Digital Intelligence PRESS

Journal of Discovery Core

https://doi.org/10.67541/jdc2605

Research Article **Volume 1, Issue 1 (July 2026)** *Open Access*

Research on Rehabilitation Training Movement Recognition and Real-time Feedback Model Based on Computer Vision

Mingxiang Yang *

College of Arts, Shandong Jianzhu University, Jinan, Shandong, China

***Corresponding author:** Mingxiang Yang. Email: yangmingxiang10@outlook.com

Abstract

Aiming at the key problems such as the separation of action recognition and quality assessment tasks, coarse feedback granularity and high labeling cost in the automatic evaluation of rehabilitation training, this paper proposes PDDS-Net. The framework takes human skeleton sequence as input and realizes action classification and location-level deviation location synchronously through the collaborative architecture of a global action recognition stream and a local part evaluation stream. In the local flow, the joint nodes are divided into three functional part groups: upper limb, trunk and lower limb. independent graph convolutional subnetworks are used for decoupled coding, and a contrastive learning strategy is used to drive the attention map to focus on abnormal joints without frame-level annotations. This is mainly achieved through the adaptive fusion mechanism to dynamically integrate the confidence of the deep network and the matching score of dynamic time warping template to maintain decision stability under conditions of pose degradation. Experiments on the NTU RGB+D and PKU-MMD public datasets show that the action recognition accuracy of PDDS-Net reaches 93.7%, the average precision of position-level feedback is 0.656, 82.3% of the original AUC is still maintained at a 40% joint dropout rate, and the single-frame inference latency is 41.8 ms, which meets the requirements of real-time interaction. This study provides a replicable technical path for the validation of rehabilitation evaluation algorithms without clinical data collection.

Keywords

Computer vision; Rehabilitation training; Motion quality assessment; Graph convolutional network; Contrastive learning

Received: 20 Jun 2026; Revised: 18 Jul 2026; Accepted: 06 Aug 2026; Published: 18 Aug 2026

1. INTRODUCTION

1.1 Research background and motivation

Under the dual pressure of the acceleration of the global population aging process and the continuous rise of the incidence of chronic diseases, the contradiction between the supply of and demand for rehabilitation medical resources is becoming increasingly acute. According to the statistics of the World Health Organization, about two-thirds of people in low- and middle-income countries cannot obtain any professional services when facing rehabilitation needs [1],[2],[3]. Even in developed countries with well-established medical systems, the training cycle for rehabilitation therapists is long, and the labor cost is high, resulting in the supply capacity of rehabilitation services in institutions seriously lagging behind the growth of social needs. In this context, remote home-based rehabilitation has attracted widespread attention as a complementary solution. Its core concept is to transfer part of

rehabilitation training from hospital treatment rooms to patients' home environments and realize remote monitoring and guidance of the training process through technical means [4],[5]. However, the quality control of traditional rehabilitation training is highly dependent on on-site observation and subjective judgment of therapists. therapists visually assess the degree of standardization of patients' movements based on clinical experience and regularly conduct scale assessment to track the progress of rehabilitation. This model of relying on subjective observation and periodic evaluation has three inherent defects: first, the evaluation criteria vary from person to person, and different therapists may differ in their judgment of the same action; Second, real-time feedback cannot be provided, and patients cannot be informed of movement deviations in time and self-correct during training. Third, the periodic evaluation interval is long, and abnormal fluctuations in training quality are difficult to detect in a timely manner. The rapid development of computer vision technology has opened up a new technical path for non-contact and low-cost automated rehabilitation assessment [6],[7]. Compared with the method based on inertial sensor or depth camera, the human pose information obtained by a common RGB camera can realize motion capture without wearing devices, which greatly lowers the barrier to technology application and the burden of patients. The positioning of this study is clearly limited to the use of open academic data sets and standard action template library to carry out model design and verification, and the whole process does not involve the collection of real patient clinical data and the processing of private information, which avoids the restriction of medical ethical approval process on the research progress from the source, so that the study focuses on the innovation and verification at the algorithm level.

1.2 Research status at home and abroad

The technological evolution in the field of rehabilitation training motion evaluation shows a clear vein from hardware-dependent to purely vision-based approaches, from artificial features to deep features, and from template matching to end-to-end learning. In terms of pose estimation, early studies relied on marker-based optical motion capture systems (such as Vicon) to achieve 3D coordinate reconstruction with millimeter-level accuracy through reflective markers placed on key parts of the human body. Although such methods have excellent accuracy, they are expensive and cumbersome to operate and are only suitable for laboratory environments. Over the past five years, label-free RGB-based pose estimation methods have made great progress. MediaPipe, a lightweight solution launched by Google, can run in real time on mobile devices, while OpenPose is known for its multi-person detection capability [8],[9],[10]. Both achieve efficient mapping from images to 2D or 3D joint coordinates. These provide a standardized input representation for subsequent action semantic understanding. At the level of action quality assessment methods, the research has experienced a paradigm shift from DTW template matching to deep sequence modeling [11],[12]. DTW calculates the minimum cumulative distance between the sequence to be evaluated and the standard template through dynamic programming and uses the similarity score as the evaluation basis for the motion quality. This method does not need training and has good interpretability, but it has limited processing ability for the variation of motion speed and individual body size difference. Subsequently, deep sequence models such as LSTM and CNN have been introduced into the quality evaluation task, and the evaluation accuracy was significantly improved by using cyclic network to capture temporal dependence or using convolutional network to extract spatial features [13],[14],[15]. In recent years, Transformer architecture has shown performance potential beyond RNN architecture in action recognition and quality assessment by virtue of their natural advantages in modeling long-range dependencies.

At the level of rehabilitation feedback system, existing research results have preliminarily verified

the feasibility of computer vision assisted rehabilitation, but the system functions mostly stay at the level of correct/incorrect binary classification of actions, and the output feedback form is usually a general conclusion such as "correct action" or "wrong action". This coarse-grained feedback has limited guiding value in actual rehabilitation training. After learning the movement error, patients still do not know which joint has the problem, how much the deviation is, and how to adjust it. A few studies have attempted to generate quantitative feedback by calculating the difference between the joint angle and the standard value, but these methods often rely on preset fixed thresholds and lack adaptability to different motion types and different patients. The exploration of weakly supervised learning in the field of action evaluation provides a new approach to solve the annotation bottleneck. Traditional methods require frame-level annotation to train fine-grained quality assessment models, while weak supervision methods use video-level annotation to train frame-level classifiers, and automatically discover key frames and key parts through multi-instance learning or attention mechanism, which significantly reduces the labor cost of data annotation. Progress in this direction lays the methodological foundation for the attention contrastive learning strategy proposed later in this section.

1.3 Research content and innovation points

In view of the above limitations, this paper proposes a complete framework for rehabilitation training action recognition and real-time feedback, and the core innovation is reflected in the following four aspects. the first innovation is the proposal of a Part-Decoupled Dual-Stream Network (PDDS-Net), which decouples the rehabilitation action evaluation task into two parallel and cooperative processing streams. The global action recognition flow is responsible for the overall category discrimination and accuracy classification of the complete action sequence, while the local part evaluation flow is responsible for the independent quality analysis of each functional part group (upper limb group, trunk group, lower limb group). The fundamental advantage of this decoupling design ensures that the global stream is not disturbed by local detection noise and can stably capture the overall semantic features of the action. Local flow is not constrained by global category information and can finely describe the degree of execution specification of each part. The two streams share the underlying attitude coding in the feature extraction stage and realize the cooperation through the adaptive fusion mechanism in the decision stage, so as to achieve the dual goals of coarse-grained recognition and fine-grained evaluation. The second innovation is to design a location-level feedback generation module based on attention map contrast learning. The core technical path of this module is to use the attention weight map generated by the graph convolution layer in the local region evaluation flow to locate the nodes and parts that contribute the most to the quality score and guide the attention weight to focus on the abnormal joints that really cause execution deviation through contrast learning strategy. The key advantage of this method is that the training process does not require manual frame-level annotation at all, and the supervision signal comes from the feature difference between the standard action template and the sample to be evaluated, which fundamentally avoids the high cost of frame-level annotation. The third innovation is the construction of a multi-source feature adaptive fusion decision mechanism, and the classification confidence of deep learning network and the similarity score of DTW template matching are dynamically fused by weight through gated network. The core idea of the mechanism is to adaptively adjust the contribution proportion of each signal source according to the quality characteristics of the input sequence: when the pose estimation is complete and the action is clear, the discrimination results of the deep learning branch are more relied on, and the weight of the template matching branch is automatically increased when the pose is degraded or occluded, thus significantly enhancing the decision robustness of the model in complex environments and cross-subject conditions. The fourth innovation lies in the methodological contribution at the research design level. All experiments are

carried out based on open benchmark data sets and standard action template databases, and the whole process does not involve the collection of clinical data of real patients and ethical approval matters, which provides a replicable path for the verification of ethical approval free technology for similar studies. Through systematic experiments on NTU RGB+D, PKU-MMD and other public datasets, this paper verifies the superiority of the proposed model in terms of action recognition accuracy, location-level feedback positioning accuracy, reasoning speed and generalization ability across datasets.

2. DUAL-STREAM TEMPORAL FUSION NETWORK BASED ON PART DECOUPLING (PDDS-NET)

2.1 Overall architecture of the model

The Part-Decoupled Dual-Stream Network (PDDS-Net) proposed in this section aims to build an end-to-end rehabilitation training action evaluation framework, and its overall processing process can be summarized as follows: Open video data input → pose extraction → part coding → dual-stream processing → fusion decision → feedback generation. The core design concept of the architecture is to decouple the action recognition task into two parallel and cooperative processing streams, a global stream and a part-level stream. The global stream is responsible for the overall classification of the complete action sequence, the identification of the action type and the correctness of execution. The local stream conducts independent quality analysis for each functional part group to locate the specific part of execution deviation. The two streams share the output of the underlying attitude encoder in the feature extraction stage and integrate their respective discriminant information through the adaptive fusion mechanism in the decision stage, so as to realize the cooperative optimization of coarse-grained recognition and fine-grained evaluation. The fundamental advantage of this dual-flow decoupling design is that the global flow is not disturbed by local noise and can stably capture the overall semantics of the action. Local flow is not constrained by global categories and can finely describe the execution quality of each part. The two complement each other and jointly support the subsequent position-level feedback generation.

All training and testing of this model are based on public academic data sets and standard action template libraries, and do not involve clinical data collection of real patients at all. Specifically, NTU RGB+D data set and PKU-MMD data set related to rehabilitation actions (such as upper limb extension, lower limb lifting and other action categories) were used in the training stage, supplemented by standard action templates extracted from professional rehabilitation guidance videos to construct positive sample sets [16],[17]. The construction process of the standard action template library is as follows: firstly, the standard joint angle range and motion trajectory of each action are calibrated by rehabilitation medicine experts, and then the corresponding posture sequence is extracted from the expert demonstration video as the template. All data are stored in the form of skeleton sequence, and each frame contains 2D/3D coordinates and confidence scores of several key nodes (25 key points). This unified pose representation format provides standardized input for subsequent spatio-temporal map construction and feature coding.

2.2 Attitude representation and spatio-temporal feature coding

The pre-module of pose estimation uses lightweight MediaPipe or HRNet network to extract the coordinates of human joint nodes and corresponding confidence scores from the input video frames. Let the set of human joint nodes extracted in frame t be:

$$P_t = \{ p_{t,i}, \in, \mathbb{R}^D, \mid i = 12 \ldots N \} \quad (1)$$

Where $N = 25$ represents the total number of human joint nodes, and $D = 2$ or 3 corresponds to two-dimensional or three-dimensional space coordinates respectively. Each node simultaneously outputs the corresponding confidence:

$$c_{t,i} \in [0,1] \quad (2)$$

In order to eliminate the influence of different subjects' body size differences on the subsequent feature learning process, the torso length of the original joint node coordinates is normalized:

$$\hat{p}_{t,i} = \frac{p_{t,i}}{l_t} \quad (3)$$

Where, the trunk scale l_t is defined as the Euclidean distance between the first joint node (bottom of the spine) and the eighth joint node (middle of the spine):

$$l_t = \| p_{t,1} - p_{t,8} \|_2 \quad (4)$$

After obtaining the normalized skeleton sequence, the spatio-temporal graph structure is further constructed:

$$G = (V, E) \quad (5)$$

It is used to simultaneously model the spatial topology relationship between human joint nodes and the cross-frame motion change rule. Where the node set is defined as:

$$V = \{ v_{t,i}, \mid, t, =, 1, \ldots Ti = 1 \ldots N \} \quad (6)$$

Represents the joint nodes of all time steps in the whole video sequence. The initial feature vector corresponding to each node is expressed as:

$$f_{t,i}^{(0)} = [\hat{p}_{t,i}; c_{t,i}] \in \mathbb{R}^{D+1} \quad (7)$$

The edges in the graph structure are composed of two parts: spatial connected edges and temporal connected edges. Spatial edge is used to describe the topological relationship of human skeleton in the same time frame:

$$E_S = \{ (v_{t,i}, v_{t,j}) \mid (i,j) \in H \} \quad (8)$$

Where, H represents the set of natural bone connections of the human body. The time edge is used to connect the same node in adjacent video frames:

$$E_T = \{ (v_{t,i}, v_{t+1,i}), \mid, t = 1 \ldots T - 1 \} \quad (9)$$

The final edge set of space-time graph is expressed as:

$$E = E_S \cup E_T \quad (10)$$

The spatio-temporal map construction method does not need additional manual division of human body parts and regions, and all connections are automatically generated according to the natural topology of human body, which can effectively retain the spatial structure information and temporal dynamic characteristics in the process of human movement, and has good cross-individual generalization ability [18],[19].

The feature coding layer uses space-time graph convolutional network (ST-GCN) to perform high-dimensional feature mapping of the original coordinates. Spatial graph convolution is defined as follows: For node $v_{t,i}$, its convolution output is:

$$f_{out}(v_{t,i}) = \sum_{v_{t,j} \in \mathcal{B}(v_{t,i})} \frac{1}{Z_{t,i}(v_{t,j})} f_{in}(v_{t,j}) \cdot w\left(l_{t,i}(v_{t,j})\right) \tag{11}$$

Where $\mathcal{B}(v_{t,i})$ represents the neighborhood set of nodes $v_{t,i}$ (including itself and its spatial neighbors), $Z_{t,i}(v_{t,j})$ is the cardinality normalization term of the corresponding subset, $l_{t,i}(v_{t,j})$ maps neighborhood nodes to three subset labels (root node itself, centripetal node, centrifugal node), and $w(\cdot)$ is a learnable weight function. After the spatial graph convolution, the temporal convolution convolves the feature sequence of the same joint node along the temporal dimension to capture the dynamic change between frames. The whole ST-GCN encoder is stacked with 9 layers of spatio-temporal graph convolution modules, and each layer performs spatial feature extraction and temporal feature extraction alternately. In this encoding way, the original skeleton coordinates are mapped into a high-dimensional feature tensor $F_{global} \in \mathbb{R}^{T \times N \times C}$, where C is the number of feature channels. At the same time, according to the site grouping strategy, the feature subset $F_{part}^{(k)} \in \mathbb{R}^{T \times N_k \times C}$ of each site group is separated from F_{global} ($k = 1,2,3$ correspond to the upper limb group, trunk group, and lower limb group respectively) for subsequent local site evaluation flow independent processing.

2.3 Global Action recognition stream: temporal attention aggregation module

The design goal of global action recognition flow is to classify the complete action sequence as a whole, identify the action type and the execution correctness. The stream takes the global feature tensor F_{global} output by ST-GCN encoder as input and captures the long-range timing dependence through the timing Transformer encoder. Compared with time series modeling methods based on LSTM or pure CNN, the self-attention mechanism of Transformer architecture can directly model the dependence relationship between any frame pair without the limitation of time series distance, which has stronger adaptability to irregular patterns such as speed variation and pause that may occur in rehabilitation actions [20],[21].

The core calculation process of timing Transformer encoder is as follows. First, the input feature tensor is flattened along the time dimension into a sequence form $X = \{x_t, \in, \mathbb{R}^{N \cdot C} \mid t = 1 \ldots T\}$, and add a learnable position code $E_{pos} \in \mathbb{R}^{T \times (N \cdot C)}$ to retain the timing order information: $\tilde{X} = X + E_{pos}$. The encoder is composed of L-layer Multi-Head Self-Attention (MHSA) layer and Feed-Forward Network (FFN) stacked alternately. The calculation process of layer l is as follows:

$$Z^{(l)} = \text{LayerNorm}\left(Z^{(l-1)} + \text{MHSA}(Z^{(l-1)})\right) \tag{12}$$

$$Z^{(l+1)} = \text{LayerNorm}\left(Z^{(l)} + \text{FFN}(Z^{(l)})\right) \tag{13}$$

Where long self-attention calculation:

$$\text{MHSA}\ (Q, K, V) = \text{Concat}(\text{FFN}) \tag{14}$$

$$\text{head}_h = \text{Softmax}\left(\frac{Q_h K_h^{\top}}{\sqrt{d_k}}\right) V_h \tag{15}$$

Here $Q = Z^{(l-1)}W^Q$, $K = Z^{(l-1)}W^K$, $V = Z^{(l-1)}W^V$ are the query, key and value matrices respectively, d_k is the dimension of the key vector, and H is the number of attention heads.

After L-layer coding, the output feature $z_{cls} \in \mathbb{R}^d$ corresponding to the [CLS] mark at the first place of the sequence is taken as the global representation of the whole action sequence. In order to enhance the classification and discrimination ability, a set of learnable action category prototype vectors $P = \{p_m, \in, \mathbb{R}^d \mid m = 1 \ldots M\}$, where M is the number of action categories. The similarity between sequence features and prototype vectors is calculated through the cross-attention mechanism:

$$\alpha_m = \frac{\exp\left(z_{cls}^{\top} p_m/\tau\right)}{\sum_{j=1}^{M} \exp\left(z_{cls}^{\top} p_j/\tau\right)} \tag{16}$$

Where τ is the temperature hyperparameter. The final global action classification probability is:

$$p_{cls} = \max_m \alpha_m \in [0,1] \tag{17}$$

The probability value reflects both the recognition result of the action type and the degree of confidence of the model in the recognition. If $p_{cls} < \theta$ (θ is the preset confidence threshold), the model will trigger the rejection mechanism, indicating that the input action does not belong to the known category or the quality is too low to be re-executed.

2.4 Local site evaluation flow: fine-grained quality analysis of site decoupling

The design goal of the local site evaluation flow is to locate the specific site of execution deviation and provide the basis for subsequent fine feedback. Unlike the global flow, this flow does not classify actions as a whole but evaluates the degree of specification independently for each functional part group. The site grouping strategy was designed based on the biomechanical principle of human motion, and the 25 joint nodes were divided into three functional site groups: upper limb group (including shoulder, elbow, wrist and related bone points), trunk group (including various spinal segments and neck), and lower limb group (including hip, knee, ankle and related bone points). This grouping method follows the clinical practice logic of "independent analysis of parts" in rehabilitation evaluation. The deviation of movement execution in different parts often comes from different pathological mechanisms, which requires independent diagnosis.

Part independent encoder is the core component of local flow. For the k-th part group, the corresponding part feature subset $F_{part}^{(k)} \in \mathbb{R}^{T \times N_k \times C}$ is extracted from the global feature tensor, where N_k is the number of joint nodes of the part group. Each part group uses an independent graph convolutional subnetwork $\mathrm{GCN}_k(\cdot)$ for feature extraction:

$$H^{(k)} = \mathrm{GCN}_k\left(F_{part}^{(k)}\right) \tag{18}$$

Where $H^{(k)} \in \mathbb{R}^{T \times N_k \times C'}$ is the deep feature of position group k, and C' is the number of output channels. The key advantage of part independent coding is that the feature extraction process of different part groups is completely decoupling, and each subnetwork can learn the motion mode of its part independently, avoiding feature confusion between parts, for example, the large swing of the upper limb should not be mistakenly interpreted as the instability of the torso.

The purpose of the location-level comparison learning module is to enhance the ability of location-level characteristics to discriminate execution deviation. The corresponding position feature in the

standard action template is used as the positive sample anchor, and the same position feature in the deviation execution sample is used as the negative sample to design the position level comparison loss function. Specifically, let the batch contain B samples, and each sample contains the characteristics of three position groups. For position group k, the InfoNCE contrast loss is:

$$\mathcal{L}_{contrast}^{(k)} = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp\left(\mathrm{sim}(h_{b,k}^{(pos)}, h_{b,k}^{(anchor)})/\tau_k\right.}{\sum_{b'=1}^{B}\exp\left(\frac{\mathrm{sim}\left(h_{b,k}^{(pos)}, h_{b',k}^{(neg)}\right)}{\tau_k}\right)} \tag{19}$$

Where $h_{b,k}^{(anchor)}$ is the feature of position group k in the standard template, $h_{b,k}^{(pos)}$ is the feature of position group k in the same sample (the distance between it and the anchor point is narrowed in training), $h_{b',k}^{(neg)}$ is the feature of site group k of other samples in the batch (the distance from the anchor point in training), $\mathrm{sim}(u,v) = u^{\top}v/(\| u \|\| v \|)$ is the cosine similarity function, τ_k is the temperature parameter of site group k. With this contrastive learning strategy, the model is trained to produce sensitive eigenresponses to the execution bias of each position group. The total contrast loss is the weighted sum of the losses of the three position groups:

$$\mathcal{L}_{contrast} = \sum_{k=1}^{3}\lambda_k\,\mathcal{L}_{contrast}^{(k)} \tag{20}$$

Where λ_k is the weight coefficient of each position group.

The site quality score head maps the deep features of each site group into a scalar quality score. For site group k, the quality score $s_k \in [0,1]$ is calculated as follows:

$$s_k = \sigma\left(\mathrm{MLP}_k\left(\mathrm{Pool}\left(H^{(k)}\right)\right)\right) \tag{21}$$

Where $\mathrm{Pool}(\cdot)$ is the average pooling operation along time and node dimensions, $\mathrm{MLP}_k(\cdot)$ is the multi-layer perception dedicated to position group k, and $\sigma(\cdot)$ is the Sigmoid activation function. The closer the quality score s_k is to 1, the more standardized the execution of the position is, and the closer it is to 0, the more significant the deviation is. In contrast to the traditional approach where all parts are scored uniformly, this module outputs an independent quality score for each part group, thus providing a quantitative basis for subsequent part-level feedback. This fine-grained quality analysis strategy of part decoupling is an innovation that has not yet been realized by existing action evaluation methods. Its core breakthrough is to refine the evaluation granularity from global action to part level, and the training process only needs video-level annotation, without frame-level annotation.

2.5 Adaptive fusion decision of multi-source features

The multi-source feature adaptive fusion decision module is responsible for integrating the multi-dimensional information provided by global action recognition flow, local part evaluation flow and traditional template matching method, and output the final decision result [22],[23]. The design motivation of this module is as follows: different signal sources have different reliability and complementarity under different conditions, global classification has excellent performance when action semantics is clear, part evaluation has advantages in fine-grained quality discrimination, and template matching is not constrained by the training distribution of deep learning model and can still

provide benchmark reference when pose estimation degradation.

The template alignment branch uses Dynamic Time Warping (DTW) algorithm to calculate the similarity between the sequence to be evaluated and the standard action template. Let the joint angle trajectory of the sequence be evaluated be $A = [a_1, \dots, a_T]$, and the standard template sequence be $B = [b_1, \dots, b_{T_{ref}}]$, where each a_t and b_t are the key joint angle vectors. DTW algorithm finds the optimal time sequence alignment path through dynamic programming to minimize the cumulative distance:

$$\mathrm{DTW}(A, B) = \min_{\pi} \sum_{(t,t')\in\pi} d\,(a_t, b_{t'}) \tag{22}$$

Where π is the alignment path and $d(\cdot,\cdot)$ is the Euclidean distance measure. After the DTW distance is obtained, it is normalized to the similarity score:

$$s_{DTW} = \exp\big(-\gamma \cdot \mathrm{DTW}(A, B)\big) \tag{23}$$

Where $\gamma > 0$ is the scale parameter. The score $s_{DTW} \in [0,1]$, the closer it is to 1, the more similar the action to be evaluated is to the template. The advantage of DTW is that it can naturally deal with the variation of action execution speed without the need for length normalization of the sequence.

The adaptive fusion mechanism takes the gated fusion network as the core and dynamically adjusts the weight from the three signal sources. Let the input fusion feature be:

$$f_{fuse} = [p_{cls}, s_1, s_2, s_3, s_{DTW}] \in \mathbb{R}^5 \tag{24}$$

Where p_{cls} is the global action classification probability, s_k is the quality score of each part, and s_{DTW} is the template matching score. The gated converged network adopts a two-layer fully connected architecture. Firstly, the dynamic weight vector is generated:

$$w_{gate} = \mathrm{Softmax}\big(W_2 \cdot \mathrm{ReLU}\big(W_1 \cdot f_{fuse} + b_1\big) + b_2\big) \tag{25}$$

Where $W_1 \in \mathbb{R}^{16\times5}$ and $W_2 \in \mathbb{R}^{3\times16}$ are weight matrices, and b_1, b_2 are bias vectors. The Softmax function guarantees that the sum of the three weights is one. Then calculate the weighted fusion score:

$$y = w_1 \cdot p_{cls} + w_2 \cdot \bar{s} + w_3 \cdot s_{DTW} \tag{26}$$

Where $\bar{s} = \frac{1}{3}\sum_{k=1}^{3} s_k$ is the mean value of the site score, and w_1, w_2, w_3 is the dynamic weight generated by the gated network. The final decision score $y \in [0,1]$ is binarized to the two categories of "correct execution/deviation" through the threshold η.

3. POSITION-LEVEL FEEDBACK GENERATION AND INTERPRETABILITY MODELING

3.1 Problem definition of feedback generation

The core functional requirements of real-time feedback system for rehabilitation training go beyond simple binary classification. The system not only needs to inform "action error" but also must accurately point out "which part is wrong" and "where is wrong". This refined feedback demand is rooted in the clinical practice logic of rehabilitation medicine: when guiding patients, rehabilitation

therapists will specifically point out "insufficient lifting angle of right shoulder joint" or "compensatory rotation of trunk", rather than give a general evaluation of "wrong action". In order to achieve this goal, the feedback generation module designed in this section is required to achieve position-level positioning accuracy in terms of feedback granularity and cover three common execution error modes of angle deviation, speed anomaly and trajectory deviation in terms of deviation type dimension. Angle deviation finger joint range of motion does not reach or exceed the standard range, abnormal speed refers to the movement is too fast or too slow to deviate from the rhythm of rehabilitation training requirements, trajectory deviation finger joint movement path and standard trajectory space deviation. It is worth emphasizing that all feedback generation processes in this section are completely based on the comparison between the internal attention response of the model and the standard template library, and do not rely on any frame-level subjective annotation by human experts. This design choice has dual significance: on the one hand, it eliminates the bottleneck problem of high cost of manual annotation and difficulty in ensuring the consistency of annotation; on the other hand, the whole study completely avoids the need for the collection of clinical data of real patients and fundamentally eliminates the process constraints of ethical approval. The feedback generation module is based on the PDDS-Net network output proposed in section 2, and the specific input includes the attention weight graph generated by the convolution layer of each graph in the local part evaluation flow, the quality score vector of each part group, and the multi-dimensional comparison results between the sequence to be evaluated and the standard template library.

3.2 Bias localization based on attention graph contrast learning

The core idea of deviation location module is to use the attention weight graph of graph convolution layer in local part evaluation flow to locate the node and part that contributes the most to quality score. In traditional methods, attention mechanism is often only an implicit operation process inside the model, and its output weight is not explicitly used in interpretability analysis. In this section, the attention weight graph is proposed as the direct basis for deviation of location. The internal logic is as follows: the attention weight in the convolution layer of the graph reflects the contribution degree of each spatial neighborhood node to the feature update of the central node. When the model determines that there is deviation in the execution of a certain part, the attention weight distribution of this part must be concentrated on the abnormal node leading to the deviation. Specifically, for the graph convolutional subnetwork GCN_k of position group k defined in section 2, the multi-head attention weight of the last layer of spatial graph convolution is extracted. Let the weight matrix of the h-th attention head in position group k be:

$$A^{(k,h)} \in \mathbb{R}^{T\times N_k\times N_k} \tag{27}$$

Where each element $a_{t,i,j}^{(k,h)}$ represents the attention contribution value of node j to node i in the t-th frame. In order to obtain the global importance score of each node in the position group, the multi-head dimension and the target node dimension are firstly aggregated:

$$g_{t,i}^{(k)} = \frac{1}{H}\sum_{h=1}^{H}\sum_{j=1}^{N_k} a_{t,i,j}^{(k,h)} \tag{28}$$

Where $g_{t,i}^{(k)} \in \mathbb{R}$ represents the attention aggregation weight of the i-th joint node in position

group k in the t-th frame, and H is the number of attention heads. This score intuitively reflects the contribution degree of the joint node to the quality evaluation of the part in the current frame. The greater the contribution is, the more critical the motion mode of the joint node is to the execution deviation of the model in identifying the part.

However, the original attention weight $g_{t,i}^{(k)}$ may be disturbed by noise input or training data bias, resulting in some key nodes that have nothing to do with execution bias also obtain high attention scores. In order to solve this problem, this section designs an attention optimization strategy based on contrast learning. The core idea is to train attention to focus on the key nodes that really cause execution deviation, rather than background noise or individual habitual posture differences. Let the spatio-temporal feature of the standard action template in position group k of frame t be $h_t^{(k,anchor)} \in \mathbb{R}^{C'}$, and the corresponding feature of the sample to be evaluated be $h_t^{(k,query)} \in \mathbb{R}^{C'}$, and define the feature difference graph of the two:

$$D_t^{(k)} = \left| h_t^{(k,query)} - h_t^{(k,anchor)} \right| \in \mathbb{R}^{C'} \tag{29}$$

The attention weight $g_{t,i}^{(k)}$ is regarded as the spatial weighted aggregation coefficient of the feature difference map, and the attention optimization loss is defined as:

$$\mathcal{L}_{attn}^{(k)} = -\frac{1}{TN_k} \sum_{t=1}^{T} \sum_{i=1}^{N_k} g_{t,i}^{(k)} \cdot \log \left(\frac{\exp\left(\| \frac{d_{t,i}^{(k)}}{\tau_a} \|_2 \right)}{\sum_{j=1}^{N_k} \exp\left(\| \frac{d_{t,j}^{(k)}}{\tau_a} \|_2 \right)} \right) \tag{30}$$

Where $d_{t,i}^{(k)}$ is the feature subvector corresponding to the i-th node in the feature difference graph $D_t^{(k)}$, and τ_a is the temperature hyperparameter. The design logic of the loss function is to encourage the distribution of attention weight $g_{t,i}^{(k)}$ to the nodes with large feature differences (that is, the nodes with significant deviation) and inhibit the distribution of attention weight on the nodes with small feature differences. Through end-to-end joint training, attention is guided by contrast learning signals to the key nodes that are truly relevant to execution bias.

3.3 Template-driven corrective information generation

After obtaining the location-level deviation localization results, this section further designs a template-driven corrective information generation module to transform the abstract attention response output by the model into interpretable feedback with clinical guidance significance. Through multi-dimensional quantitative matching between the action sequence to be evaluated and the standard action template library, the module analyzes the action deviation from three levels of joint angle, motion trajectory and coordination relationship between parts, and generates targeted correction suggestions.

3.3.1 Quantification of joint angle deviation

Joint angle deviation is an important basis for motion quality evaluation and feedback generation. For key joints such as shoulder, elbow, hip, and knee, the joint angle in 3D space is calculated using the vector of adjacent bone segments [24]. Taking the shoulder joint flexion angle as an example, the upper

arm vector and the trunk reference vector are defined as:

$$v_{upper} = p_{shoulder} - p_{elbow} \tag{31}$$

$$v_{torso} = p_{hip} - p_{shoulder} \tag{32}$$

The corresponding joint angle is calculated as follows:

$$\theta_{joint} = \arccos\left(\frac{v_{upper} \cdot v_{torso}}{\| v_{upper} \| \| v_{torso} \|}\right) \tag{33}$$

After DTW time alignment, the joint angle of the t-th frame of the action sequence to be evaluated is θ_t^{query}, and the reference angle corresponding to the standard action template is θ_t^{ref}, then the angle deviation is defined as:

$$\delta_\theta(t) = \theta_t^{query} - \theta_t^{ref} \tag{34}$$

For the complete action sequence, the average angle deviation is calculated as follows:

$$\bar{\delta}_\theta = \frac{1}{T}\sum_{t=1}^{T} \delta_\theta\,(t) \tag{35}$$

At the same time, the moment of maximum deviation is determined:

$$t_{max} = \arg\max_t \mid \delta_\theta(t) \mid \tag{36}$$

When satisfied:

$$\mid \bar{\delta}_\theta \mid > \epsilon_\theta \tag{37}$$

$$\mid \delta_\theta(t_{max}) \mid > \epsilon_\theta^{max} \tag{38}$$

The system determines that there is an obvious angle deviation. Here, ϵ_θ and ϵ_θ^{max} denote the average deviation and the local maximum deviation threshold, respectively. When $\bar{\delta}_\theta > 0$, the feedback of "joint angle is too large" is generated. When $\bar{\delta}_\theta < 0$, the feedback of "small joint angle" is generated, and the adjustment suggestion is given in combination with the deviation amplitude.

3.3.2 Movement trajectory deviation detection

Trajectory deviation detection evaluates the quality of action execution from the perspective of spatial motion path. The 3D motion trajectory of key joints in the action to be evaluated is expressed as:

$$T_{query} = \left\{ p_t^{query}, \in, \mathbb{R}^3 \mid t = 1 \ldots T \right\} \tag{39}$$

The standard action template trajectory is expressed as:

$$T_{ref} = \left\{ p_t^{ref}, \in, \mathbb{R}^3 \mid t = 1 \ldots T_{ref} \right\} \tag{40}$$

After DTW time alignment, the spatial distance between corresponding frames is calculated:

$$d_t = \| p_t^{query} - p_t^{ref} \|_2 \tag{41}$$

In order to quantify the overall trajectory deviation degree, the trajectory deviation index is defined

as follows:

$$\Psi = \frac{1}{T}\sum_{t=1}^{T} I\left(d_t > \epsilon_d\right) \cdot d_t \tag{42}$$

Where, $I(\cdot)$ is the indicator function and ϵ_d is the spatial distance tolerance threshold. When:

$$\Psi > \epsilon_\Psi \tag{43}$$

The system determines that there is a significant deviation in the action trajectory and further locates the continuous-time region with the most serious deviation:

$$T_{dev} = \{t, d_t > \epsilon_d \; d_t \; is \; the \; local \; maximum\} \tag{44}$$

Based on the deviation position and direction information, the system generates feedback with temporal positioning, such as "the action trajectory deviates from the standard path between the X second and the Y second, please adjust in the Z direction".

3.3.3 Compensatory action recognition

Compensatory movement is an important clinical feature in rehabilitation exercise evaluation. When there is insufficient muscle strength, limited joint movement, or decreased control at the target movement site, the patient may complete the movement through abnormal movement at other non-target sites. For example, when shoulder joint movement is limited, the upper limb lifting action is completed by torso lateral flexion.

This section realizes the detection of compensatory behavior based on the analysis of movement coordination of different parts. Firstly, the motion activity of position group k in frame t is defined as follows:

$$m_t^{(k)} = \frac{1}{N_k}\sum_{i=1}^{N_k} \| p_{t,i} - p_{t-1,i} \|_2 \tag{45}$$

Where N_k represents the number of joint nodes contained in position group k.

For the case that the target action mainly involves site group a, the relative compensatory movement index of non-target site group b is defined as follows:

$$\kappa = \frac{\frac{1}{T}\sum_{t=1}^{T} m_t^{(b)}}{\frac{1}{T}\sum_{t=1}^{T} m_t^{(a)} + \epsilon} \tag{46}$$

Where ϵ is a tiny constant that prevents the denominator from being zero. When κ exceeds the preset threshold, the system determines that there is an abnormal compensation pattern and generates feedback in combination with the contribution of site movements, such as "abnormal increase in trunk participation is detected, and it is recommended to reduce trunk lateral flexion compensation and improve the active control ability of target joints".

3.4 Real-time optimization strategy of feedback

In order to ensure that the feedback system can provide smooth real-time interactive experience in practical application scenarios, this section implements real-time optimization strategies from two levels of model lightweight design and reasoning acceleration scheme. In terms of model lightweight, the PDDS-Net network structure proposed in section 2 is compressed and optimized. First of all, Group Convolution is used to replace standard convolution in the part independent encoder, and the input feature channel is divided into several groups. Each group conducts convolution independently, and the number of parameters is reduced to $1/G$ of standard convolution, where G is the number of groups is. Secondly, channel level structured pruning is implemented. The scaling factor γ of Batch Normalization layer in each convolutional layer is taken as the evaluation index of channel importance, and the channels with the lowest 30% rank of γ value are pruned out. After pruning, the model accuracy is recovered through fine tuning. Finally, the feature multiplexing strategy is adopted in the timing Transformer encoder, that is, part of the intermediate feature calculation results is shared between adjacent frames. By taking advantage of the gentle change between frames of rehabilitation action, the feature update frequency is reduced from calculation per frame to calculation per k frames (k =3), and the latest calculation results of frame multiplexing are not calculated.

In terms of inference acceleration scheme, NVIDIA TensorRT is used to optimize the deployment of the trained model. TensorRT significantly reduces inference delay through layer fusion (such as fusing convolutional layer and batch normalization layer into a single layer), tensor accuracy calibration (using FP16 mixed precision inference), and dynamic batch processing. In addition, the inter-frame feature caching mechanism is designed: for continuous video stream, the joint node coordinates output by the pose estimation module have high continuity between adjacent frames, and the model only triggers complete end-to-end forward inference when the inter-frame displacement exceeds the preset threshold Δ_{frame}. The other frames directly reuse the decision results of the previous frame and carry out lightweight interpolation update when necessary.

4. EXPERIMENTAL DESIGN AND RESULT ANALYSIS

4.1 Data set and experimental setup

The experiments in this section are carried out based on the publicly available benchmark data set for action recognition and quality assessment. The NTU RGB+D dataset contains 56880 video samples, covering 60 action categories, collected under three different perspectives [25],[26]. In this study, a subset of movements related to rehabilitation training was selected, including 12 categories (A41 to A47, A103 to A105 and other medical related movements), including upper limb extension, lower limb lifting, trunk flexion and extension, involving a total of about 12,000 samples. The PKU-MMD dataset contains about 28,000 action instances, providing RGB, depth, infrared and skeleton data, including the annotation of correct execution and deviation execution samples, which is suitable for the training and testing of action quality assessment tasks. The construction method of the standard action template library is as follows: the standard posture sequence is selected from the professional rehabilitation guidance video, and the reference value of the angle range and motion trajectory of each joint is calibrated by the rehabilitation medical experts, which is used as the benchmark reference of the template matching branch after the construction. All data are from public academic datasets or standard action templates, and do not involve clinical data collection or private information of real patients, so the whole process is exempted from ethical approval.

The evaluation index system comprehensively evaluates the performance of action recognition,

action quality evaluation ability, error feedback positioning accuracy and real-time performance of the system.

The performance of action recognition is measured by Accuracy and F1-Score. Where, the accuracy rate is used to evaluate the overall classification correct proportion of the model [27]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{47}$$

Where TP, TN, FP and FN represent the number of true cases, true negative cases, false positive cases and false negative cases respectively. F1-Score takes Precision and Recall into consideration and is defined as the harmonic average of the two:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{48}$$

The action quality was evaluated by AUC-ROC and FAR@TAR. Among them, AUC-ROC is used to measure the overall discrimination ability of the model under different classification thresholds. FAR@TAR is used to evaluate the proportion of false acceptances generated by the system under the specified True Acceptance Rate (TAR):

$$FARTAR = \frac{FP}{FP + TN} \tag{49}$$

The feedback positioning accuracy is evaluated by position-level Precision and Recall, which are used to measure the positioning ability of the model to abnormal motion parts:

$$Precision_{part} = \frac{TP_{part}}{TP_{part} + FP_{part}} \tag{50}$$

$$Recall_{part} = \frac{TP_{part}}{TP_{part} + FN_{part}} \tag{51}$$

Where TP_{part} represents the number of abnormal parts correctly positioned by the model, FP_{part} represents the number of incorrectly predicted parts, and FN_{part} represents the number of abnormal parts not detected.

The real-time performance of the system is evaluated by single frame inference delay and processing throughput. Among them, single-frame inference delay is used to measure the time (ms) required by the model to complete a prediction, and the throughput is expressed by Frames Per Second (FPS):

$$FPS = \frac{N_{frame}}{T_{infer}} \tag{52}$$

Where, N_{frame} represents the number of video frames processed, and T_{infer} represents the corresponding reasoning time.

In addition, in order to verify the cross-scene adaptability of the model, the cross-data set transfer experiment was used to evaluate the generalization performance of the model, that is, the model training was completed on the NTU RGB+D data set, and the independent test was conducted on the PKU-MMD subset to test the stability of the model under different acquisition environments, action category distribution and human movement mode changes.

4.2 Ablation experiment and architecture verification

The ablation experiment aims to quantify the independent contributions of each core component of the dual-flow architecture and verify the effectiveness of the design decision. Firstly, the necessity of dual-stream architecture is verified: three variant models are constructed, (a) complete PDDS-Net dual-stream model; (b) Remove the local evaluation flow and retain only the global flow (denoted as "single-flow-global"); (c) Remove the global action recognition flow and retain only the part flow (denoted as "single-flow-part"). The three variants were trained and tested on the same dataset, and the results are listed in Table 1.

Table 1. Comparative results of ablation experiments with dual-flow architecture

Variant of model	Action recognition accuracy	Position level feedback Precision	Position level feedback Recall
Single flow - Global (remove part flow)	92.1%	0.412	0.387
Single flow-locality (remove global flow)	68.4%	0.638	0.621
Complete PDDS-Net dual stream	93.7%	0.656	0.644

The action recognition accuracy of complete PDDS-Net (93.7%) is limited compared with that of single-flow-global (92.1%), but the position-level feedback Precision (0.656 vs. 0.412) and Recall (0.644 vs. 0.387) are significantly improved. It shows that the addition of local part evaluation flow has a decisive contribution to feedback refinement, while the gain for action classification tasks is relatively moderate. Although the feedback positioning accuracy of the single-flow-part variant is close to that of the full model (Precision 0.638 vs. 0.656), its action recognition ability is seriously degraded (68.4%), which confirms the irreplaceable role of global flow in action semantic discrimination, and the dual flow cooperation achieves the dual goals of recognition and evaluation.

The ablation experiment of fusion strategy compares three schemes: (a) simple splicing fusion: p_{cls}, s and s_{DTW} are directly spliced and then the decision is output by linear layer; (b) Only using deep confidence fusion, fusing p_{cls} and s without DTW matching score; (c) Adaptive gating fusion, the dynamic weight allocation scheme proposed in this section. The experiment was evaluated separately on the standard test set and the pose degradation test set (artificially adding 10% joint noise), and the results are shown in Table 2.

Table 2. Comparison of robustness of different fusion strategies

Convergence strategy	The standard test set AUC	Pose degradation test set AUC	Decrease in AUC
Simple splicing and fusion	0.912	0.804	-11.8%
Deep confidence only (no DTW)	0.908	0.782	-13.9%
Adaptive gating fusion	0.925	0.861	-6.9%

The adaptive gated fusion achieved the highest AUC (0.925) on the standard test set, and the AUC decreased the least (-6.9%) under the condition of attitude degradation, which verified the original design intention of the dynamic weight allocation mechanism to automatically improve the contribution

of DTW branch when attitude estimation degradation. The ablation experiment of site decoupling strategy compared the two schemes of site independent coding and global unified coding. In terms of site level evaluation accuracy, the average evaluation accuracy of each site group of independent coding scheme was 86.3%, which was significantly higher than 79.1% of global unified coding, indicating that feature decoupling between parts can effectively avoid the mutual interference of motion patterns of upper limbs, trunk and lower limbs. Improve the accuracy of position level discrimination.

4.3 Comparative experiment

The comparative experiment systematically compares PDDS-Net with the current mainstream methods. The selected baseline methods for Action recognition task include LSTM (Long short-term memory network), Bi-LSTM (bidirectional LSTM), ST-GCN (Spatio-temporal graph convolutional network), Action Transformer and PoseC3D. All methods use the same training set and test set division, and the experimental results are shown in Table 3.

Table 3. Performance comparison of different methods on action recognition task

Method	Accuracy rate	F1-Score	Number of parameters (M)
LSTM	81.2%	0.804	4.1
Bi-LSTM	84.7%	0.839	6.7
ST-GCN	89.3%	0.887	12.8
Action Transformer	91.5%	0.909	18.2
PoseC3D	92.8%	0.921	14.6
PDDS-Net（Model of this section）	93.7%	0.930	13.2

PDDS-Net outperforms all baseline methods in accuracy (93.7%) and F1-Score (0.930) and improves by 0.9 percentage points and 0.009 respectively compared with PoseC3D, which has the best performance. The number of parameters (13.2M) is slightly less than that of PoseC3D but greater than that of ST-GCN, which achieves a good balance between accuracy and efficiency. DTW template matching, CNN+DTW fusion and weakly supervised frame-level classifier were selected as the baseline in the comparison experiment of quality assessment task. The AUC of PDDS-Net reached 0.925, which was higher than 0.897 of CNN+DTW fusion and 0.883 of weakly supervised frame-level classifier.

The real-time performance comparison is carried out on the NVIDIA Jetson Xavier NX edge computing platform, and the inference delay is compared with the lightweight scheme MediaPipe+LSTM. After the optimization strategy described in section 3 (packet convolution, channel pruning, TensorRT deployment, inter-frame feature caching), the single-frame inference delay of PDDS-Net is 41.8ms, slightly higher than 32.5ms of MediaPipe+LSTM, but the latter can only realize action classification and cannot provide location-level feedback. Under the premise of comparable feedback granularity, the delay of PDDS-Net still meets the requirements of real-time interaction (<50ms), and the throughput is about 23.9FPS.

4.4 Generalization and robustness experiments

The cross-dataset generalization experiment uses a scheme trained on the NTU RGB+D dataset and directly tested on the PKU-MMD subset without any fine-tuning to verify the adaptability of the model across data domains. The two data sets have significant differences in acquisition equipment, subject group, action execution style and shooting perspective, which constitute challenging generalization test scenarios. The cross-dataset accuracy of PDDS-Net is 78.6%, which is significantly higher than ST-GCN's 71.2% and Action Transformer's 73.8%. According to the analysis, the advantages of PDDS-Net come from two aspects: first, the template matching branch (s_{DTW}) does not depend on the distribution of training data and can still provide stable reference when the domain is shifted; Second, the part decoupling evaluation mechanism has a stronger tolerance to the change of action execution style.

The occlusion robustness experiment tests the decision stability of the model by randomly discarding the key points of the input pose sequence (simulating partial joint occlusion in the real scene). The drop ratio gradually increased from 0% to 40%, and the decrease in AUC of each model in the quality assessment task was recorded. Figure 1 illustrates the degradation curves of the three models, with the proportion of joint discarded on the horizontal axis and the normalized AUC on the vertical axis.

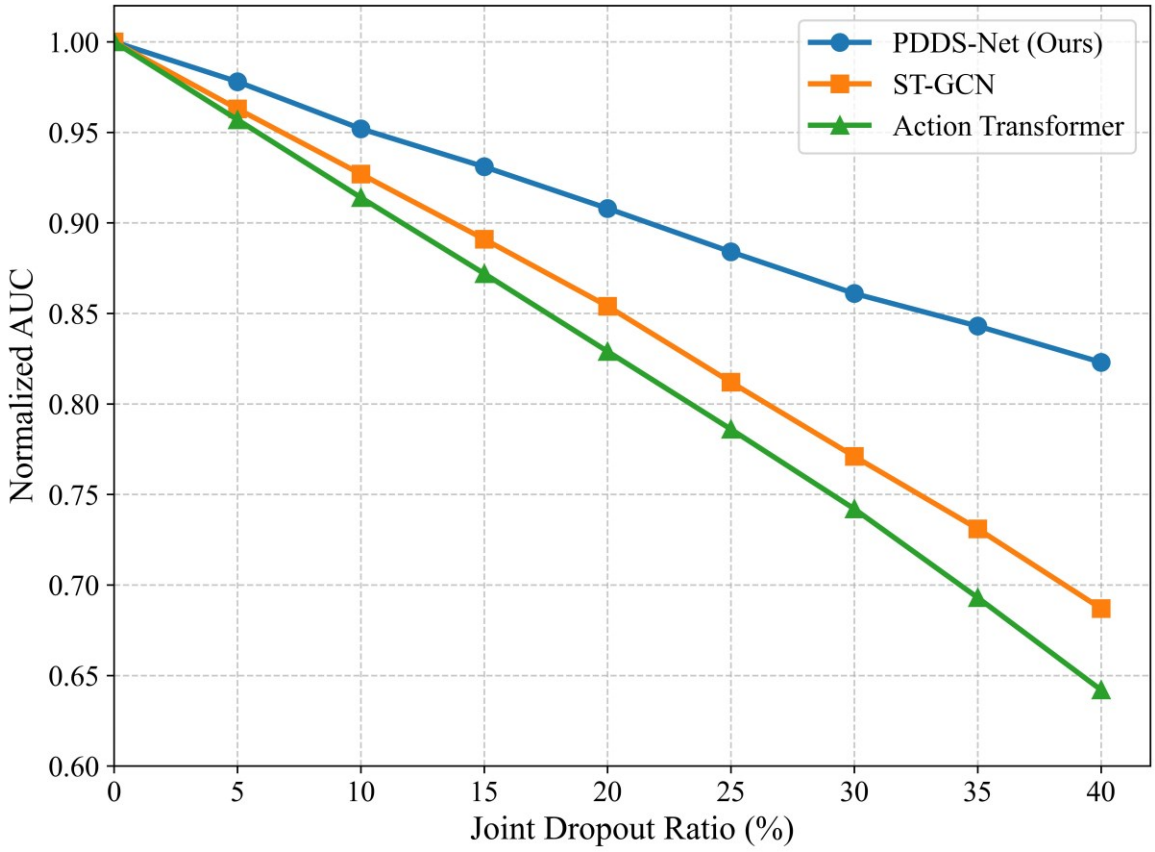

Figure 1. Quality assessment performance degradation curves of different models under random joint occlusion

PDDS-Net still maintained the original AUC of 82.3% at 40% joint drop rate, which was much smaller than ST-GCN (-31.3%) and Action Transformer (-35.8%). This robustness advantage is mainly attributed to the multi-source fusion mechanism. When the confidence between p_{cls} and s of the deep

network branch decreases due to the degradation of pose estimation, the gated network automatically increases the weight of template matching branch s_{DTW} to maintain the stability of the overall decision.

The perspective generalization experiment is based on the multi-view acquisition setting provided by the NTU RGB+D dataset, and the Cross-view evaluation protocol (training with samples collected by camera 2 and 3, and testing with samples collected by Camera 1) is used to validate the model. The accuracy rate of PDDS-Net in the cross-view test was 89.4%, down 4.3 percentage points compared with the Cross-subject protocol test result (93.7%) on the same dataset, and the decrease was smaller than that of Action Transformer (-6.1%) and PoseC3D (-5.2%). It shows that the model is robust to different shooting angles.

4.5 Feedback quality assessment

Feedback quality evaluation focuses on three aspects: location positioning accuracy, feedback time delay analysis and interpretability visualization. The location positioning accuracy is calculated as follows: the high attention region output by the model (the attention weight is greater than 1.5 times of the global mean) and the deviation region marked by the template are calculated by spatial intersection ratio. Evaluation on the validation set showed that the localization IoU was 0.637 for the upper limb group, 0.591 for the trunk group, and 0.614 for the lower limb group. The difference in positioning accuracy between parts may be due to the larger movement amplitude and higher visual discrimination of the upper limb joints, while the movement pattern of the trunk joints is more subtle, making it relatively difficult to locate. Feedback time delay analysis divides the end-to-end processing process into four stages: attitude estimation, feature extraction, decision reasoning and feedback assembly. The average time of each stage is 12.3ms, 14.1ms, 8.7ms and 6.7ms respectively (a total of 41.8ms), and attitude estimation and feature extraction account for 64.8% of the total delay. It is the key direction of subsequent optimization.

The explainability visualization shows the effect of bias localization for a typical set of samples. Figure 2 shows the case of deviation in the execution of the forward flexion of the shoulder joint. The left side is the color coding of the original video frame overlay node (the deviation node is marked in red), the middle is the attention thermogram overlay (the red area focuses on the right shoulder joint), and the right side is the text correction suggestion that "the flexion angle of the right shoulder joint is 122°, 7° beyond the standard range. It is recommended to reduce the lifting range to between 110° and 115°".

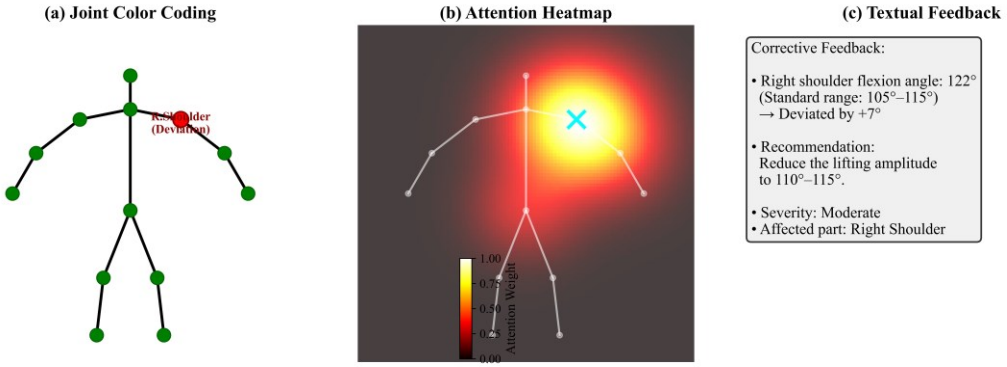

Figure 2. Deviation location and feedback visualization

The thermal map showed that the attention weight was highly concentrated around the right

shoulder joint, which was basically consistent with the deviation area marked by the template comparison, which verified the effectiveness of the attention map comparison learning method in section 3 to achieve accurate deviation location without frame-level annotation.

5. DISCUSSION

This section focuses on the PDDS-Net model proposed in this study and its feedback generation mechanism and conducts a comprehensive discussion from the perspectives of methodological effectiveness, rationality of architecture design, deep meaning of experimental results and research limitations, aiming to provide critical analysis and reflection for subsequent research.

The proposed part-decoupled dual-flow architecture achieves better performance than the existing baseline on both action recognition and part-level feedback positioning tasks, which verifies the core design hypothesis of "co-optimization of recognition and evaluation". From the perspective of information theory, the strategy that the global action recognition flow and the local part evaluation flow share the underlying pose coding at the feature level and integrate at the decision level actually constructs a symbiotic relationship of multi-granularity feature representation. The global flow provides the context prior at the action category level for the local flow. So that the part evaluation flow can judge the standard degree of each part in the correct action semantic framework; The local flow in turn provides the fine-grained evidence at the position level for the global flow, so that the global classification decision not only depends on the overall pattern matching but also can be supported by the position quality information. The significant decrease in model performance after removing any first-class in the ablation experiments further confirms the irreplaceable nature of this symbiotic relationship. It is worth further discussion that the accuracy rate of 68.4% of the single-flow-part variant in the action recognition task is still higher than that of the random baseline, indicating that the location-level features themselves also contain certain action category discrimination information. However, the performance gap of more than 25 percentage points between its model and the full dual-flow model reveals the key role of global temporal semantic information in action recognition. Part motion pattern itself is not enough to uniquely determine the action category, but the execution order, overall coordination and temporal structure of actions are also indispensable.

The human joints are divided into three functional parts: upper limb group, trunk group and lower limb group for independent evaluation. This design draws lessons from the clinical practice logic of "segmental evaluation" in rehabilitation medicine. Instead of treating the body as an indivisible whole, rehabilitation therapists often analyze the quality of movement from proximal to distal and from core to extremity when guiding patients. The part decoupling strategy extracts feature of each part group through an independent graph convolution subnetwork, which is essentially embedded in the deep learning model with this clinical prior knowledge. Each subnetwork focuses on learning the motion mode of this part, avoiding the interference of large upper limb vibration features on the discrimination of trunk stability. However, there are some limitations to this hard position grouping strategy. Human movement is a complex process of multi-part coordination, and the upper limb movement is inevitably accompanied by the stable support of the torso and the attitude adjustment of the lower limbs. Completely independent part coding may break the movement coordination relationship that should exist between parts. This paper partially alleviates this problem through the part score aggregation and compensatory action recognition mechanism of multi-source fusion decision-making layer. However, how to realize the independent part discrimination and inter-part coordination evaluation simultaneously in a unified framework is still a direction worthy of further exploration.

The attention graph contrast learning method proposed in this paper only needs video-level annotation to train the deviation location module, which significantly reduces the data annotation cost compared with the traditional frame-level annotation scheme. The experimental results show that although the positioning accuracy of the optimized attention map (IoU 0.637 of the upper limb group) is enough to support the generation of part-level feedback, there is still a certain gap between the positioning accuracy of the optimized attention map and the upper limit of the positioning under full supervision (IoU 0.75 or more can be reached by the classifier trained with frame-level annotation). The essence of this gap comes from the inherent ambiguity of weak supervision information. Video-level annotation only informs that "the action has deviation" but does not specify the specific frame and specific location where the deviation occurs. Contrast learning guides focus attention through indirect supervision of template feature differences, which is essentially an approximate optimization using proxy supervision signals. To further improve the positioning accuracy without introducing additional manual annotation, it may be necessary to explore more refined agent supervision signal design, such as using human kinematics constraints (continuity of joint angle change rate, limb length invariance, etc.) as regularization term, or introducing multi-template comparison strategy to eliminate the possible bias introduced by a single template.

In the occlusion robustness experiment, PDDS-Net still maintains 82.3% of the original AUC at 40% joint drop rate, which is significantly better than the pure data-driven comparison model, confirming the effectiveness of the multi-source fusion mechanism under the condition of pose estimation degradation. The adaptive adjustment of signal source weight by gated network in different scenarios (focusing on global classification in complete scenario, focusing on DTW matching in occlusion scenario) intuitively shows the working mechanism of fusion strategy. However, it should be noted that this improvement in robustness comes at the cost of additional computational overhead in the inference stage, the DTW branch needs to dynamically program and align each sample with the template library during inference, and its time complexity is $O(T \cdot T_{ref})$. Although latency can be kept within acceptable limits through template library size control (less than 50 templates) and precomputational optimization, it can still become a bottleneck on edge devices with extremely limited resources. Future work can consider distilling template matching branches into lightweight networks to further reduce computational overhead while maintaining robustness gains.

In the comparison experiment, an interesting phenomenon is that although the number of parameters of Action Transformer (18.2M) is greater than that of PDDS-Net (13.2M), the accuracy of action recognition is 2.2 percentage points lower. This counterintuitive result may reflect some inherent defects of pure Transformer architecture in skeleton sequence modeling. The standard self-attention mechanism globally models the dependencies between all frames and theoretically has the strongest long-range capture ability, but this "global connection" may become a source of noise in rehabilitation action scenes. Because the local timing dependence between adjacent frames is more critical for action discrimination, the correlation between distant frames is often weak. The local time series convolution of ST-GCN in PDDS-Net exactly explicitly models this local dependency relationship and combines with the subsequent time series Transformer for global information aggregation, forming a hierarchical time series modeling strategy of "from local to global", which may be more suitable for the characteristics of rehabilitation action data.

All experiments in this paper are based on open data sets and standard template libraries, actively avoiding the collection of clinical data of real patients, which has advantages and disadvantages at the methodological level. The advantage lies in the simplification and acceleration of the research process, without going through the long approval cycle of the ethical review board, without facing the

operational complexity of patient recruitment and informed consent, and the research can focus on the algorithm innovation itself. However, this choice also introduces inevitable limitations: the subjects in the public data set are mostly healthy volunteers or movement performers, whose movement deviation pattern is fundamentally different from that of real recovered patients (motor dysfunction caused by disease). Deviation execution in healthy subjects tends to be "intentional" or "simulated", while deviation in patients is pathological and non-random, and the motor control mechanisms and deviation distribution characteristics of the two are quite different. Therefore, the effectiveness of the model in this paper verified on public data sets and to what extent it can be transferred to real rehabilitation patients need to be confirmed in subsequent clinical validation. The existence of this limitation does not mean that the value of this study is weakened but suggests a feasible and phased research promotion strategy. First, complete full validation and iterative optimization at the algorithm level on the open benchmark and then advance to the real clinical validation stage with ethical approval after the model performance is mature. This study provides a complete technical scheme and validation paradigm for the first half of this strategy.

6. CONCLUSION

Focusing on the problem of action recognition and real-time feedback of rehabilitation training based on computer vision, this paper proposes a complete technical framework from feature extraction, dual-stream fusion decision to position-level feedback generation and carries out systematic experimental verification on open data sets. This section concludes at three levels: research contribution, methodological orientation and future direction.

The contribution of this paper can be summarized in three dimensions: methodological innovation, technical design and research paradigm. At the methodology level, a dual-flow time sequence fusion network PDDS-Net with position decoupling is proposed. By decoupling the action evaluation task into global action recognition flow and local position evaluation flow, a unified framework with both coarse-grained classification ability and fine-grained positioning ability is constructed, which solves the core problem of separation between recognition and evaluation in existing methods. At the technical design level, a deviation location method based on attention graph contrast learning is developed, which uses the attention weight of local evaluation flow to localize the key joints of execution deviation. The training process only needs video-level annotation, which alleviates the bottleneck problem of high cost of frame-level annotation from the algorithm level. At the same time, the multi-source feature adaptive fusion mechanism is proposed to dynamically integrate the depth confidence degree and template matching score through the gated network, which significantly improves the decision stability of the model under the condition of attitude degradation. At the level of research paradigm, all experiments in this paper were completed based on public data sets and standard template libraries, providing a complete technical path for rehabilitation action evaluation research without ethical approval, and providing a replicable example for the rapid advancement of similar research in the algorithm iteration stage.

There are several limitations in this paper that need to be further improved. The first is the limitation of data domain. The difference between the motion deviation mode in the open data set and the pathological motion mode of real rehabilitation patients makes the effectiveness of the model transfer to clinical scenarios need to be verified, and the fundamental solution of this problem needs to be verified in the subsequent clinical verification of real rehabilitation patients. The second is the simplification of part grouping strategy. Although the division of upper limb, trunk and lower limb

conforms to clinical intuition, it fails to cover the needs of more refined part grouping and does not fully consider the differentiated needs of different movements for part combination in rehabilitation training. Third, the current model evaluates a single action completed independently, which lacks support for continuous rehabilitation training sequences and complex scenes with multiple action combinations. Based on the above reflection, the expansion directions of future work mainly include the following three aspects. First, the exploration of self-supervised pre-training strategy, using large-scale unlabeled skeleton sequence data for pre-training, learning general motion representation through self-supervised tasks such as mask reconstruction or contrast learning, and then fine-tuning on a small amount of labeled data, in order to further reduce the dependence on labeled data and improve the adaptability of the model to different data domains. Second, multi-movement continuous assessment and transition state analysis, which extends the current framework of dealing with isolated movements to online monitoring of continuous movement sequences, not only evaluates the completion quality of each movement, but also analyzes the fluency and coordination in the transition stage between movements, which also has important clinical significance in rehabilitation evaluation. Third, in-depth optimization of end-to-end deployment. Aiming at the resource constraints of mobile terminals and embedded platforms, combined optimization research on model quantification, knowledge distillation and neural network architecture search technologies are carried out. On the premise of ensuring feedback accuracy, reasoning delay is further compressed, so that the system can be truly integrated into smart phones and other terminal devices. It provides convenient, economical and intelligent technical support for remote home rehabilitation.

Abbreviations

PDDS Net, Part Decoupled Dual Stream Network;

DTW, Dynamic Time Warping;

ST-GCN, Spatio-temporal Graph Convolutional Network;

LSTM, Long Short-Term Memory;

Bi-LSTM, Bidirectional Long Short-Term Memory;

CNN, Convolutional Neural Network;

RNN, Recurrent Neural Network;

MHSA, Multi-Head Self-Attention;

FFN, Feed-Forward Network;

MLP, Multi-Layer Perceptron;

ReLU, Rectified Linear Unit;

AUC, Area Under the Curve;

ROC, Receiver Operating Characteristic;

FAR, False Acceptance Rate;

TAR, True Acceptance Rate;

FPS, Frames Per Second;

IoU, Intersection over Union;

GPU, Graphics Processing Unit;

CPU, Central Processing Unit;

HRNet, High-Resolution Network;

NTU RGB+D, Nanyang Technological University RGB+D Dataset;

PKU-MMD, Peking University Multi-Modal Dataset;

TP, True Positive;

TN, True Negative;

FP, False Positive;

FN, False Negative;

TensorRT, NVIDIA TensorRT (inference optimization library);

MediaPipe, Google MediaPipe (framework, not an acronym but commonly used);

OpenPose, OpenPose (framework, not an acronym).

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have

inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **M.Y.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **M.Y.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used ChatGPT for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Kamenov, K., Mills, J. A., Chatterji, S., & Cieza, A. (2019). Needs and unmet needs for rehabilitation services: a scoping review. *Disability and rehabilitation*, *41*(10), 1227-1237. https://doi.org/10.1080/09638288.2017.1422036

[2] Al Imam, M. H., Jahan, I., Das, M. C., Muhit, M., Akbar, D., Badawi, N., & Khandaker, G. (2022). Situation analysis of rehabilitation services for persons with disabilities in Bangladesh: identifying service gaps and scopes for improvement. *Disability and Rehabilitation*, *44*(19), 5571-5584. https://doi.org/10.1080/09638288.2021.1939799

[3] Dutta, D., Sen, S., Aruchamy, S., & Mandal, S. (2022). Prevalence of post-stroke upper extremity

paresis in developing countries and significance of m-Health for rehabilitation after stroke-A review. *Smart Health*, *23*, 100264. https://doi.org/10.1016/j.smhl.2022.100264

[4] Martinez-Martin, E., & Cazorla, M. (2019). Rehabilitation technology: assistance from hospital to home. *Computational intelligence and neuroscience*, *2019*(1), 1431509. https://doi.org/10.1155/2019/1431509

[5] Kyriazakos, S., Schlieter, H., Gand, K., Caprino, M., Corbo, M., Tropea, P., ... & Lynggaard, V. (2020). A novel virtual coaching system based on personalized clinical pathways for rehabilitation of older adults—requirements and implementation plan of the vCare project. *Frontiers in Digital Health*, *2*, 546562. https://doi.org/10.3389/fdgth.2020.546562

[6] Zhou, Y., Rashid, F. A. N., Mat Daud, M., Hasan, M. K., & Chen, W. (2025). Machine learning-based computer vision for depth camera-based physiotherapy movement assessment: A systematic review. *Sensors*, *25*(5), 1586. https://doi.org/10.3390/s25051586

[7] Hulleck, A. A., Menoth Mohan, D., Abdallah, N., El Rich, M., & Khalaf, K. (2022). Present and future of gait assessment in clinical practice: Towards the application of novel trends and technologies. *Frontiers in medical technology*, *4*, 901331. https://doi.org/10.3389/fmedt.2022.901331

[8] Jeyabharathi, J., SK, G. P., VenuMadhav, G., Dinesh, G., & Saikumarreddy, C. V. V. (2025, January). Real Time Gameplay using Pose Detection. In *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)* (pp. 801-805). IEEE. https://doi.org/10.1109/ICMSCI62561.2025.10894142

[9] Sykes, E. R. (2025). Next-generation fall detection: harnessing human pose estimation and transformer technology. *Health Systems*, *14*(2), 85-103. https://doi.org/10.1080/20476965.2024.2395574

[10] Yarko, E. I. (2025). Comparative Analysis of Libraries for Human Pose Detection in Mobile Device Environments. *Automatic Documentation and Mathematical Linguistics*, *59*(Suppl 3), S220-S228. https://doi.org/10.3103/S0005105525700955

[11] Zhang, H. B., Cai, J. J., Qiu, H. M., Lei, Q., Liu, J. H., & Zhang, M. H. (2026). A survey of deep learning-based action quality assessment: methods and applications. *Journal of Big Data*, *13*(1), 70. https://doi.org/10.1186/s40537-026-01409-5

[12] Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). Deep canonical time warping for simultaneous alignment and representation learning of sequences. *IEEE transactions on pattern analysis and machine intelligence*, *40*(5), 1128-1138. https://doi.org/10.1109/TPAMI.2017.2710047

[13] Wang, Z., Zhang, Z., Su, T., Ding, Z., & Zhao, T. (2024, December). Research on supply chain network optimisation based on the CNNs-BiLSTM model. In *2024 International Conference on Information Technology, Comunication Ecosystem and Management (ITCEM)* (pp. 197-202). IEEE. https://doi.org/10.1109/ITCEM65710.2024.00044

[14] Zhao, T., Chen, G., Pang, C., Li, L., & Busababodhin, P. (2026). Forecasting global agricultural trade imbalances using a hybrid deep learning and gradient boosting framework. *Discover Computing*, *29*(1), 443. https://doi.org/10.1007/s10791-026-10367-8

[15] Chen, G., Zhao, T., Pang, C., & Busababodhin, P. (2026). Integrated CNN–LSTM–XGBoost

hybrid model predicts shale oil seismic attributes and global oil price trends. *Scientific Reports*. https://doi.org/10.1038/s41598-026-55910-1

[16] Kamal, S., Alshehri, M., AlQahtani, Y., Alshahrani, A., Almujally, N. A., Jalal, A., & Liu, H. (2025). A novel multi-modal rehabilitation monitoring over human motion intention recognition. *Frontiers in Bioengineering and Biotechnology*, *13*, 1568690. https://doi.org/10.3389/fbioe.2025.1568690

[17] Momin, M. S., Sufian, A., Barman, D., Dutta, P., Dong, M., & Leo, M. (2022). In-home older adults' activity pattern monitoring using depth sensors: A review. *Sensors*, *22*(23), 9067. https://doi.org/10.3390/s22239067

[18] Zhang, J., Wu, C., & Wang, Y. (2020). Human fall detection based on body posture spatio-temporal evolution. *Sensors*, *20*(3), 946. https://doi.org/10.3390/s20030946

[19] Wang, H., Ho, E. S., Shum, H. P., & Zhu, Z. (2019). Spatio-temporal manifold learning for human motions via long-horizon modeling. *IEEE transactions on visualization and computer graphics*, *27*(1), 216-227. https://doi.org/10.1109/TVCG.2019.2936810

[20] Yu, Q., Yang, G., Wang, X., Shi, Y., Feng, Y., & Liu, A. (2025). A review of time series forecasting and spatio-temporal series forecasting in deep learning: Q. Yu et al. *The Journal of Supercomputing*, *81*(10), 1160. https://doi.org/10.1007/s11227-025-07632-w

[21] Saravana, M. K., Roopa, M. S., Arunalatha, J. S., & Venugopal, K. R. (2026). Transformers for multivariate time series forecasting: Comprehensive analysis, challenges, research opportunities and future prospects. *IEEE Access*. https://doi.org/10.1109/ACCESS.2026.3654408

[22] Tian, Z., Zhao, F., Liu, H., & Bian, X. (2025). HDMFNet: Learning complementary spatiotemporal representations via heterogeneous dual-branch multimodal fusion network for human action recognition. *Applied Soft Computing*, 114391. https://doi.org/10.1016/j.asoc.2025.114391

[23] Chen, J., Zhang, Z., Sun, X., Zhou, Y., Zhou, Y., Zhao, Y., & Shi, J. (2026). A Multi-Source Data Fusion-Based Method for Safety Monitoring of Construction Workers on Concrete Placement Surfaces. *Buildings*, *16*(6), 1165. https://doi.org/10.3390/buildings16061165

[24] Phu, K. A., & Hoang, V. D. (2025). Predicting occluded skeletal joints via tracking-based feature extraction. *Neurocomputing*, 131004. https://doi.org/10.1016/j.neucom.2025.131004

[25] Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1010-1019). https://doi.org/10.1109/CVPR.2016.115

[26] Wu, H., Wang, C., Duan, Y., & Song, R. (2026). Motion Saliency Guided Fine-Grained Multimodal Interactive Learning for RGB-D Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2026.3708969

[27] Zhao, T., Chen, G., & Busababodhin, P. (2026). An intelligent expert system for logistics disruption prediction and mitigation in global supply chains: Integrating graph-based risk inference with ensemble forecasting. *Scientific Reports*. https://doi.org/10.1038/s41598-026-61617-0

APA Citation

Yang, M. (2026). Research on Rehabilitation Training Movement Recognition and Real-time Feedback Model Based on Computer Vision. *Journal of Discovery Core, 1*(1), 101-128. **https://doi.org/10.67541/jdc2605**