Transformer-based Multi-task Learning For Table Tennis Motion Feature Recognition

Tianfang Ma*

Physical Education Teaching and Research Department, Harbin Finance University, Harbin 150030 China

Corresponding author. E-mail: xdwangxd@163.com

Received: May. 12, 2025; Accepted: Jun. 10, 2025

In the process of multi-task sports motion behavior feature recognition, it is prone to be affected by few-shot samples, resulting in catastrophic forgetting phenomena, which leads to poor processing ability of variability. In order to solve the above-mentioned problems, this paper proposes a novel table tennis motion feature recognition method based on Transformer-based multi-task learning. This model adopts a grouped attention structure to enhance the extraction ability of local features, and adds the spatial information embedding and temporal information embedding modules to enhance the extraction of spatial and temporal features by the original Transformer model. The extracted chaotic invariant features are classified and recognized through the multi-task learning method by support vector machine to achieve the accurate recognition of multi-task table tennis motion features. The experiment results show that this new method can efficiently identify the motions of table tennis movement, accurately capture the subtle changes of joints, and perform excellently in both single/complex multi-tasks and cross-individual scenarios.

Keywords: multi-task table tennis motion; feature recognition; Transformer; multi-task learning; support vector machine

http://dx.doi.org/10.6180/jase.202603_29(3).0005

1. Introduction

The characteristics of human movement behavior refer to the dynamic features and behavioral patterns exhibited by the human body during various activities. These characteristics include but are not limited to the movement trajectories of joints, the contraction and relaxation of muscles, and changes in body posture, etc. With the advancement of technology and the deepening of applications, the recognition of human movement behavior characteristics has demonstrated tremendous potential and value in multiple fields such as video surveillance, human-computer interaction, intelligent rehabilitation, and sports analysis [1]. However, due to the complexity and diversity of human movement, recognition methods often have difficulty coping with the challenges in multi-task scenarios, such as occlusion, illumination changes, and diverse movement patterns. The recognition of human movement behavior characteristics, as an important research direction in the fields of computer vision and artificial intelligence, has received extensive attention and research from many scholars [2, 3].

Luo et al. [4] proposed a method for recognizing human motion behavior features that combined machine learning and multi-scale local spatio-temporal feature extraction. This method constructed a rich set of local spatio-temporal features by capturing the movement changes of the human body at different time and spatial scales. The Kalman filtering technique was utilized to predict the joint position and optimize the state estimation. The wavelet transform was adopted to optimize the neural network. The extracted feature points were taken as input to train the model to achieve efficient recognition of local feature points of human movements. Although this method had improved the

recognition accuracy, the steps such as multi-scale feature extraction, Kalman filtering prediction and neural network training all involved complex calculations, which could affect the real-time performance of the overall algorithm and make it difficult to meet the application scenarios with high real-time requirements. Lovanshi and Tiwari [5] employed a method that combined pose estimation, shift graph convolution, and dense residual connection to extract human motion behavior features. Due to the fact that human movement behaviors in different fields (such as sports, dance, daily life, etc.) had their own characteristics, this method faced challenges in adapting to new data distributions and features when applied across fields. Wang et al. [6] proposed a two-stream network architecture based on adaptive fusion weights to solve the problems of insignificant joint features and noise interference in human behavior recognition. This method combined time-domain segmentation and entropy method, which segmented the video in the time domain and used the dual-stream network of BN-Inception to extract the static spatial information and dynamic motion features respectively. In each video clip, the entropy method was used to initially fuse these two kinds of information. The network weights were automatically adjusted according to the saliency of the joint points to emphasize the salient features and suppress the noise, thereby identifying the joint features in human behavior. However, the time flow network was highly dependent on the accurate extraction of bone joint points. If there were errors or noises in the bone data, it would directly affect the construction of the motion feature vector and the subsequent recognition performance. Xiao [7] adopted the improved version of DD-Net (dual-feature dual-motion network) for human behavior feature recognition. This method enriched spatial features by introducing new branches and normalized joint coordinates (NCJ), and realized the recognition of human action features by fusing global trajectory information and enhanced spatial information. However, this method had difficulties in identifying actions that had a weak connection with the global trajectory. This method could not be sensitive enough when capturing local subtle movements or those that did not rely on the overall body movement pattern, limiting the wide range of its application.

For the recognition of multi-task human motion behavior characteristics, it is necessary to handle multiple different motion patterns and scenarios, and the recognition method must have a high degree of complexity and diversity processing capability [8, 9]. Chaotic invariants, as a kind of characteristic parameters that can describe complex dynamic characteristics, can extract key information from nonlinear dynamic systems and better describe the complexity and variability of human motion. To better extract the motion characteristics of the human body by using the self-attention mechanism, Transformer-based methods have emerged [10]. These methods can adaptively capture the non-physical connections between joints and optimize the modeling ability of skeleton features such as HyperFormer [11], IGFormer [12] and TAG network [13]. Although these methods integrate the spatial and temporal information of the skeleton data to a certain extent, it fails to fuse the inherent adjacency relationship between the joint diagram structures, resulting in a decrease in the recognition accuracy of occlusion and similar joint movements. So this paper proposes a novel table tennis motion feature recognition method based on Transformer-based multi-task learning.

2. Materials and methods

The skeletal data of athlete body movements can be decomposed from two dimensions: chronological order and space. When decomposed in the dimension of time, each frame in the action time series represents a certain moment of athlete movement. In this paper, it is represented by X_t, specifically referring to the set of three-dimensional coordinates of each joint point on the athlete skeleton of this frame, that is, $X_t \in R^{25\times3}$. The three-dimensional coordinate set of the skeleton joint points can be obtained by the KinectV2 depth camera in combination with the joint point estimation algorithm, and the spatial coordinates of the 25 identified athlete skeleton joint points and each joint point can be obtained.

It combines the arrays of these single-frame athlete body joint point data according to the time dimension. After adding the time dimension, the complete skeleton data X of athlete movements can be obtained, that is, $X = \sum X_t \in R^{T\times25\times3}$.

The model extracts and dimensionalizes the spatial data of each frame skeleton layer by layer through five layers of spatial feature encoding modules. It enhances the spatial dimension characteristics of the joint points to obtain high-dimensional spatial vectors. Then, the dynamic information of the 15-layer time series feature encoding module on the time series is obtained, and the residual alternating connection is used to save the features at different levels. Finally, the output of the classification results is achieved through the classification head [14, 15].

2.1. Skeleton space feature encoding module

The spatial characteristics of the movement of the athlete skeleton are contained in the changes of the spatial coordi-

nates of each joint point. If the spatial coordinates of the joint points are taken as the eigenvalues of the nodes in the graph structure, the skeleton motion information of the graph structure can be obtained, that is, the skeleton graph of athlete motion. By using the topological kinematic characteristics of the skeleton graph and the node aggregation ability of graph convolution, the spatial characteristics of the skeleton motion can be calculated. The self-attention mechanism can calculate the interaction relationships among all global joint points to achieve feature aggregation. Therefore, in this paper, the skeleton spatial feature encoding module is utilized to extract the spatial feature information in athlete motion.

In order to optimize the ability of the self-attention model to extract the spatial information of the skeleton, this paper proposes a hierarchical attention mechanism to optimize the spatial feature extraction ability of the model through the local attention of the limbs and the global self-attention of the skeleton. The athlete skeleton can be divided into five groups according to the left arm, right arm, left leg, right leg and trunk. Movements can also be classified into two types based on the amplitude of the movement: large amplitude and small amplitude [16, 17]. When identifying large-scale movements, attention should be paid to the relationship between the limbs and the trunk. For small-scale movements, it is necessary to pay attention to the spatial position change relationship among each joint point within a single skeleton group. Therefore, in order to extract athlete motion features of different amplitudes, this paper proposes a hierarchical attention mechanism for modeling the spatial features of the skeleton.

This method sets up a multi-layer spatial structure modeling mechanism, starting from the local and then the global. For the modeling of local features within the group, this paper achieves it through the multi-head attention module within the group. First, it groups the original input X according to the limb part, and adds the specified label to each group. Then it obtains $X_{la}, X_{ra}, X_{ll}, X_{rl}$, and X_{bd} respectively, and takes them as the inputs of each local feature extraction module. Taking the local feature extraction of the left arm part as an example. There are a total of 6 joint points in the left arm part. The input feature matrix is X_{la}. Each point in the feature matrix contains three spatial dimension coordinates. The spatial features of the joint points within the group are extracted in the self-attention mechanism module.

Firstly, the original skeleton data X_{la} of the left arm is processed through the learnable weight matrix W_q, W_k, W_v to obtain the three dimensional components q, k, and v, that is,

$$q = X_{la} W_q \tag{1}$$

$$k = X_{la} W_k \tag{2}$$

$$v = X_{la} W_v \tag{3}$$

Then, the mutual attention scores between each joint point are calculated respectively using q and k, that is, the importance of the correlation in the action. Then, the complete attention weight matrix A is obtained by organizing and mapping through the softmax function. During the process, it is also necessary to expand the information dimension of the joint points through the multi-head attention mechanism, extend the input information to a higher information dimension, and model the input features comprehensively. Specifically as shown in equation (4), where C and h are the joint feature dimension and the number of self-attention heads respectively.

$$A = \text{softmax}\left(qk^T / \sqrt{C/h}\right) \tag{4}$$

Finally, through equation (5), it multiplies A by v to obtain the output X'_{la} calculated by attention.

$$X'_{la} = Av = \text{ selfattention } (X_{la}) \tag{5}$$

After the calculation of the multi-head attention module within the group mentioned above, the features of each point contain its position information on the skeleton and the correlation features among them.

For the modeling of the overall features, this paper achieves it through the global attention module. After completing the local feature extraction of the skeleton grouping, it is also necessary to restore it to the overall skeleton form, that is, to combine the output vectors to obtain the vector X_{full} of the complete skeleton mode as the input of the global feature extraction module.

2.2. Skeleton Transformer

To solve the problem that the original spatial information of the skeleton graph structure is not utilized in the traditional Transformer method, this paper optimizes the extraction ability of the Transformer for the spatial features of the skeleton graph by adding the spatial position encoding module and the centrality encoding module.

The original topology of the nodes in the skeleton graph can be represented by the adjacency matrix $A \in R^{V \times V}$. Black squares indicate that there are physical connections between the corresponding 2-joint nodes. It is shown in figure 1.

The input X_{in} of the spatial feature encoding module is obtained by combining the human body joint point array X_{full} of a single frame with its adjacency matrix A. At

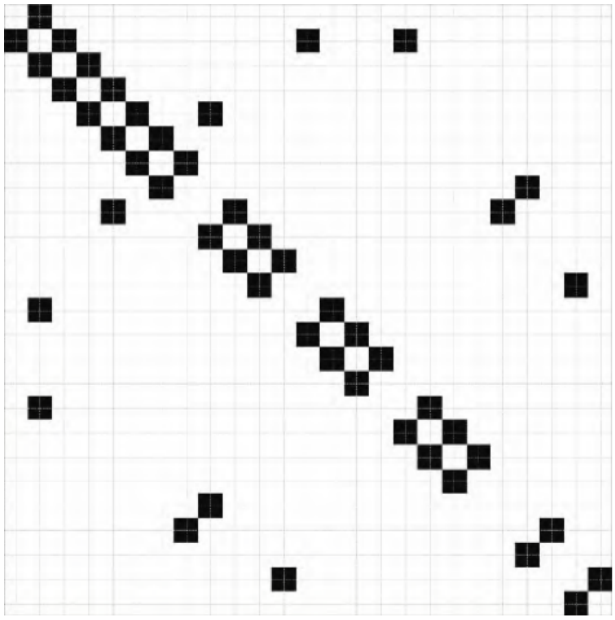

Fig. 1. Diagram of the skeleton adjacency matrix

this point, X_{in} does not yet have the complete topological information of the skeleton graph. To improve the expressive ability of its graph structure data, this paper proposes the centrality encoding of the athlete skeleton and uses centrality to quantify the connection importance between skeleton joint points. Taking the degree of each joint as the initial input Z_{deg} of the centrality of the joint, it extends the dimension and the learnable coefficient matrix to make it the same dimension as the original feature matrix, and then adds the two together. The centrality information of the skeleton graph is embedded into the original features of the nodes as the input of the skeleton's self-attention.

$$X_{in} = X_{in} + W_c \times Z_{deg} \tag{6}$$

The X_{in} after adding the centrality encode contains information on the importance of each joint point in athlete movement. However, the model still needs to learn the relative spatial positional relationship between the joint points. Therefore, it is also very necessary to introduce the spatial relative geometric relationship between the joint points as a priori.

This paper proposes a learnable spatial position encoding method for skeleton graphs. The core is to construct the shortest path distance (SPD) function in space. By calculating the Euclidean spatial distance between each group of nodes, the spatial dependence between nodes is measured. This method can obtain the initial position matrix D as:

$$D_{d(i,j)} = W_{S(i,j)} \times \left(SPD\left(V_i - V_j \right) \right) \tag{7}$$

Where SPD is a function for calculating the shortest Euclidean distance between two joint points. Furthermore, it is necessary to attach learnable parameters to the distance matrix in the $V \times V$ dimension and transform it into a learnable spatial position embedding matrix D_d, which further enriches the amount of information of the input model in the calculation process of the global attention of the skeleton.

$$A_{MHSA(i,j)} = \frac{\left(h_i W_q\right)\left(h_j W_k\right)^T}{\sqrt{d}} + D_{d(i,j)} \tag{8}$$

$$X_{S(i,j)} = A_{MHSA(i,j)} \times X_{in(i,j)} \tag{9}$$

In the input and calculation processes of the module, Z_{deg} and D_d are introduced respectively, and the summary of the local relationships extracted in the previous step is achieved through the multi-head self attention (MHSA) mechanism. It obtains the contribution degree of each node to feature extraction in this motion and the correlation degree between the joint nodes. Finally, the computational output of the spatial module is obtained, namely the weighted feature X_S of the skeleton joint points after spatial feature extraction.

3. Results and discussion

3.1. Experiment data sets

The experiment employed the UCF101 dataset, which contains over 13000 video clips and covers 101 movement categories, ranging from sports and dance to a wide range of human activities in daily life. These movements are accomplished collaboratively by multiple parts of the human body, such as the upper limbs, legs and trunk. Some samples are shown in Figure 2. The GPU used in the experiment is RTX4090 and the video memory capacity is 24GB. The Pytorch framework with Python version 3.9 is used. The batch size used for training is 32 , the initial learning rate is 0.02 , and the cosine annealing learning rate optimizer is used. The model is trained for a total of 168 rounds.

3.2. The recognition effect of chaotic invariant features in motion recognition

Five typical athlete movement behaviors, namely walking, running, jumping, falling and sitting/standing, are selected from the experimental dataset as the recognition tasks. For each task, the corresponding chaotic invariant features are extracted respectively, and the corresponding motion behavior features are identified. The recognition performance

Table 1. Recognition results based on chaotic invariant features

Task type	Features of chaotic invariants	Recognition rate/%
Walking	Related integral, Lyapunov index	90.61
Running	Related integrals, Kolmogorov-sinai entropy	92.12
Jumping	Lyapunov index, maximum entropy	93.63
Falling	Chaotic invariant combination	95.21
Sitting/standing	Correlation integral, correlation dimension	91.94

Fig. 2. Sample images in the UCF101 dataset

of this paper is evaluated through tests on different tasks. The results are shown in Table 1.

Table 1 shows that the proposed method has the highest accuracy rate of fall recognition among different task types, reaching 95.21%. This may be related to the relatively significant chaotic characteristics of the fall action itself. Meanwhile, jumping recognition and running recognition also achieve good results, reaching 93.63% and 92.12% respectively. Although the accuracy rate of walking and sitting/standing recognition is slightly lower than that of other tasks, it still remains above 90%, demonstrating the good adaptability of this method among different tasks. It is indicated that the proposed method in this paper has stable and efficient recognition capabilities in the identification of different types of athlete movement characteristics.

3.3. Other comparison results

The NTU-RGB+D dataset is a commonly used benchmark dataset for the recognition of human skeleton modal actions [18]. It has 60 types of human actions and contains 56880 complete skeleton action sequences. Each action contains the three-dimensional information of 25 skeleton joint points for each people. And three KinectV2 depth cameras simultaneously capture images from different perspectives at the same horizontal height. The final generated data contains two subsets, CS (cross-subject) and CV (cross-view), which are respectively used to test the model performance on the two evaluation benchmarks of cross-subject CS and cross-view CV. In the CS subset, the subject characters in the training set and the test set come from two groups respectively, and there are 20 subjects in each group. The training set of the CV subset includes 37920 samples captured by depth camera perspectives 2 and 3, while the test set is 18960 samples captured by depth camera perspective 1 . We test the robustness of the model in terms of human body shape and camera perspective respectively.

On the NTU-RGB+D dataset, the performance of the proposed model is compared with that of other advanced action recognition methods. The recognition accuracy and parameter count of the model are evaluated respectively under the two criteria of CS and CV. The specific experimental results are shown in Table 2.

Table 2. Comparison of experimental results on the NTU60RGB+D dataset

Model	CS/%	CV/%	Parameter/MB
P-LSTM [19]	63.0	70.4	8.7
IndRNN [20]	81.9	88.1	10.2
ST-GCN [21]	81.6	88.4	3.2
2s-AGCN	88.6	95.2	7.8
ST-TR [22]	90.0	96.2	12.5
MSSTNet [23]	89.7	95.4	18.5
Sem-GCN [24]	86.3	94.3	16.3
Proposed	90.5	96.2	5.0

The proposed model has an action recognition accuracy rate of 90.5% under the CS standard, and has the highest recognition accuracy rate for large-scale actions such as one-legged jumps and hugs. It also has a relatively high recognition accuracy rate for minor movement faults and hand-waving movements.

Compared with the CS standard, the proposed model has a higher accuracy rate under the CV evaluation index, reaching 96.2%. During the training stage, this model learns the skeleton action data from different perspectives. Moreover, the centrality encoding and spatial position embedding of nodes can enable the model to pay more attention to the interaction relationships among the key points inside the skeleton, thereby reducing the impact brought

by different perspectives. It has good generalization in occasions where perspectives are prone to change, such as human-computer interaction and monitoring. Moreover, due to the smaller number of layers of the model, the number of parameters can be controlled within a more excellent range. The total number of parameters of 5.0 MB makes it more conducive to deployment.

Furthermore, compared with the ST-GCN baseline model, the proposed model in this paper utilizes a hierarchical attention mechanism to focus on features of different scales and has achieved better classification results for actions with highly similar motion paths, such as eating and drinking, reading and writing, etc. The specific improvement effects of each movement category are shown in Figure 3.

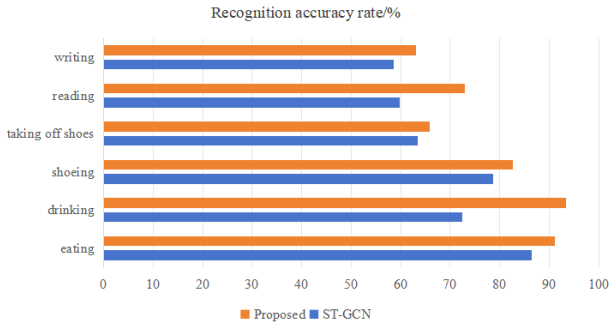

Fig. 3. Partial action recognition accuracy comparison

To verify contributions of the added central-EM (CEM), spatial-EM (SEM), group fusion strategy (G), temporal convolutional network (TCN), and temporal self-attention temporal-SA (TSA) modules to the model, the ablation comparison experiments are respectively designed. The comparison of the ablation experiment results is shown in Table 3.

It can be known from Table 3 that different functional components added to the model have different degrees of contribution to the accuracy of action recognition. According to experiments 1,3 and 6 , it can be known that the grouping strategy contributes significantly to the accuracy of model training. After replacing the hierarchical attention mechanism with the single-layer self-attention mechanism, the recognition accuracy of the model decreases. Moreover, it can be known from experiments 1 and 2 that the temporal self-attention model also makes a significant contribution in the modeling of sequence features. After replacing all of them with one-dimensional convolution, the data volume increases and the accuracy rate decreases by about 1.3%. It can be known from experiments 1, 4, 5 and 6 that the functional module embedded with graph structure information has also made corresponding contributions in the model, effectively improving the modeling ability of the Transformer method on the skeleton graph as the structural data.

Table 3. Comparison of ablation experiment results of spatial position encoding module

No.	Model	CS/%	CV/%
1	G+CEM+SEM+TSA	90.5	96.2
2	G+CEM+SEM+TCN	89.2	94.8
3	CEM+SEM+TSA	88.5	94.2
4	G+SEM+TSA	89.8	94.7
5	G+CEM+TSA	87.9	93.7
6	G+TSA	87.2	93.2

4. Conclusions

The proposed model is based on the dual characteristics of space and time of human motion skeleton data. It utilizes the spatial attention and sequence feature extraction capabilities of the Transformer to extract features and fuse them respectively from the two dimensions of space and time. This model fully exploits and utilizes the structural characteristics inherent in the data itself, achieving a high accuracy rate in action recognition at different scales and amplitudes, and improving the accuracy rate in long sequence action recognition. In the subsequent work, by extracting the chaotic invariants of the motion data, the originally high-dimensional and complex original motion data are refined into low-dimensional feature vectors. While retaining the key motion features, the computational complexity of the subsequent processing is also significantly reduced. Chaotic invariants, as a kind of characteristic parameters capable of describing complex dynamic characteristics, can meet the high complexity and diversity processing capabilities required for multi-task human motion behavior feature recognition.

References

[1] S. Edriss, C. Romagnoli, L. Caprioli, A. Zanela, E. Panichi, F. Campoli, E. Padua, G. Annino, and V. Bonaiuto, (2024) *"The role of emergent technologies in the dynamic and kinematic assessment of human movement in sport and clinical applications"* **Applied Sciences *14***(3): 1012. DOI: 10.3390/app14031012.

[2] R. Leib, I. S. Howard, M. Millard, and D. W. Franklin, (2024) *"Behavioral motor performance"* **Comprehensive Physiology *14***(1): 5179–5224. DOI: 10.1002/j.2040-4603.2024.tb00286.x.

[3] A. Jisi, S. Yin, et al., (2021) *"A new feature fusion network for student behavior recognition in education"* **Journal of Applied Science and Engineering *24***(2): 133–140. DOI: 10.6180/jase.202104_24(2).0002.

[4] J. Luo, W. Wang, and H. Qi, (2014) *"Spatio-temporal feature extraction and representation for RGB-D human action recognition"* **Pattern Recognition Letters *50***: 139–148. DOI: 10.1016/j.patrec.2014.03.024.

[5] M. Lovanshi and V. Tiwari, (2024) *"Human skeleton pose and spatio-temporal feature-based activity recognition using ST-GCN"* **Multimedia Tools and Applications *83***(5): 12705–12730. DOI: 10.1007/s11042-023-16001-9.

[6] Z. Wang, H. Lu, J. Jin, and K. Hu, (2022) *"Human action recognition based on improved two-stream convolution network"* **Applied Sciences *12***(12): 5784. DOI: 10.3390/app12125784.

[7] M. Xiao, (2024) *"The best angle correction of basketball shooting based on the fusion of time series features and dual CNN"* **Egyptian Informatics Journal *28***: 100579. DOI: 10.1016/j.eij.2024.100579.

[8] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, and Y.-D. Zhang, (2018) *"Multi-domain and multi-task learning for human action recognition"* **IEEE Transactions on Image Processing *28***(2): 853–867. DOI: 10.1109/TIP.2018.2872879.

[9] A. Laghari, H. He, A. Khan, R. Laghari, S. Yin, and J. Wang, (2022) *"Crowdsourcing platform for QoE evaluation for cloud multimedia services"* **Computer Science and Information Systems *19***(3): 1305. DOI: 10.2298/csis220322038l.

[10] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges. "A spatio-temporal transformer for 3d human motion prediction". In: *2021 International Conference on 3D Vision (3DV)*. IEEE. 2021, 565–574. DOI: 10.1109/3DV53792.2021.00066.

[11] K. Ding, A. J. Liang, B. Perozzi, T. Chen, R. Wang, L. Hong, E. H. Chi, H. Liu, and D. Z. Cheng. "HyperFormer: Learning expressive sparse feature representations via hypergraph transformer". In: *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*. 2023, 2062–2066. DOI: 10.1145/3539618.3591999.

[12] Y. Pang, Q. Ke, H. Rahmani, J. Bailey, and J. Liu. "Igformer: Interaction graph transformer for skeleton-based human interaction recognition". In: *European Conference on Computer Vision*. Springer. 2022, 605–622. DOI: 10.1007/978-3-031-19806-9_35.

[13] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, (2002) *"TAG: A tiny aggregation service for ad-hoc sensor networks"* **ACM SIGOPS Operating Systems Review *36***(SI): 131–146. DOI: 10.1145/844128.844142.

[14] H. Rao, S. Wang, X. Hu, M. Tan, Y. Guo, J. Cheng, X. Liu, and B. Hu, (2021) *"A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification"* **IEEE Transactions on Pattern Analysis and Machine Intelligence *44***(10): 6649–6666. DOI: 10.1109/TPAMI.2021.3092833.

[15] J. Jiang, J. Chen, and Y. Guo. "A dual-masked auto-encoder for robust motion capture with spatial-temporal skeletal token completion". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, 5123–5131. DOI: 10.1145/3503161.3547796.

[16] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani. "Infogcn: Representation learning for human skeleton-based action recognition". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 20186–20196. DOI: 10.1109/CVPR52688.2022.01955.

[17] S. Guan, H. Lu, L. Zhu, and G. Fang, (2022) *"AFE-CNN: 3D skeleton-based action recognition with action feature enhancement"* **Neurocomputing *514***: 256–267. DOI: 10.1016/j.neucom.2022.10.016.

[18] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, (2019) *"Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding"* **IEEE transactions on pattern analysis and machine intelligence *42***(10): 2684–2701. DOI: 10.1109/TPAMI.2019.2916873.

[19] X. Li, X. Zhang, L. Zhang, X. Chen, and P. Zhou, (2023) *"A transformer-based multi-task learning framework for myoelectric pattern recognition supporting muscle force estimation"* **IEEE Transactions on Neural Systems and Rehabilitation Engineering *31***: 3255–3264. DOI: 10.1109/TNSRE.2023.3298797.

[20] W. Xin, R. Liu, Y. Liu, Y. Chen, W. Yu, and Q. Miao, (2023) *"Transformer for skeleton-based action recognition: A review of recent advances"* **Neurocomputing *537***: 164–186. DOI: 10.1016/j.neucom.2023.03.001.

[21] C. Shi and S. Liu, (2024) *"Human action recognition with transformer based on convolutional features"* **Intelligent Decision Technologies *18***(2): 881–896. DOI: 10.3233/IDT-240159.

[22] Y. Xing, Z. Hu, X. Mo, P. Hang, S. Li, Y. Liu, Y. Zhao, and C. Lv, (2024) *"Driver steering behaviour modelling based on neuromuscular dynamics and multi-task time-series transformer"* **Automotive Innovation** *7*(1): 45–58. DOI: 10.1007/s42154-023-00272-x.

[23] C. Fan, S. Lin, B. Cheng, D. Xu, K. Wang, Y. Peng, and S. Kwong, (2024) *"EEG-TransMTL: A transformer-based multi-task learning network for thermal comfort evaluation of railway passenger from EEG"* **Information Sciences** *657*: 119908. DOI: 10.1016/j.ins.2023.119908.

[24] W. Li, N. Zhou, and X. Qu. "Enhancing eye-tracking performance through multi-task learning transformer". In: *International Conference on Human-Computer Interaction*. Springer. 2024, 31–46. DOI: 10.1007/978-3-031-61572-6_3.