation (14).

$$\mathcal{L}_{sm} = \frac{1}{M-1} \sum_{m=1}^{M-1} \| \hat{Y}_{m+1} - \hat{Y}_m \|_2^2 \tag{14}$$

Here, $\mathcal{L}_{sm}$ represents the stage smoothing loss and $\hat{Y}_{m+1} - \hat{Y}_m$ represents the state difference between adjacent segments. Equation (14) is used to maintain the smooth transition of gait style on the time axis, so that the local action changes are consistent with the overall rhythm, and also let the style embedding have better stage continuity.

After the above steps, segment-level pose features are organized into style representations with temporal continuity, spatial synergy, and phase separability. The framework integrates local motion, regional linkage and global rhythm into the same relationship coding process, so that the style differences in the gait of the model can enter the subsequent performance analysis and ablation verification in a more stable way.

4 Results

4.1 Performance analysis of model gait style modeling based on posture feature learning

In this section, the proposed model is evaluated from three levels: classification performance, feature stability, and computational efficiency. The evaluation metrics include accuracy, macro average F1 score, cluster purity, and average inference time. The precision is used to measure the overall accuracy of the style category determination, the macro-average F1 value is used to reflect the recognition balance between each class, the cluster purity is used to observe the compactness of the style samples in the embedding space, and the average inference time is used to describe the response efficiency of the model in the actual video analysis. The experimental platform uses Intel Core i7 processor, RTX 3090 graphics card and 24GB video memory. The dataset consists of 4680 gait sequences covering six types of typical style states. The training set, validation set and test set are 3276, 468 and 936 sequences respectively. All samples are processed by keypoint normalization, gait segment

segmentation and pose image reconstruction before input.

In order to verify the effectiveness of the proposed model, ST-MLP, Pose-Transformer and Gait-RelationNet are selected as comparison models in this section, and the proposed model is denoted as PoseStyle-Net. ST-MLP mainly describes the basic Pose sequence mapping, Pose-Transformer emphasizes long-range dependency modeling, and Gait-RelationNet focuses on relationship propagation and local structural constraints. The four models are compared under the same data partitioning and training Settings, and the training results are shown in Fig. 5.

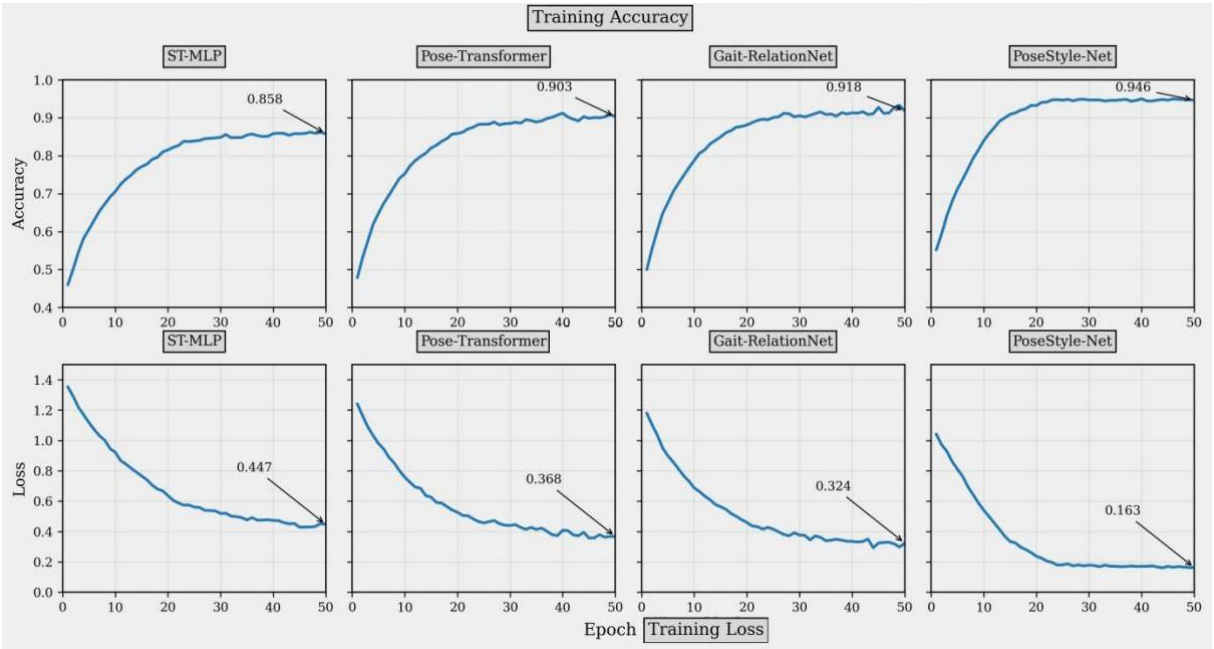


Figure 5: Accuracy versus loss change of four models during training

It can be seen from Fig. 5 that PoseStyle-Net has the fastest convergence speed, and the accuracy increases from 0.712 in the fifth round to 0.907 in the 15th round, stabilizes around 0.946 in the 25th round, and remains 0.946 in the 50th round. In contrast, ST-MLP, Pose-Transformer, and Gait-RelationNet achieve accuracies of 0.858, 0.903, and 0.918 at round 50, respectively. The loss curve also shows the same trend, the loss of PoseStyle-Net decreases from 1.041 to 0.181, and stabilizes to 0.163 in the 50th round, which is lower than 0.447, 0.368, and 0.324 of the other three models. The results show that the proposed model has more advantages in training stability and convergence efficiency.

Table 2 further compares the different models in terms of embedding representation quality and deployment complexity. Here, instead of repeating the accuracy curve or category recognition results, the comprehensive performance of each model is evaluated from three levels: feature clustering effect, model size and running resources.

Table 2: Comparison of representation quality and deployment complexity of different models

Model	NMI	Silhouette Coefficient	Number of Parameters / M	GPU Memory Usage / GB	Throughput (seq/s)
ST-MLP	0.781	0.412	6.3	3.8	57.4
Pose-Transformer	0.826	0.468	9.7	5.1	43.8
Gait-RelationNet	0.844	0.491	8.5	4.6	49.2
PoseStyle-Net	0.879	0.537	8.1	4.2	61.7

According to Table 2, PoseStyle-Net achieves the best results in two representation quality indicators, NMI and silhouette coefficient, reaching 0.879 and 0.537 respectively, indicating that the style embeddings learned by this model are stronger in category aggregation and boundary separation. At the same time, the parameter quantity of the proposed model is 8.1M, which is lower than the Pose-Transformer, the video memory occupation is 4.2GB, which is also lower than the two relational modeling baselines, and the throughput rate reaches 61.7 seq/s, indicating that the model has a good advantage in deployment efficiency while maintaining high representation quality. This indicates that the spatio-temporal relation encoding not only improves the classification results, but also enhances the organization effect of style features in the embedding space and maintains a good level of resource utilization.

The recognition performance of the model on different style categories is shown in Fig. 6.

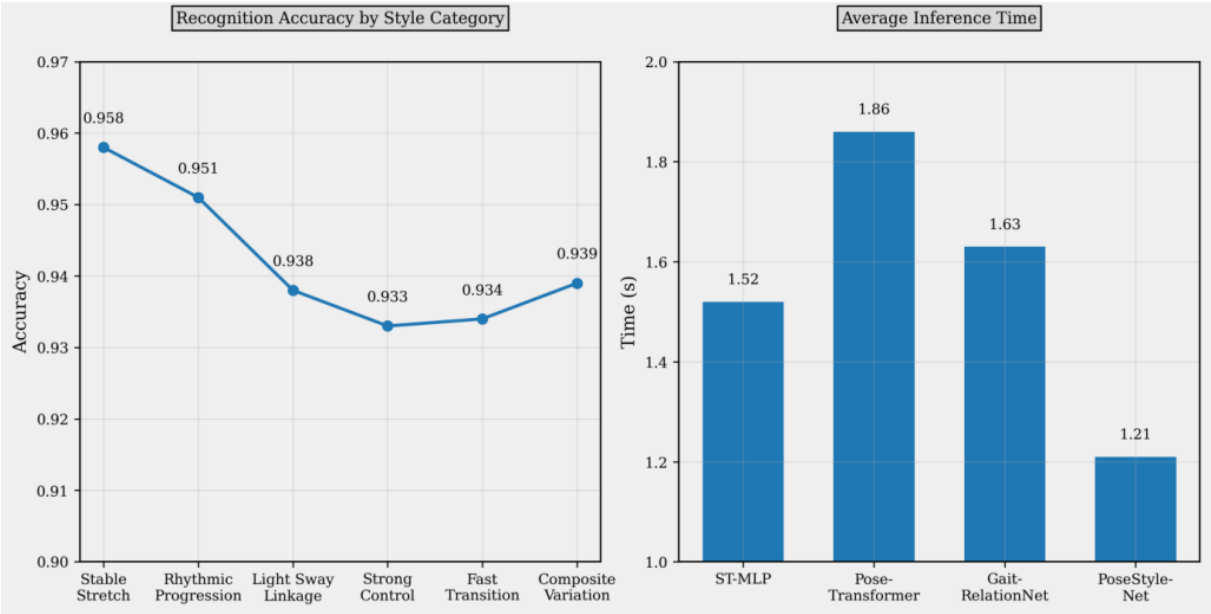


Figure 6: Recognition accuracy versus average inference time under different style categories

Fig. 6 shows that PoseStyle-Net maintains a high recognition rate on the six types of style states, and the recognition results of "stable stretching type" and "rhythm pushing type" are the most concentrated, with an accuracy of 0.958 and 0.951 respectively. On the two types of samples of "fast transition type" and "strong control type", the proposed model still maintains the recognition level above 0.93, while the comparison model decreases more significantly on these two types. This indicates that the spatio-temporal relation coding has a stronger ability to depict stage boundaries and local linkage, and can maintain a clearer discriminative boundary between similar style categories.

Table 3 further statistics the performance of PoseStyle-Net under a variety of perturbation test scenarios, which is used to examine the adaptation ability of the model under changes in actual video conditions. Instead of repeating the category accuracy, the table looks at how well the performance is maintained under perturbed conditions.

Table 3: Robustness statistics of PoseStyle-Net under different perturbation scenarios

Perturbation Scenario	ACC Retention Rate / %	Macro-F1 Retention Rate / %	Mean Confidence	Performance Fluctuation Level
Slight camera shift	98.6	98.1	0.941	Low
Enhanced clothing swing	97.9	97.4	0.933	Low
Accelerated rhythm	97.1	96.8	0.927	Medium
Slowed rhythm	97.4	97.0	0.929	Medium
Turning transition	96.8	96.3	0.921	Medium
Partial background occlusion	96.5	96.1	0.918	Medium

Table 3 shows that the distribution of precision, recall and F1 value of the model on different style categories is relatively balanced, and no category is significantly lower, which indicates that the model can not only identify significant style features, but also deal with samples with relatively close style boundaries. Pairwise t-test shows that compared with Pose-Transformer and Gait-RelationNet, PoseStyle-Net achieves significant improvements in accuracy and macro average F1 value ($p<0.01$). In general, the modeling method based on posture feature learning and spatio-temporal relationship coding can more effectively organize the local motion, regional linkage and global rhythm in the model gait, so as to provide a stable basis for the subsequent gait style discrimination results and ablation experiment analysis.

To show the stability differences of different models under repeated experimental conditions more intuitively, Table 4 shows the accuracy standard deviation, Macro-F1 standard deviation and average inference time after five repeated training. It can be seen that PoseStyle-Net also exhibits a smaller fluctuation range while maintaining a high recognition accuracy, which indicates that the model has more stable parameter convergence characteristics under different batches of training.

Table 4: Stability comparison of repeated experiments with different models

Model	ACC Standard Deviation	Macro-F1 Standard Deviation	Average Inference Time / s
ST-MLP	0.019	0.021	1.52
Pose-Transformer	0.014	0.016	1.86
Gait-RelationNet	0.011	0.013	1.63
PoseStyle-Net	0.007	0.008	1.21

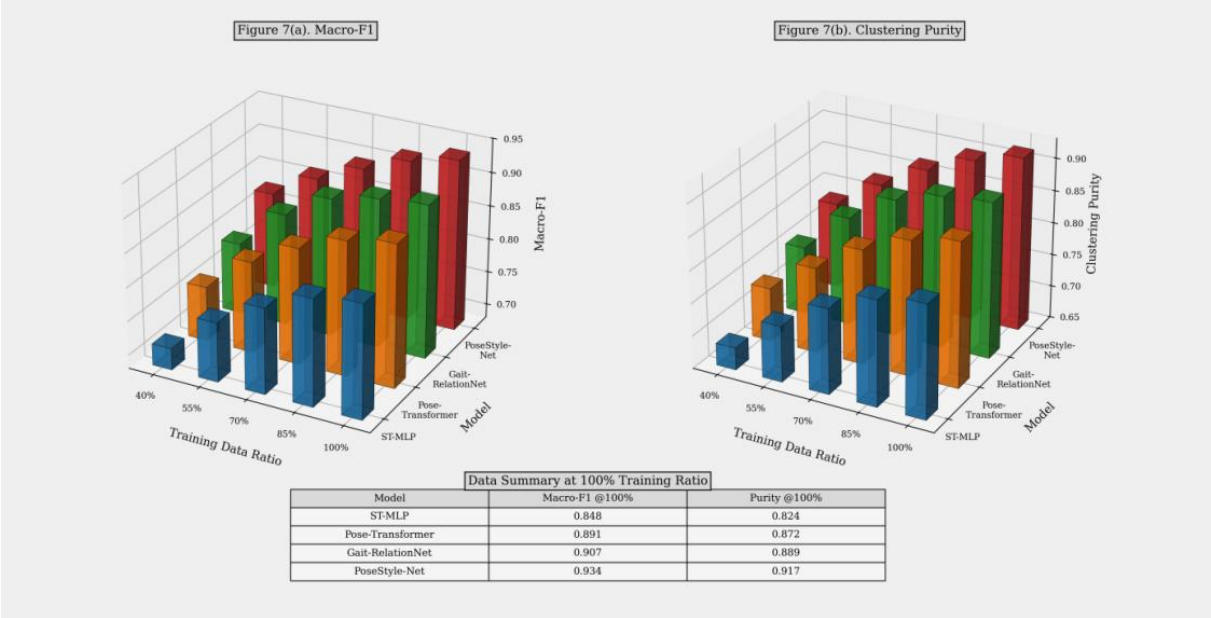
Table 4 further shows that the fluctuation amplitude of PoseStyle-Net in accuracy and macro-average F1 value is lower than that of the other three models, indicating that spatio-temporal relation encoding can form a more stable way of feature organization during training. Combined with the overall performance results in Table 2, the proposed model not only dominates in accuracy, cluster purity and inference time, but also maintains good consistency under repeated experimental conditions, which indicates that the performance improvement of the model does not come from accidental convergence, but from the synergy

between relationship propagation, stage aggregation and weight assignment. Combined with Table 3, it can be found that the recognition accuracy of the model on different style categories does not appear category bias due to the enhanced stability, but still maintains high Precision and Recall on complex categories, indicating that the spatio-temporal relation encoding not only improves the overall recognition performance, but also enhances the ability to maintain boundaries between different style categories. Taken together, PoseStyle-Net shows strong advantages in accuracy, stability and efficiency. The model not only maintains high accuracy in the overall recognition results, but also shows smaller performance fluctuations under repeated experimental conditions, which indicates that the spatio-temporal relation coding can improve the style discrimination ability while maintaining good computational stability.

4.2 Gait style discrimination results and ablation experiment analysis

In order to further test the ability of the model to distinguish fine-grained style differences, this section conducts the analysis from three levels: style boundary preservation, neighborhood category separation, and module marginal contribution. ST-MLP, Pose-Transformer and Gait-RelationNet are still selected as comparison models, and the model in this paper is denoted as PoseStyle-Net. In addition to the overall performance, this section pays particular attention to the case of confusion between style-close categories and the actual contribution of different modules to the stability of style representation.

Fig. 7 shows the Macro-F1 versus cluster purity variation of each model with different proportions of training data. Fig. 7(a) corresponds to Macro-F1 changes, and Fig. 7(b) corresponds to cluster purity changes.



Data Summary at 100% Training Ratio

Model	Macro-F1 @100%	Purity @100%
ST-MLP	0.848	0.824
Pose-Transformer	0.891	0.872
Gait-RelationNet	0.907	0.889
PoseStyle-Net	0.934	0.917

Figure 7: Comparison of Macro-F1 and cluster purity for each model with different proportions of training data

As can be seen from Fig. 7, Macro-F1 and cluster purity of the four models show an upward trend as the proportion of training data increases from 40% to 100%, but there are differences in the increase of different models. The Macro-F1 of ST-MLP increased from 0.712 to 0.848, and the cluster purity increased from 0.684 to 0.824, and the increase slowed down after 85%. The Macro-F1 of Pose-Transformer increases from 0.761 to 0.891, and the

cluster purity increases from 0.732 to 0.872. Gait-RelationNet grows slowly after 70%, Macro-F1 is 0.907, and the cluster purity is 0.889. The bottom data frame lists the final values of the model at 100% scale. In contrast, Macro-F1 of PoseStyle-Net increases from 0.824 to 0.934, and cluster purity increases from 0.783 to 0.917, and keeps increasing in the range of 85% to 100%, indicating that spatio-temporal relationship coding is more stable for the organization of style boundaries.

In order to further investigate the ability of the model to distinguish between confusing style categories, this paper selected three gait style combinations with similar boundaries and cross movement organization methods for comparative analysis, and the specific results are shown in Table 5.

Table 5: Discrimination results for confusable style pairs

Style Pair	ST-MLP F1	Pose-Transformer F1	Gait-RelationNet F1	PoseStyle-Net F1
Stable Stretching Type – Rhythmic Progression Type	0.81	0.87	0.89	0.93
Light Swing Linkage Type – Integrated Variation Type	0.78	0.84	0.86	0.91
Strong Control Type – Rapid Transition Type	0.74	0.82	0.85	0.90

Table 5 shows that PoseStyle-Net achieves the highest F1 value on the three groups of confusing style pairs, among which "strong control type - fast transition type" has the most obvious improvement, which is 0.16 higher than ST-MLP and 0.08 higher than Pose-Transformer. This shows that the proposed model not only improves the overall recognition accuracy, but also enhances the discrimination strength between similar style states. To visually observe the differences in model outputs, Fig. 8 shows the style representation visualization results of different models on typical samples.

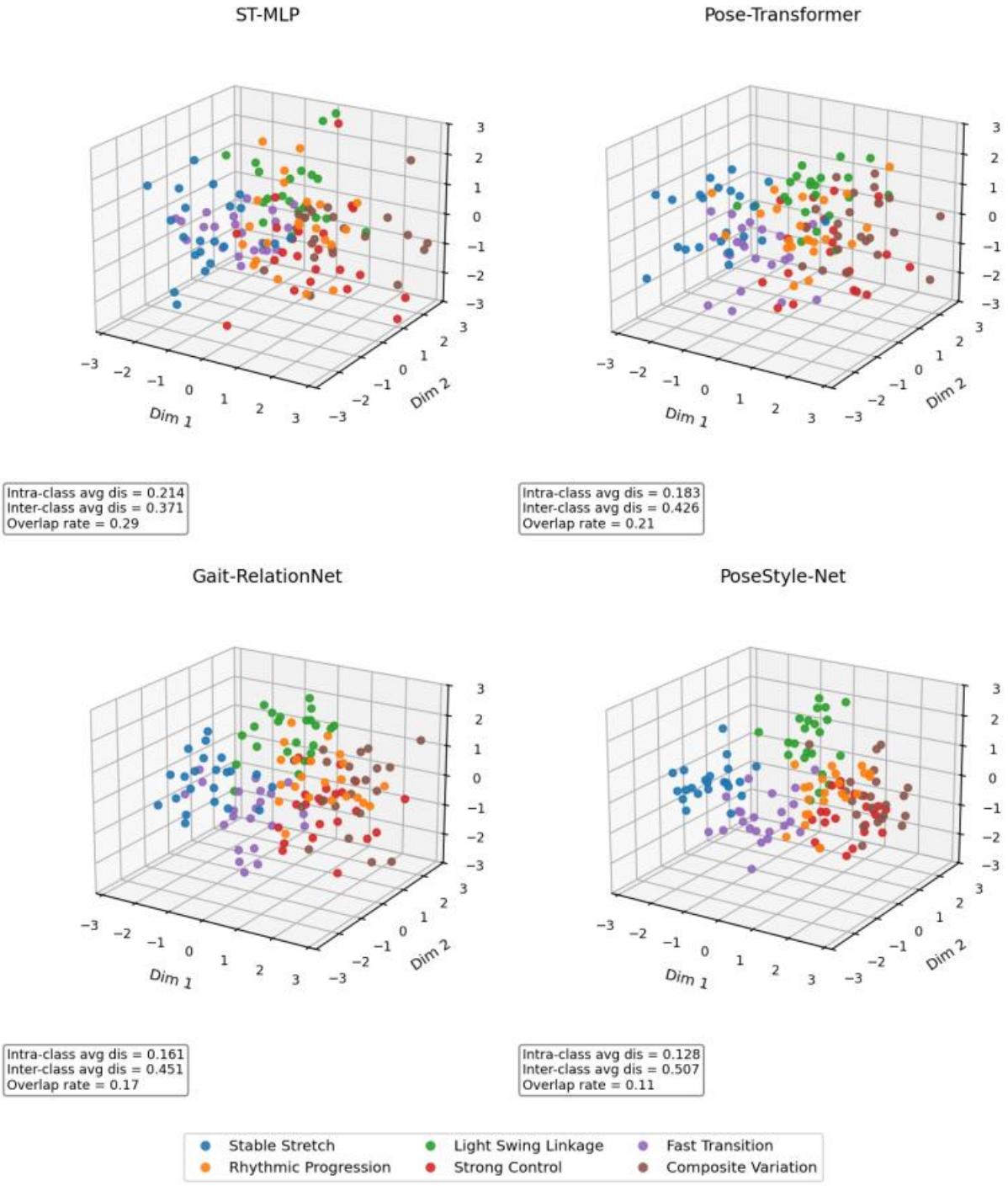


Figure 8: Visualization results of style representation of four models on typical samples

Fig. 8 shows that the intra-class average distance of ST-MLP is 0.214 and the inter-class average distance is 0.371. Pose-Transformer is 0.183 and 0.426, respectively; Gait-RelationNet is 0.161 and 0.451; PoseStyle-Net is 0.128 and 0.507, respectively. Meanwhile, the category overlap rate of PoseStyle-Net decreases to 0.11, which is lower than 0.29 of ST-MLP, 0.21 of Pose-Transformer, and 0.17 of Gait-RelationNet. It indicates that the fragment relation propagation and phase gating mechanism can compress the intra-class dispersion more effectively and enhance the separability between style states. To verify the marginal contribution of each component module, Table 6 presents the ablation experiment

results.

Table 6: Results of the PoseStyle-Net ablation experiments

Model Configuration	ACC	Macro-F1	Clustering Purity	Average Inference Time / s
Complete Model	0.946	0.934	0.917	1.21
Without Temporal Relationship Encoding	0.921	0.907	0.884	1.16
Without Spatial Collaborative Propagation	0.928	0.914	0.891	1.18
Without Stage Gating	0.933	0.919	0.896	1.17
Without Structural Constraint Term	0.937	0.923	0.901	1.21

According to Table 6, the performance of the model decreases most obviously after removing the temporal relationship encoding, and Macro-F1 decreases from 0.934 to 0.907, and cluster purity decreases from 0.917 to 0.884, indicating that rhythm continuity modeling plays a fundamental role in maintaining style boundaries. After removing the spatial collaborative propagation, the model's ability to describe the linkage of body regions is weakened, and the cluster purity is reduced to 0.891. After removing stage gating, the average inference time does not change much, but Macro-F1 decreases to 0.919, indicating that the gating mechanism mainly acts on style discrimination rather than computational efficiency. After removing the structural constraint term, the overall accuracy is still higher than most of the baselines, but the aggregation quality of the embedding space shows a decrease.

In summary, the spatio-temporal relationship coding, stage gating and structural constraint terms jointly support the stable discrimination of gait styles, and the role of temporal relationship coding is the most obvious. The results show that the modeling of gait style is not dependent on a single local action, but is based on the cooperative organization of multiple action stages and multiple body regions. Especially in the neighboring style categories, this collaborative modeling method can more effectively maintain the boundary differences, thereby improving the overall discrimination effect.

5 Discussion

Related experimental results and comparative analysis show that PoseStyle-Net exhibits higher recognition stability and clearer style boundaries in the modeling gait style task. Compared with ST-MLP, Pose-Transformer and Gait-RelationNet, the ACC of the proposed model reaches 0.946, Macro-F1 reaches 0.934, cluster purity reaches 0.917, and the average inference time is 1.21s. It is better than 0.861, 0.903, 0.918 and 1.52s, 1.86s, 1.63s and other comparison results. In the experiment of confusing style pairs, the F1 value of "strong control type and fast transition type" is increased to 0.90, which is 0.16 higher than that of ST-MLP and 0.08 higher than that of Pose Transformer, indicating that spatio-temporal relation coding has a stronger role in neighbor category discrimination. Ablation experiments further show that after removing the temporal relationship encoding, Macro-F1 decreases from 0.934 to 0.907, and the cluster purity decreases from 0.917 to 0.884, with the most significant performance degradation. After removing the spatial collaborative propagation, the cluster purity was reduced to 0.891. After removing stage gating, Macro-F1 decreases to 0.919. This indicates that the advantage of the proposed model does not come from a single module, but

from the synergy of temporal relation encoding, spatial co-propagation, and phase gating. At the same time, the relationship matrix construction and cross-stage aggregation also bring some computational overhead, but this cost is acceptable under the premise of ensuring high recognition accuracy and embedding stability. From the overall results, the proposed method maintains a more balanced performance among accuracy, discrimination and stability, which also indicates that the gait style modeling of models needs to establish a multi-stage and multi-region relationship representation.

6 Conclusions

In this paper, a gait style modeling method based on posture feature learning and spatio-temporal relationship encoding is proposed to solve the problem that fine-grained style boundaries are difficult to express stably and local actions and global rhythms are difficult to organize uniformly in gait style modeling. The method takes the keypoint sequence as input and constructs a style representation that can simultaneously depict local actions, regional collaboration and stage order through pose normalization, segment segmentation, multi-layer feature learning and relationship propagation. Experimental analysis shows that the constructed model performs well in style recognition stability, neighbor category discrimination ability and embedding spatial organization effect, and the spatio-temporal relationship coding plays an obvious role in maintaining the clarity of style boundaries. On the whole, the proposed method can well organize the local motion, regional linkage and global rhythm in the gait of the model, which provides a computable technical path for style recognition, style retrieval and digital motion modeling. However, the current methods still have room for further improvement under the conditions of complex clothing occlusion, cross-scene illumination changes and ultra-long sequence input. Relationship matrix construction and cross-stage aggregation also bring certain computational overhead. Future research can continue to focus on lightweight relationship coding, adaptive segment division, cross-scene generalization training, and multi-modal style joint modeling, so as to enhance the adaptation ability and deployment efficiency of the model in real runway scenes.

References

[1] Liu X, You Z, He Y, et al. Symmetry-driven hyper feature GCN for skeleton-based gait recognition[J]. Pattern Recognition, 2022, 125: 108520.

[2] Li H, Qiu Y, Zhao H, et al. GaitSlice: A gait recognition model based on spatio-temporal slice features[J]. Pattern recognition, 2022, 124: 108453.

[3] Li N, Zhao X. A strong and robust skeleton-based gait recognition method with gait periodicity priors[J]. IEEE Transactions on Multimedia, 2022, 25: 3046-3058.

[4] Rashmi M, Guddeti R M R. Human identification system using 3D skeleton-based gait features and LSTM model[J]. Journal of Visual Communication and Image Representation, 2022, 82: 103416.

[5] Zhang C, Chen X P, Han G Q, et al. Spatial transformer network on skeleton-based gait recognition[J]. Expert Systems, 2023, 40(6): e13244.

[6] Jun K, Lee K, Lee S, et al. Hybrid deep neural network framework combining skeleton

and gait features for pathological gait recognition[J]. Bioengineering, 2023, 10(10): 1133.

[7] Xu C, Makihara Y, Li X, et al. Occlusion-aware human mesh model-based gait recognition[J]. IEEE transactions on information forensics and security, 2023, 18: 1309-1321.

[8] Yan J, Wang S, Lin J, et al. Gaitsg: Gait recognition with smpls in graph structure[J]. Sensors, 2023, 23(20): 8627.

[9] Han K, Li X. Research method of discontinuous-gait image recognition based on human skeleton keypoint extraction[J]. Sensors, 2023, 23(16): 7274.

[10] Yousef R N, Khalil A T, Samra A S, et al. Model-based and model-free deep features fusion for high performed human gait recognition[J]. The Journal of Supercomputing, 2023, 79(12): 12815-12852.

[11] Chen J, Wang Z, Zheng C, et al. GaitAMR: Cross-view gait recognition via aggregated multi-feature representation[J]. Information Sciences, 2023, 636: 118920.

[12] Pan H, Chen Y, Xu T, et al. Toward complete-view and high-level pose-based gait recognition[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 2104-2118.

[13] Wang R, Shi Y, Ling H, et al. Gait recognition with multi-level skeleton-guided refinement[J]. IEEE Transactions on Multimedia, 2023, 26: 4515-4526.

[14] Deng M, Fan Z, Lin P, et al. Human gait recognition based on frontal-view sequences using gait dynamics and deep learning[J]. IEEE Transactions on Multimedia, 2023, 26: 117-126.

[15] Li Z, Li S, Xiao D, et al. Gait recognition based on multi-feature representation and temporal modeling of periodic parts[J]. Complex & Intelligent Systems, 2024, 10(2): 2673-2688.

[16] Wei S, Liu W, Wei F, et al. Gaitdlf: global and local fusion for skeleton-based gait recognition in the wild: S. Wei et al[J]. The Journal of Supercomputing, 2024, 80(12): 17606-17632.

[17] Zhang Z, Wei S, Xi L, et al. GaitMGL: Multi-scale temporal dimension and global–local feature fusion for gait recognition[J]. Electronics, 2024, 13(2): 257.

[18] Li R, Li H, Qiu Y, et al. Gaitae: a cognitive model-based autoencoding technique for gait recognition[J]. Mathematics, 2024, 12(17): 2780.

[19] Hu L, Zhang Z, Ye Y, et al. Diffusion-based Human Motion Style Transfer with Semantic Guidance[C]//Computer Graphics Forum. 2024, 43(8): e15169.

[20] Zhao Q, Li P, Yifan W, et al. Pose-to-Motion: Cross-Domain Motion Retargeting with Pose Prior[C]//Computer Graphics Forum. 2024, 43(8): e15170.