Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

# Construction and optimization method of modern dance movement style feature classification model based on deep learning

Duoduo Wang[1], Dongsheng Shi[1,✉], Molin Li[2]

[1] *School of Art and Design, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China*

[2] *College of the Arts, Guangxi University, Nanning, Guangxi, 530004, China*

ABSTRACT

Human gesture estimation and action recognition are the current research hotspots in the field of computer vision. In this paper, we propose a dance pose estimation method based on multiple dilation convolution with hybrid attention and a dance action recognition algorithm based on spatio-temporal map convolution. In the pose estimation, the residual module is used to reduce the computational load of the network, while the LAM module is employed to fuse different dance features to improve the model's representation of effective features. The GAM module is then utilized to superimpose the global feature information, capturing richer joint point information. The graph attention mechanism is embedded in the action recognition algorithm to achieve better neighbor aggregation, followed by the construction of a new partitioning strategy to assign different weights to the limbs, which enhances the recognition ability of the model. The gesture estimation algorithm in this paper reduces the number of parameters by 36.73% compared with the Cross Former model, and is able to realize high accuracy reconstruction of modern dance movements such as jumping, lowering, kicking, and stretching, etc. The ST-GCN converges faster than the PA-LSTM, and the convergence trend is more stable. The recognition accuracies in ten modern dance movement style features are 90% and above, which shows that the design of the dance gesture estimation method and the dance movement recognition algorithm in this paper achieves the task of estimating and recognizing modern dance movement style features in real time.

*Keywords:* Attention mechanism, spatio-temporal graph convolutional network, dance gesture estimation, movement recognition

## 1. Introduction

Movement is the life and soul of modern dance art. The composition of modern dance language and the expressiveness of art are all completed by "movement". The "language of modern dance" is fundamentally the "language of movement". Therefore, the study of "movement" is a compulsory

✉ Corresponding author.
*E-mail address:* pskjsgzy@163.com (D. Shi).
Received 15 February 2024; Revised 25 March 2024; Accepted 15 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-239

course for learning and choreographing modern dance [1-2]. Different art forms have different material carriers and means of expression, and the material carriers and means of expression of modern dance are the human body and human movement. Modern dance movement is the movement of human body in space. The human body movements are used as a form of language to express the choreographer's profound experience and feelings about social life, characters' thoughts, emotions, environment and so on [3-4]. Dance movements generally come from three aspects, namely, life movements, traditional dance movements and creative movements. Life movements are movements based on life, traditional dance movements are movements that have been programmed, and creative dance movements are movements that are not limited to a certain style and programmed, but are innovative according to the choreographer's experience and knowledge. Choreographers can refine, edit and process the movements according to the above three kinds of movements to form the language material of modern dance [5-7].

Modern dance is perhaps usually the most difficult to understand in comparison with other arts, but the material medium he uses, human movement, is relatively closer to the surrounding environment and to life experience than any other art. This fact, however, does not make dance easy to understand, but actually does the opposite [8-9]. The movement of the human body is so taken for granted that we are fundamentally and unconsciously oblivious to its scope and potential, and that which is most familiar can also be the most forgettable. Movement is the body's most active medium of expression, the most perceptive of the world around it. Therefore, in the process of learning modern dance, it is important to become extremely aware of movement, and to fully express one's thoughts in terms of movement as perceived by the human body [10-11].

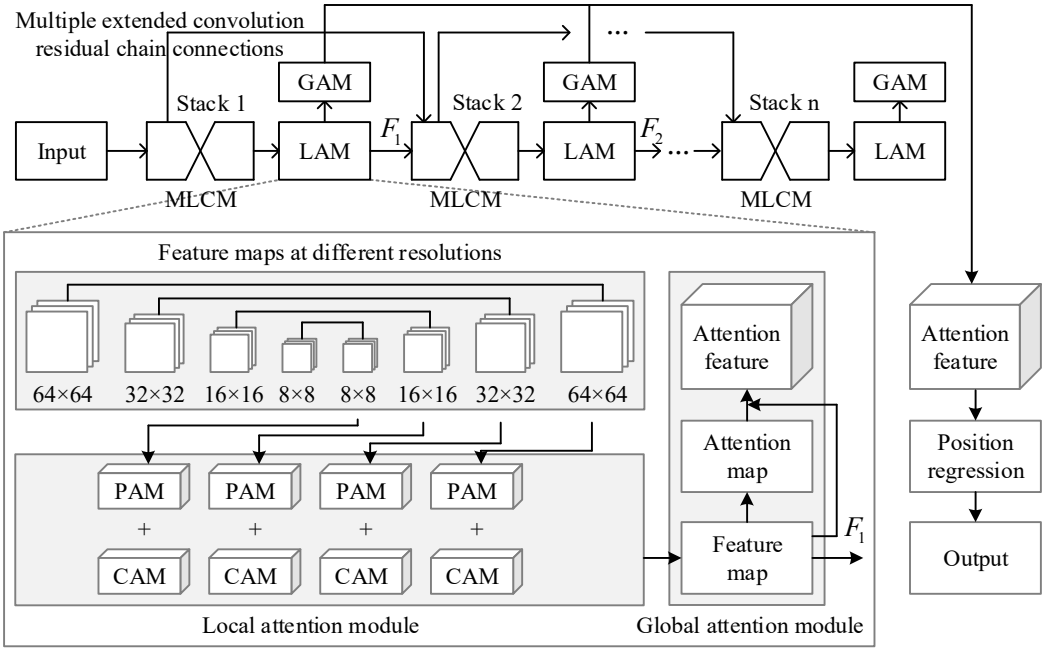
The style of dance has different concepts and various meanings, which can be used for a certain choreographer and his works, as well as for the dance works of a certain genre, a certain era or a certain nation. Therefore, in the theory of dance creation, the analysis of style generally contains the style of the work, the style of the choreographer, the style of the nation, the style of the era, the style of the genre and many other different meanings [12-13]. The most basic language of dance is human physical action, dance vocabulary is divided into classical, ballet, ethnic, modern and other different types, distinctive style categories, and each style of dance has its own artistic characteristics [14]. In terms of the artistic style of a specific dance work, it is mainly rooted in the content and form of the work, i.e., the characteristics of the dance movement, rhythm, rhyme, performance and many other aspects, which are the important signs of the aesthetic judgment of the dance art, and these artistic characteristics are the creative personalities of the choreographers or performers in the creation of the work, and for some excellent choreographers and performers, such distinctive personalities have become the unique mark of only belonging to them, that is to say, they have their own unique mark. For some excellent choreographers and performers, this distinctive personality becomes a mark of their own uniqueness, that is to say, they have formed their own artistic style in creation [15-18]. From Wu Xiaobang's condensed and subtle, to Dai Ailian's delicate and elegant, to Yang Liping's fresh and beautiful, all of them let us see the different style qualities. The formation and development of artistic styles of dance creation shows that the styles of artistic creation are diversified, and the basis of such diversification comes from the richness and colorfulness of social life. Only rooted in the fertile soil of life, the dance works will show a strong vitality, and will have a broad mass base [19-20].

In this paper, we design a coding network that incorporates multiple dilation convolution and hybrid attention for dance pose estimation. The network adopts an encoder-decoder structure, and the jump connection and residual module are added to the encoder structure to reduce the computational load of the network. In the next step, a dance movement recognition algorithm based on a spatio-temporal graph convolutional network is constructed, and a graph attention mechanism is introduced to enhance the performance of the model. Subsequently, a new partitioning strategy is proposed based on the existing partitioning strategy, taking into account the extraction of first-order neighbor relations and second-order neighbor relations as a way to increase the learning ability of the model.

Finally, the two designed algorithms are applied to the estimation and style recognition of actual modern dance movements, which provides scientific data for further research of the subsequent algorithms.

## 2. Multiple dilation convolution combined with hybrid attention for dance pose estimation

Aiming at the challenges of the stacked hourglass network structure, such as the problem of large number of parameters and high computational effort, and the difficulty of accurately capturing part of the joints' information due to occlusion in the process of dance gesture estimation, this chapter proposes a dance gesture estimation method based on multiple dilation convolution with hybrid attention. The method consists of a lightweight stacked hourglass module (MCLM) for multiple dilation convolution, a local attention module (LAM) [21], and a global attention module (GAM) [22], and the method is abbreviated as MLG, and the overall network structure of the MLG model is shown in Fig. 1.




MLCM: Lightweight hourglass module with multiple expansion convolution

PAM: location attention module

CAM: Channel attention module

**Fig. 1.** Human pose estimation network structure

### *2.1. Stacked Hourglass Networks and Attention Mechanisms*

Stacked Hourglass Network (SH-Net) acquires multi-scale information and fuses it by cascading multiple hourglass network structures, and this network has achieved better results in dance pose estimation.

The hourglass stacking network shows superior recognition performance in the dance pose estimation task, but because of the increased computational effort of the network, it is necessary to consider how to optimize the residual module in the stacked hourglass network.

In computer vision, attention remembering can be categorized into two types: soft attention mechanisms and strong attention mechanisms. The soft attention mechanism pays more attention to

the channel and region information of the target and realizes the task of allocating the attention weights to the two parts of the channel and region through the microscopicity in the neural network model. The strong attention mechanism pays more attention to the pixel points and focuses the attention on one of the regions through a process of random dynamic changes.

The SE attention module is an effective mechanism used to enhance the performance of the model by exploiting the dependencies between the modeled channels and adaptively calibrating the channel-based response features to substantially improve the quality of the representation.The working principle of the SE attention module can be divided into two main steps: compression and excitation.

In the SE attention module, the input $X \in R^{H' \times W' \times C'}$ features are mapped to $U \in R^{H \times W \times C}$ and compressed by Squeeze on $U$. This step generates a channel descriptor with a global receptive field, which not only contains global information, but also retains enough spatial information to finally obtain the statistic $z \in R^{C}$, $z$ for which the $c$ th element of the computational formula is shown in (1):

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{1}$$

In order to achieve full utilization of the information aggregated after the above operations, the relationships between the feature channels are further learned through two fully connected layers after compression for excitation processing. One of the fully-connected layers is used to significantly reduce the computational effort of the model by means of dimensionality reduction, while the other fully-connected layer restores the number of feature channels to their original size, allowing the network to learn the weights of each channel. Finally, an activation function is used to normalize these weights to between 0 and 1. The computational formula is shown in equation (2):

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{2}$$

where $\delta$ is the ReLU function, $W_1 \in R^{\frac{C}{r} \times C}, W_2 \in R^{C \times \frac{C}{r}}$.

Considering that the model complexity should not be too high, the original feature map is Scaled as shown in equation (3):

$$\tilde{x} = F_{Scale}(u_c, s_c) = s_c u_c \tag{3}$$

where $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_C]$; $F_{Scale}(u_c, s_c)$ denote the channel multiplication between scalar $s_c$ and feature mapping $u_c \in R^{H \times W}$.

### *2.2. Lightweight Stacked Hourglass Module with Multiple Expansion Convolution*

In order to ensure that the network parameters are small while improving the detection of the model, the method in this chapter introduces a lightweight stacked hourglass module (MCLM) based on multiple dilation convolution.

In the hourglass module proposed in this chapter, the encoder is directly connected to the decoder through a hopping connection, which eliminates the need to repeat the feature processing in the next stack, and can improve the encoder's ability to extract features.

The residual module is a very important part of the hourglass network, and the structure of the pre-activated residual block is different from that of the traditional residual block, which can efficiently construct the deep network and realize a significant increase in the training speed of the network.

The residual block based on multiple dilation convolution in the hourglass module structure proposed in this chapter solves the above problems well. Blindly realizing the expansion of the sensory field by increasing the size of the convolution kernel will lead to a large computational cost of the network, and this paper constructs a residual module using dilation convolution.

The number of parameters in the standard convolution is shown in equation (4):

$$\#\ param = K^2CM \tag{4}$$

where $K$ denotes the size of the core; $C$ denotes the size of the input channel; and $M$ denotes the size of the output channel. If the size of the input and output channels is the same as $H \times W$, the required computation is shown in equation (5):

$$Computational\ Cost = K^2CMHW \tag{5}$$

Since dilated convolution uses zero-padding within the kernel, this drastically reduces the computational complexity and computational cost.

### *2.3. Local attention module*

The method in this chapter adds a local attention module (LAM) to the features with different resolutions, which is divided into two main parts: the positional attention module enables the model to focus on spatial contextual information by attentively weighting the features in each channel, which facilitates the interconnection between similar features. The local attention module reintegrates the features extracted from the hourglass network at each level through two different attention module branches to obtain new attention features, which are then fused with the original features to enhance the characterization ability of the features. The up-sampling process in the hourglass network will generate different scale features, which will correspond to different feature mappings $f_r$, while their corresponding different attentional features $h_r^{an}$ will be obtained by the attention module.

First, the positional attention module processes the feature mapping $f_r$ to generate the local attention mapping $\phi_r^p$ and subsequently outputs the attention features $h_r^{pant}$ Then, the local attention features are passed through the channel attention module to obtain the attention mapping $\phi_r^c$ corresponding to each channel then matrix-products it with the previous local attention features to obtain the attention features $h_r^{cat}$ for each channel; finally, the attention features (i.e., $h_r^{cant}$) for each channel are integrated into the attention features $h_r^{an}$ as the output of each layer in the hourglass network. The computational formulas are shown in Eq. (6) to Eq. (8):

$$\phi(\lambda) = \frac{e^{s(\lambda)}}{\sum_{\lambda' \in Q} e^{s(\lambda')}} \tag{6}$$

$$h_r^{patt} = f_r * \sum_r \phi_r^p \tag{7}$$

$$h_r^{an} = \sum_c (h_r^{pan} * \phi_r^c) \tag{8}$$

where $Q = \{(x, y), \ell \mid x = 1, \cdots, W, y = 1, \cdots, H, \ell = 1, \cdots, C\}$ ; $\phi$ denote the attention mapping; $\sum_{\lambda_i = 1} \phi(\lambda) = 1$, $S(\lambda)$ denote the feature mapping of the channel.

The channel attention module can establish the correlation between the features of different parts of the human body, enhance the constraints on the posture of dance movements, and improve the accuracy of posture estimation.

### *2.4. Global attention module*

Global attention can filter and integrate different semantic information to further align features. Attention features generated by the attention modules in each layer of the stacked hourglass network will be integrated into the next layer, layer by layer, which not only enables the network to better deal with the global context, but also the deeper network layers facilitate the recovery of occluded information.

The stacked hourglass network receives images as data inputs, and each unit of the hourglass network outputs a feature map $f_i$. And for each feature map, the attention module is firstly utilized to convert it into an attention mapping $\phi_i$, and then the feature map and the attention mapping are combined to generate an attention feature $h_i^{an}$. Finally, the fusion of the feature mapping and the attention feature is used as inputs for the next hourglass network unit. The final feature map $g$ is obtained, which can be expressed as Eq. (9) and Eq. (10):

$$g = f_n + h_n^{an} \tag{9}$$

$$h_i^{an} = f_i * \phi_i \tag{10}$$

where $f_i$ denotes the feature map generated by the $i$ nd hourglass network, $i = 1, 2, \cdots, n$; $h_i^{an}$ denote the corresponding attentional features obtained by the attentional module.

## 3. Modern dance movement style recognition based on spatio-temporal graph convolution

### *3.1. Spatio-temporal map convolutional networks*

The main idea of ST-GCN [23] is to extract graph information in dance movements by alternating feature information in both spatial and temporal dimensions.

A movement sequence is represented using 2D or 3D coordinate position information of human dance movement points in a segment of consecutive frames. The main idea of ST-GCN is to extract feature information alternately in both spatial and temporal dimensions, using graph convolution fusion for action points in the image space of a single frame to obtain spatial topological structure information, and using one-dimensional temporal convolution to aggregate successive action information in the sequence dimension of consecutive multiple frames.

The input to the ST-GCN network is the action spatio-temporal map obtained from the action sequence, which is first normalized to the input and then fed into the ST-GCN unit for processing.

The convolution operation is abstracted into a nonlinear mapping from the input feature map to the output feature map, defined as shown in equation (11):

$$f_{out}(x) = \sum_{h=1}^{K} \sum_{w=1}^{K} f_{in}(p(x,h,\omega)) \cdot w(h,\omega) \tag{11}$$

3.1.1. Sampling function and weighting function. Subsequently the set of neighboring node information of node $v_{ti}$ is defined on the graph $B(v_{ti}) = \{v_{ti} \mid d(v_{ti}, v_{tj}) \le D\}$ and further on $B(v_{ti})$ the sampling function $p$ is defined. Where $d(v_{ti}, v_{tj})$ denotes the smallest distance in the set of paths from node $v_{ti}$ to node $v_{tj}$. In this chapter, the neighbors of the sampling function are defined as second order i.e. $D = 2$, from which the set of neighboring node information of node $v_{ti}$ is constructed $B(v_{ti})$ and the sampling function $p(v_{ti}, v_{tj})$ is defined as shown in equation (12):

$$p(v_{ti}, v_{tj}) = v_{tj} \tag{12}$$

The weight function $w$ is defined with reference to the 2D convolutional kernel, which operates on each node in the graph $l_{ti} : B(v_{ti}) \to \{0, \ldots . K-1\}$ to map the information set of its neighboring nodes, and at the same time introduces a partitioning strategy to divide $B(v_{ti})$ into multiple sub-regions, making the mapping operation simpler. The weight function is defined as shown in equation (13):

$$w(v_{ti}, v_{tj}) = w'(l_{ti}(v_{tj})) \tag{13}$$

3.1.2. Constructing spatial map convolutions. Considering the factor that different subregions contribute differently to the recognition task, this section introduces the normalization term $Z_{ti}(v_{tj})$

defined as $Z_{ti}(v_{tj}) = |\{v_{tk} \mid l_{ti}(v_{tk}) = l_{ti}(v_{tj})\}|$, and the new convolutional output is defined as shown in equation (14):

$$f_{out}(x) = \sum_{v_{ti} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}((v_{ti}, v_{tj})) \cdot w(v_{ti}, v_{tj}) \tag{14}$$

After redefining the $p$ and $w$ functions, the graph convolution operation on the space is obtained, as shown in equation (15):

$$f_{out}(x) = \sum_{v_{ti} \in B(v_{ti})} \frac{1}{z_{ti}(v_{tj})} f_{in}(v_{tj}) \cdot w'(l_{ti}(v_{tj})) \tag{15}$$

3.1.3. Constructing the Time Map Convolution. The input action spatio-temporal graph of ST-GCN contains not only action sequences in the spatial dimension, but also in the temporal dimension, so it is also necessary to define the graph convolution operation in time. The action sequences in the temporal dimension can be interpreted as a collection of actions of the same target within consecutive video frames, and this collection can be represented using the information changes of the same action points within consecutive video frames, i.e., connecting the same class of action points within neighboring video frames. Therefore, when constructing the action spatio-temporal graph, a sequence of N frames of video frames can be selected first, and the human body actions within the selected N frames can be connected.

### *3.2. Modern dance movement recognition algorithm based on spatio-temporal graph convolution*

The attention mechanism (AM) in neural networks mainly by is to mimic the way humans pay attention to the target of interest, enabling the network to also pay attention to the characteristics of the target.

The self-attention mechanism to video recognition is proposed to generalize the Non-local non-local network. A generalized Non-local operation in deep neural networks is defined in the Non-local network, and the specific operation expression is shown in equation (16):

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{16}$$

In the dance movement recognition task, previous approaches are generally adapted based on the framework of spatio-temporal graph convolution networks, which first extract the spatial information by raw graph convolution, then aggregate the temporal frame features by temporal convolution, and finally input the extracted high-level spatio-temporal features into the classifier for movement classification. Referring to the graph convolution model in 2s-AGCN, this chapter introduces an adaptive matrix to do element-level addition instead of the previous dot product. The specific adaptive graph convolution operation formula is shown in equation (17):

$$Y = \sum_{k=1}^{K} W_k X (A_k + B_k + C_k) \tag{17}$$

where $A_k$ is the original adjacency matrix based on the physical connections of the human body, and $B_k$ is a global attention matrix of $N \times N$, added to enable unconstrained learning of the model. $c_k$ is an adaptive matrix learned through data-driven self-attention, which helps capture the implicit relationships between different joints. The graph convolution described above is used in this chapter.

In the modern dance movement recognition task, a reasonable partitioning strategy can improve the model's ability to learn the movement information features, so the design of a more reasonable partitioning strategy can make the model's classification ability effectively improved. Here, the three original ST-GCN partitioning strategies are first introduced separately.

The main practice of the unlabeled configuration partitioning strategy is to divide the action points and their first-order neighbors into the same region, and set these action points to the same weight.

The main approach of the distance configuration partitioning strategy is also to divide the action points and their first order neighbors into the same region, but unlike the unlabeled configuration partitioning strategy that assigns the same weight, the distance configuration partitioning strategy categorizes the action points within the same region according to the distance and configures different weight parameters.

The main practice of the spatial configuration partitioning strategy is also to first divide the action points and their first-order neighbors into the same region, and also classify the action points within the same region by distance and configure different weight parameters, but in the distance configuration partitioning strategy, the region is centered on the action points, while in the spatial configuration partitioning strategy, it is centered on the center of gravity of the action.

Among them, the spatial configuration partitioning strategy is the most effective, which divides the adjacent region of the action point into three sub-regions, the first one is the root node itself, the second one is the action point whose distance to the center among the first-order neighboring action points is less than the distance from the root action point to the center, which is known as centripetal swarm, and the last one is the action point whose distance to the center among the first-order neighboring action points is greater than the distance from the root action point to the center, which is known as centrifugal swarm. The partitioning strategy approach is shown in equation (18):

$$l_{ti}(v_{ti}) = \begin{cases} 0 \, r_j = r_i \\ 1 \, r_j < r_i \\ 2 \, r_j > r_i \end{cases} \tag{18}$$

Where $r_j$ is the distance from action point $j$ to the center in the partitioned area, and $r_i$ is the distance from root action point $i$ to the center in the partitioned area.

In the dance action recognition task, a reasonable partitioning strategy can improve the model's ability to learn action information features. Therefore, a new partitioning strategy is adopted in this chapter to construct the ST-GCN model.

The new partitioning strategy can be shown in equation (19):

$$l_{ti}(v_{ti}) = \begin{cases} 0 & r_j = r_i \\ 1 & r_j < r_i \\ 2 & r_j < r_i, r_{k,j} = 1 \\ 3 & r_j > r_i \\ 4 & r_j > r_i, r_{k,j} = 1 \end{cases} \tag{19}$$

Where, $r_j$ is the distance from action point $j$ to the center in the divided area, $r_i$ is the distance from root action point $i$ to the center in the divided area. $r_{k,j}$ is the distance from action point $k$ to action point $j$.

The new partitioning strategy, similar to the spatial configuration partitioning strategy, sets the center of the region as the center of gravity of the action, and at the same time expands the criteria for dividing the region from first-order neighbors to second-order neighbors.

The new partitioning strategy proposed in this chapter not only takes into account the first-order neighbor relationship of action points, but also adds the extraction of the second-order neighbor relationship of action points. In this way, the sensory field is expanded, not only focusing on the local motion changes of the human body, but also extracting the relationship information between the localities through the second-order neighbor relationships.

By using the new partitioning strategy, the neighboring regions of the action points can be realized by the unit matrix $l$ of the root action point and the adjacency matrix $A$ of the action points in the neighboring regions, which are calculated as shown in Equation (20):

$$f_{out} = \sum_j \Lambda_j^{\frac{1}{2}} A_j \Lambda_j^{\frac{1}{2}} f_{in} W \tag{20}$$

where the adjacency matrix is split into several $A_j$ matrices as shown in equation (21).

$$A + I = \sum_j A_j \tag{21}$$

## 4. Experimental results of modern dance posture estimation

The challenging benchmark dataset Human3.6M was selected for evaluation, which is the largest 3D human Pose estimation dataset in common use and contains 3.6 million video frames captured by an accurate motion capture system, and 15 daily activities such as sitting, walking, and talking on the phone were performed by 7 professional subjects, and the same experimental setup was used based on previous work, with the selection of 5 of the dataset's subjects (S1, S5, S6, S7, S8) for training and 2 subjects S9 and S11 for testing.

Experimental validation was performed on the public dataset Human3.6M, using the most commonly used evaluation protocol: MPJPE in mm, and comparing it with some other representative methods in terms of mean Avg, parametric quantity Param, and arithmetic complexity FLOPs for the 15-action MPJPE, respectively, from the use of the real 2D keypoints (GT) and the tandem pyramid network (CPN) detected 2DPoses as inputs, the experimental results are shown in Tables 1 and 2, and this paper's method achieves 47.9 mm and 36.5 mm MPJPE, respectively. in terms of GT, the number of computational Parameters and FLOPs drop by 67.35% and 51.89%, respectively, and the model is much lighter compared to the MHF ormer model, even though there is no difference in the error . Compared with the latest model Pose Aug, the error decreased by 3.3mm and the positional accuracy was improved. In terms of CPN, the error of the improved model decreased by 5.3mm compared to the model Pose Aug. Compared to the latest Cross Former model using Transformer for Pose estimation, the computationally costly parameter count of the model decreased by about 36.73% and FLOPs decreased by about 62.22%, which speeds up the inference.

**Table 1.** Human 3.6M data set experimental results (GT)

| GT | Avg/mm | Param/M | FLOPs/G |
|---|---|---|---|
| Mohsen | 54.6 | 18 | / |
| Pose Aug | 39.8 | / | / |
| Pose Former | 34.6 | 9.8 | 1.35 |
| MHF ormer | 30.9 | 18.99 | 1.06 |
| This method | 36.5 | 6.2 | 0.51 |

**Table 2.** Human 3.6M data set experimental results (GPN)

| GPN | Avg/mm | Param/M | FLOPs/G |
|---|---|---|---|
| Pose Aug | 52.8 | 18 | / |
| Pavllo | 46.7 | / | / |
| Cross Former | 42.8 | 9.8 | 1.35 |
| MHF ormer | 43.2 | 18.99 | 1.06 |
| This method | 47.5 | 6.2 | 0.51 |

The MPJPE was compared with other representative methods in the field of Pose estimation on the Human3.6M dataset for eight actions, numbered 1-8 representing eight actions such as curling, stretching, rotating, jumping, twisting, throwing, lowering, and kicking, respectively. From the 2DPoses detected using real 2D keypoints (GT) and crosstalk pyramid network (CPN) as inputs, respectively, the experimental results are shown in Figs. 2 and 3, with smaller values of the horizontal coordinates representing better positional accuracy. The positional accuracy of this paper's dance pose estimation is ahead of other methods on all the eight mentioned maneuvers, especially in CPN, the error about the throwing maneuver is decreased by 31.82% compared with the model CNN, which indicates the positional accuracy of this paper's model is improved.

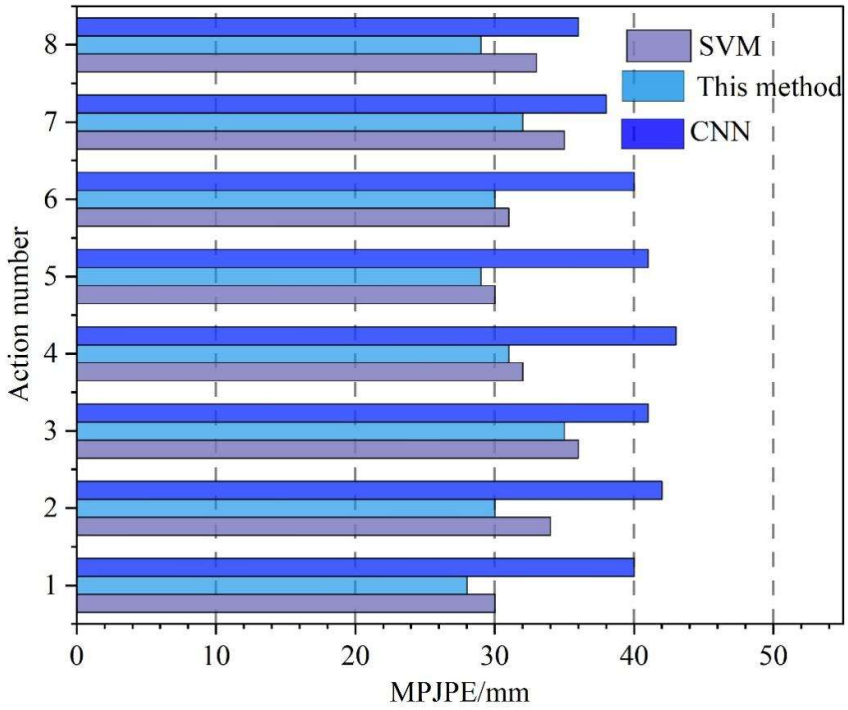



**Fig. 2.** Human 3.6M data set experimental results (GT)

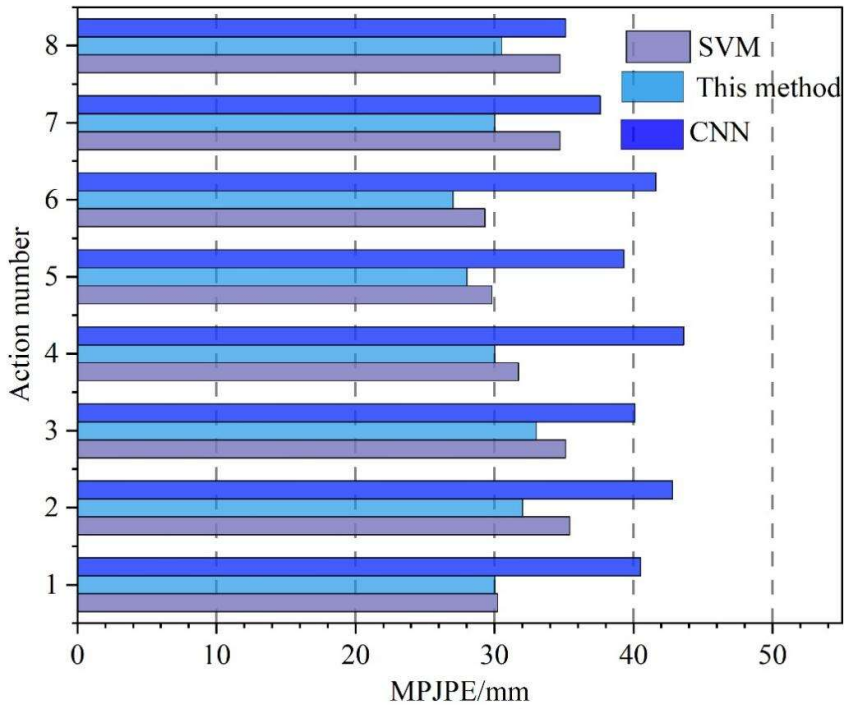
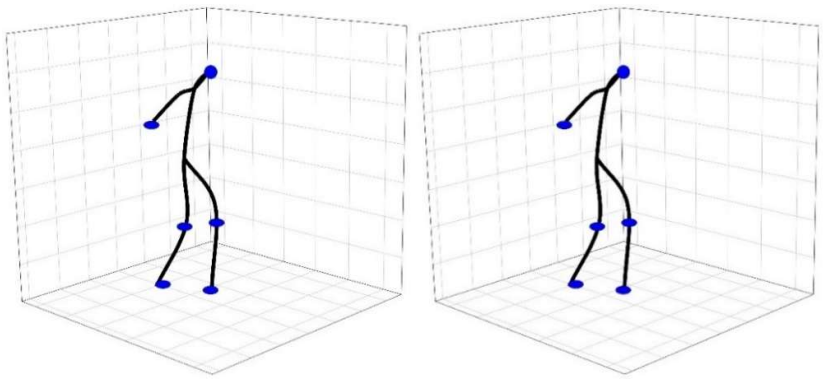


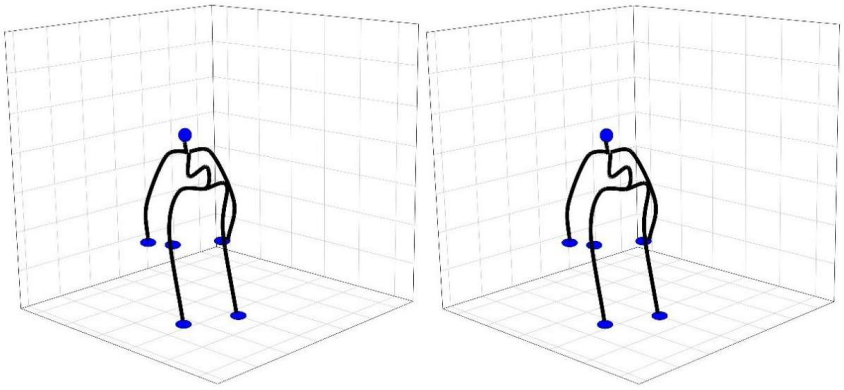
**Fig. 3.** Human 3.6M data set experimental results (GPN)

The test results of this paper's network on the Human3.6M test set were visualized by randomly selecting four modern dance movements, namely, jumping, lowering, kicking, and stretching, and extracting any frame images from them. The visualization results of the four movements are shown in Fig. 4-Fig. 7, respectively. (a) Figure is the real dance action, (b) Figure is the reconstruction result of this paper, the method in this paper can detect the key points of human body more accurately. There is a self-obscuring problem in the right arm part of the human body in Fig. 5(a) and Fig. 5(b), but the reconstructed movements of this paper's pose estimation method can also get better obtaining prediction results compared with the real Pose. It shows that the pose estimation method in this paper can reconstruct all parts of the human body with high accuracy when performing dance movements, and can effectively perform human Pose estimation.



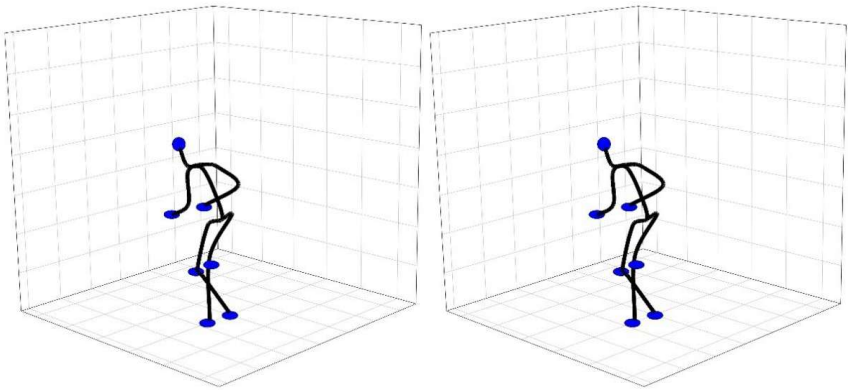
(a)True action (b)This method rebuilds the action

**Fig. 4.** Jumping action



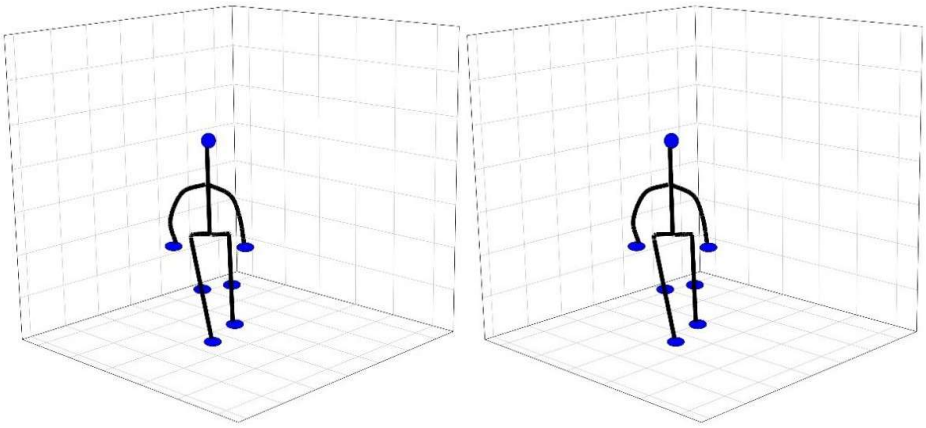
(a)True action (b)This method rebuilds the action

**Fig. 5.** Lower back motion



(a)True action (b)This method rebuilds the action

**Fig. 6.** Kicking



(a)True action (b)This method rebuilds the action

**Fig. 7.** Stretching

## 5. Experimental results of modern dance movement style recognition classification

The specific experimental environment in this chapter is shown in Table 3, in this study Ubuntu 18.04 64-bit operating system was used as the experimental environment, the deep learning framework was chosen as Pytorch, and CUDA was utilized to accelerate the training of the network, and a total of 100 rounds of iterative training was performed. Where the batch size was set to 32 and the initial learning rate was 0.1. The learning rate was reduced to 0.01 at the 30th round of training and again to 0.001 at the 60th round.

**Table 3.** Experimental environment

| Experimental environment | |
|---|---|
| Operating system | Ubuntu 18.04 LST,64-bit |
| CPU | Intel(R) Core(TM) i7-6700K CPU @ 4.00GHz |
| Memory size | 32G |
| GPU | NVIDIA TITAN RTX,display 24g |
| Software platform | Python3.6 |
| Depth learning framework | Pytorch 1.13.0 |
| Accelerating environment | CUDA 8.7.0 |

| optimizer | Adam |
|---|---|

In this paper, on the basis of spatio-temporal graph convolutional neural network, we optimize the network by adding graph attention mechanism to improve the recognition performance of spatio-temporal graph convolutional network. Fig. 8 shows the variation of loss rate of ST-GCN network and PA-LSTM network on NTU-RGB+D dataset. From the figure, it can be seen that ST-GCN converges faster and more stably than PA-LSTM, and can reach the optimal level by training with fewer epochs. It can be seen that by adding the graph attention mechanism, it effectively improves the action recognition effect and is more robust than the PA-LSTM network.

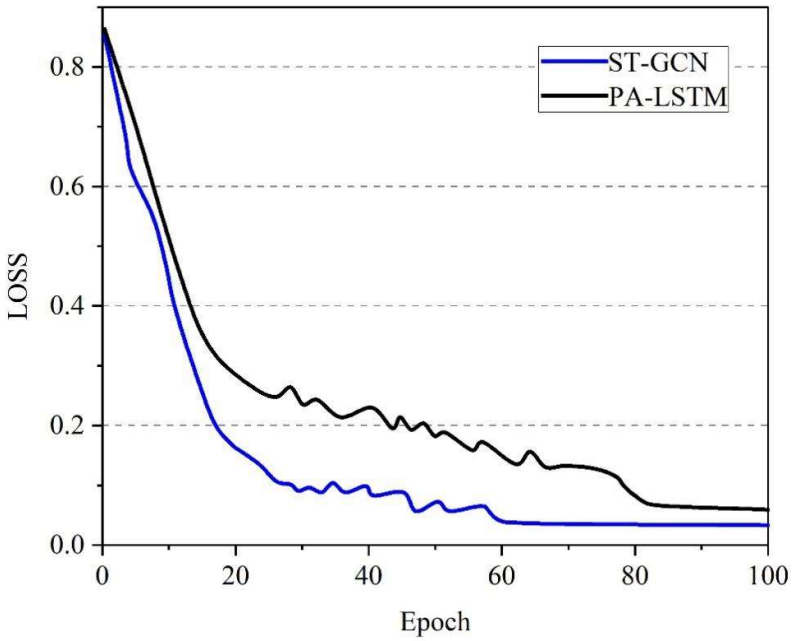



**Fig. 8.** Comparison results of loss rate

This experiment is to test the performance of this paper's spatio-temporal graph convolution-based dance movement recognition method on real-time recognition of modern dance movement style features on individual movement data streams. We use these test action data streams to simulate the real-input scene, and each data stream is composed of 10 types of basic dance movement styles and some illegal movements, with each type of movement containing 30. These 10 types of movement styles include free and flexible, abstraction, realism, simplicity and naturalness, creativity, recklessness and exuberance, implicit and introverted, skill diversity, softness and soothingness, and close to the body, which are numbered as movements 1-10.

Table 4 shows the detailed recognition results of this paper's method for recognizing modern dance movement data streams on various types of movement patterns. Recognition errors are mainly caused by substitution errors, and there are as many as three movements about substitution errors for simple and natural and technique diversity dance movement styles. This is due to the fact that there are also similarities between movements in the same category and thus they can be easily confused with each other. Deletion errors are mainly caused by differences between the movement data that make the recognition process susceptible to prematurely fulfilling the conditions presented in the chapters but not the conditions. Due to the existence of transition frames between two action patterns in the input data stream, these transition frames also have some similarity with the action patterns before and after them, so that some frames of them and the patterns before and after them are easily misrecognized as legitimate action patterns, i.e., resulting in insertion errors. However, on the whole, the style recognition accuracy of ten modern dance movements is 90% and above, and the overall recognition accuracy is high, which shows that the dance movement recognition algorithm based on

spatio-temporal graph convolution designed in this paper is effective, and the error range brought by the method is acceptable in practical applications.

**Table 4.** Identification results in various action patterns

| Action category | The total number of action patterns | Insert error number | Delete number | Substitution error number | The number of action patterns that are correctly identified | Detection rate/% |
|---|---|---|---|---|---|---|
| Move1 | 30 | 0 | 0 | 1 | 29 | 96.67 |
| Move2 | 30 | 1 | 1 | 1 | 27 | 90 |
| Move3 | 30 | 0 | 1 | 2 | 27 | 90 |
| Move4 | 30 | 0 | 0 | 3 | 27 | 90 |
| Move5 | 30 | 0 | 0 | 1 | 29 | 96.67 |
| Move6 | 30 | 0 | 0 | 2 | 28 | 93.33 |
| Move7 | 30 | 0 | 1 | 0 | 29 | 96.67 |
| Move8 | 30 | 0 | 0 | 3 | 27 | 90 |
| Move9 | 30 | 1 | 0 | 1 | 28 | 93.33 |
| Move10 | 30 | 0 | 0 | 2 | 28 | 93.33 |

Figure 9 shows the confusion matrix of the above modern dance movement recognition results, three of the simple and natural movements are recognized as soft and soothing movements, and three of the skillful and diverse movements are recognized as reckless and unrestrained movements, which indicates that the more similar movements in style are easily misrecognized, but the error rate is not too high, which suggests that the subsequent refinement and improvement of the spatio-temporal graph convolution-based dance movement recognition algorithm This suggests that the spatio-temporal graph convolution-based dance movement recognition algorithm can be improved in this area.

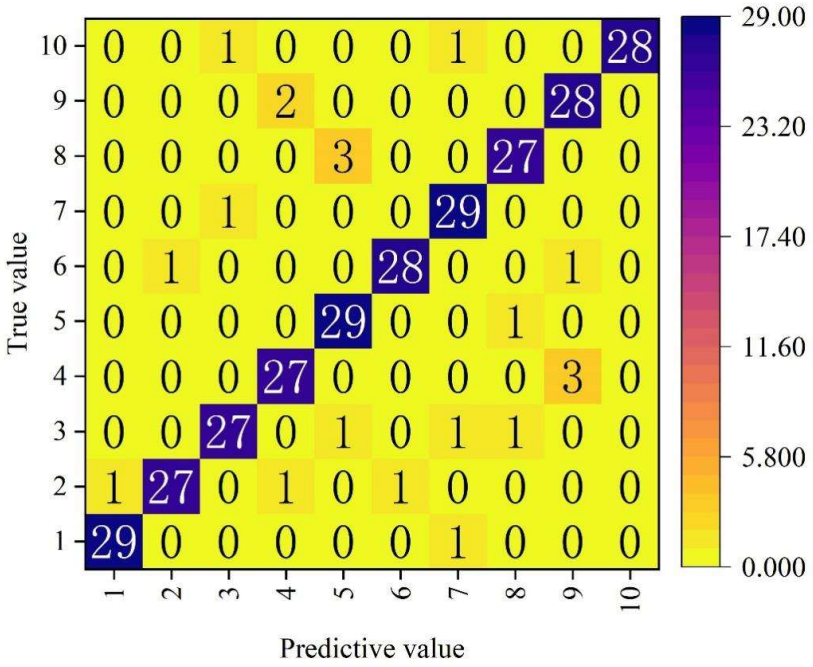



**Fig. 9.** Action identification confusion matrix

## 6. Conclusion

In this paper, we investigate the modern dance movement style classification and recognition algorithm based on deep learning, and propose a dance gesture estimation algorithm combining multiple dilation convolution and hybrid attention and a modern dance movement recognition algorithm based on spatio-temporal graph convolution.

Experimental results on the publicly available dataset Human3.6M show that the dance gesture estimation algorithm combining multiple dilation convolution and hybrid attention reduces the computational cost parametric number by 36.73% compared with the latest Cross Former model using Transformer for gesture estimation. In terms of CPN, the estimation error regarding throwing movements is reduced by 31.82% compared to the model CNN. The high accuracy reconstruction of the four modern dance movements, namely jumping, lowering, kicking, and stretching, illustrates that the dance pose estimation model in this paper can perform 3D pose estimation better.

In this paper, on the basis of spatio-temporal graph convolutional neural network, we optimize the network by adding the graph attention mechanism to improve the recognition performance of spatio-temporal graph convolutional network.ST-GCN converges faster than PA-LSTM and is more stable. It indicates that adding the graph attention mechanism to the spatio-temporal graph convolutional neural network effectively optimizes the action recognition effect of the network.

The recognition accuracy of this paper's recognition algorithm is 90% and above in different modern dance movement style features recognition, which shows that the spatio-temporal graph convolutional de dance movement recognition algorithm designed in this paper is effective and can achieve a higher detection rate.

## References

[1] Snowber, C. (2018). Living, moving, and dancing. Handbook of arts-based research, 247-262.

[2] Gotman, K. (2017). Choreomania: Dance and disorder. Oxford University Press.

[3] Xiaochen, Q., & Nakok, M. (2023). The Development of Teaching of Modern Dance and Choreography Techniques in China. Journal of Modern Learning Development, 8(9), 350-361.

[4] Pakes, A. (2020). Choreography invisible: the disappearing work of dance. Oxford University Press.

[5] Aristidou, A., Zeng, Q., Stavrakis, E., Yin, K., Cohen-Or, D., Chrysanthou, Y., & Chen, B. (2017, July). Emotion control of unstructured dance movements. In Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation (pp. 1-10).

[6] Zafeiroudi, A. (2023). Analyzing and discussing the evolution of arabesque movement according to dance elements and aesthetics. Acade. J. Interdisc. Stud, 12, 41.

[7] Puleio, D. (2017). Mid-Twentieth Century Modern Dance in the Twenty-first Century. University of California, Irvine.

[8] Kwan, S. (2017). When is contemporary dance?. Dance Research Journal, 49(3), 38-52.

[9] Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2019). Everybody dance now. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5933-5942).

[10] Haas, J. G. (2024). Dance anatomy. Human kinetics.

[11] Kassing, G. (2024). Discovering dance. Human Kinetics.

[12] Christensen, J. F., Cela‐Conde, C. J., & Gomila, A. (2017). Not all about sex: neural and biobehavioral functions of human dance. Annals of the New York Academy of Sciences, 1400(1), 8-32.

[13] Duncan, I. (2019). MODERNIST DANCE, WAR, AND MODERNITY. Dance, Modernism, and Modernity.

[14] Wang, X., & Zhang, C. (2018, November). Connotation interpretation on modern value of national dance. In 2018 5th International Conference on Education, Management, Arts, Economics and Social Science (ICEMAESS 2018) (pp. 1145-1149). Atlantis Press.

[15] Demian, N. (2022). Reframing of contemporary dance. Învăţământ, Cercetare, Creaţie, 8(1), 65-70.

[16] Yin, W., Yin, H., Baraka, K., Kragic, D., & Björkman, M. (2023). Multimodal dance style transfer. Machine Vision and Applications, 34(4), 48.

[17] Dickinson, E. R. (2017). Dancing in the Blood: Modern Dance and European Culture on the Eve of the First World War. Cambridge University Press.

[18] Lau, R., & Krause, B. (2022). Preferences for perceived attractiveness in modern dance. Journal of Cultural Economics, 46(3), 483-517.

[19] Kempe, M., & Heinen, T. (2022). Aesthetic perception of stage setups in dance. European Journal of Sport Sciences, 1(4), 29-35.

[20] Abra, A. (2017). Who makes new dances?: The dance profession and the evolution of style. In Dancing in the English style (pp. 44-76). Manchester University Press.

[21] Xie Yinghua,Zhou Yuntong,Wang Chen,Ma Yanshan & Yang Ming. (2023). Multi-scale feature fusion network with local attention for lung segmentation. Signal Processing: Image Communication.

[22] Rui Qian & Yong Ding. (2024). An Efficient UAV Image Object Detection Algorithm Based on Global Attention and Multi-Scale Feature Fusion. Electronics(20),3989-3989.

[23] Qi Lu. (2024). Sports-ACtrans Net: research on multimodal robotic sports action recognition driven via ST-GCN. Frontiers in Neurorobotics1443432-1443432.