



OPEN ACCESS

EDITED BY
Xianmin Wang,
Guangzhou University, China

REVIEWED BY
Arnab Dey,
Indian Institute of Engineering Science and
Technology, Shibpur, India
Yapeng Gao,
Taiyuan University of Technology, China
Zhiqiang Feng,
Lanzhou University of Technology, China

*CORRESPONDENCE
Yan Tong
✉ malibo2588@163.com

RECEIVED 03 June 2024
ACCEPTED 22 July 2024
PUBLISHED 21 October 2024

CITATION
Ma L and Tong Y (2024) TL-CStrans Net: a
vision robot for table tennis player action
recognition driven via CS-Transformer.
Front. Neurobot. 18:1443177.
doi: 10.3389/fnbot.2024.1443177

COPYRIGHT
© 2024 Ma and Tong. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC
BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

TL-CStrans Net: a vision robot for table tennis player action recognition driven via CS-Transformer

Libo Ma¹ and Yan Tong^{2*}

¹Guangdong Polytechnic of Environmental Protection Engineering, Foshan, China, ²Hunan Labor and Human Resources Vocational College, Changsha, China

Currently, the application of robotics technology in sports training and competitions is rapidly increasing. Traditional methods mainly rely on image or video data, neglecting the effective utilization of textual information. To address this issue, we propose: TL-CStrans Net: A vision robot for table tennis player action recognition driven via CS-Transformer. This is a multimodal approach that combines CS-Transformer, CLIP, and transfer learning techniques to effectively integrate visual and textual information. Firstly, we employ the CS-Transformer model as the neural computing backbone. By utilizing the CS-Transformer, we can effectively process visual information extracted from table tennis game scenes, enabling accurate stroke recognition. Then, we introduce the CLIP model, which combines computer vision and natural language processing. CLIP allows us to jointly learn representations of images and text, thereby aligning the visual and textual modalities. Finally, to reduce training and computational requirements, we leverage pre-trained CS-Transformer and CLIP models through transfer learning, which have already acquired knowledge from relevant domains, and apply them to table tennis stroke recognition tasks. Experimental results demonstrate the outstanding performance of TL-CStrans Net in table tennis stroke recognition. Our research is of significant importance in promoting the application of multimodal robotics technology in the field of sports and bridging the gap between neural computing, computer vision, and neuroscience.

KEYWORDS

neural computing, computer vision, neuroscience, multi-modal robot, table tennis stroke recognition

1 Introduction

Table tennis is a highly technical and fast-paced sport, and the recognition of players' movements during matches can not only help coaches and athletes analyze technique and improve training effectiveness but also provide a better viewing experience for spectators (Kim et al., 2021). Additionally, the motion recognition technology for table tennis players can be applied to automated scoring systems, enhancing the fairness and accuracy of the game (Munro and Damen, 2020). With the rapid development of artificial intelligence technologies, the precise recognition of table tennis players' movements using advanced methods such as deep learning has become possible. This not only promotes the integration of sports and technology but also provides insights and references for motion recognition in other sports (Van Amsterdam et al., 2022). Therefore, researching motion recognition technology for table tennis players holds significant practical significance and wide-ranging application prospects.

The recognition of table tennis player's movements traditionally depends on symbolic AI and knowledge representation. Firstly, expert systems simulate the decision-making process of human experts by encoding their knowledge, and they offer strong interpretability, providing clear justifications for each recognition result. For instance, Wang (2022) proposed a knowledge-based expert system for table tennis motion analysis. Furthermore, Kulkarni and Shenoy (2021) conducted a comprehensive review showcasing various applications and developments of expert systems in the field of sports. Secondly, rule-based methods employ a set of predefined rules for motion recognition. These methods demonstrate high determinism and reliability, performing well even in the face of complex or varied motions. Carvajal and Garcia-Colon (2003) presented a rule-based system for automated sports analysis, while Huang et al. (2023) provided a rule-based framework for analyzing sports performance. In addition, logistic regression, as a statistical method, performs classification decisions by learning features from training data. It not only finds significant applications in motion recognition but also improves classification accuracy. Yen et al. (2023) demonstrated the application of logistic regression in sports motion recognition, and Martin et al. (2018) further investigated its use in motion analysis in sports science to enhance recognition accuracy. While these approaches have the benefit of being easily understood and having clear decision-making processes, they are limited in their ability to handle diverse and intricate movements, as well as their capacity to process extensive amounts of data.

In order to overcome the shortcomings of traditional algorithms in dealing with intricate and diverse movements, as well as their restricted capacity to handle extensive amounts of data, motion recognition algorithms for table tennis players now heavily depend on adaptive learning and big data analysis. Firstly, decision tree-based methods construct classification models by recursively partitioning the dataset. These methods offer advantages of ease of understanding and implementation. For example, Wang et al. (2023) effectively recognized table tennis motions using a decision tree model in their study. Sha et al. (2021) also demonstrated the application of decision trees in motion recognition. Secondly, random forest-based methods improve classification accuracy and robustness by constructing multiple decision trees, exhibiting high precision and resistance to overfitting. Tabrizi et al. (2020) showcased the excellent performance of the random forest model in table tennis motion recognition in their research. Additionally, multilayer perceptrons (MLPs), as deep learning models, extract high-level features through multiple hidden layers and possess strong nonlinear mapping capabilities. Song et al. (2024) study demonstrated that MLPs outperformed traditional methods in motion recognition, and Zhang (2017) presented similar results. Nevertheless, these techniques rely heavily on extensive annotated data and require intricate training procedures.

To overcome the limitations of statistical and machine learning algorithms in handling complex temporal data and high-dimensional features, deep learning-based algorithms in motion recognition for table tennis players primarily rely on large-scale neural network models and multimodal data fusion. This approach offers stronger feature extraction capabilities and the advantage of end-to-end learning. Firstly, convolutional neural

networks (CNNs) extract spatial features through multiple layers of convolutional operations and have been widely applied in action recognition for image and video data. For example, Martin et al. (2021) successfully recognized table tennis motions using a CNN model in their research. Yen et al. (2023) demonstrated the efficiency of CNNs in motion recognition, and Wang et al.: This study presents a multimodal audio-visual robot using 3D CNN and CRNN for player behavior recognition and prediction in basketball matches. It demonstrates significant advancements in integrating auditory and visual cues to enhance action recognition systems in dynamic sports environments (Wang, 2024). Secondly, long short-term memory networks (LSTMs) address the issue of long-term dependencies in sequence data by introducing memory cells, making them suitable for handling time series data. Dey et al. (2024b) showcased the application of LSTMs in table tennis motion recognition, Lim et al. (2018) improved recognition accuracy using LSTM models, and Hu et al. (2023) demonstrated the superiority of LSTMs in handling temporal data. Dey et al.: This research introduces an attention-based DC-GRU network specifically designed for recognizing umpire signals in cricket matches. It highlights the use of attention mechanisms to improve the temporal feature extraction crucial for accurately interpreting referees' signals in sports (Dey et al., 2024a). Additionally, Transformer models process sequence data using self-attention mechanisms, offering advantages such as parallel computation and capturing global dependencies. Bian et al. (2024) achieved high-precision action recognition using Transformer models in their research, Wang and Tan (2023) showcased the advantages of Transformers in multimodal data fusion, and Yenduri et al. (2024) demonstrated the powerful performance of this model. However, these methods require substantial computational resources and involve high model training complexity.

Dey et al. (2024): This research introduces an attention-based DC-GRU network specifically designed for recognizing umpire signals in cricket matches. It highlights the use of attention mechanisms to improve the temporal feature extraction crucial for accurately interpreting referees' signals in sports.

However, traditional methods for table tennis stroke recognition primarily rely on image or video data, lacking the utilization of textual information. Moreover, these traditional methods exhibit limitations in handling complex and diverse actions, as well as in processing large-scale data. Therefore, this paper aims to propose a multimodal table tennis stroke recognition method that integrates image and textual information: TL-CStans Net, a visual robotic system for table tennis player action recognition driven by CS-Transformer. This method combines Swin Transformer, CLIP, and transfer learning. Firstly, Swin Transformer is employed to process the visual information of table tennis matches. Swin Transformer divides the image into smaller patches and applies self-attention mechanisms to capture global and local dependencies, thereby extracting key visual features from the image. Next, we utilize the CLIP model for joint representation learning of images and text. The CLIP model performs contrastive learning on a large number of image-text pairs, aligning images and text in both visual and semantic spaces. This enables the machine to understand the visual and semantic significance of table tennis stroke actions and facilitate comprehensive integration.

Finally, through transfer learning, we apply the pretrained Swin Transformer and CLIP models to the task of table tennis stroke recognition. By fine-tuning on annotated table tennis stroke action datasets, the model can adapt to specific tasks and achieve accurate stroke recognition.

- By employing the self-attention mechanism of Swin Transformer, the model can capture global and local dependencies, thereby improving the accuracy of action recognition.
- By combining the CLIP model's large-scale image-text contrastive learning, it can effectively utilize large-scale data and enhance the model's generalization ability.
- Through the multimodal learning of the CLIP model, it integrates both image and text information, enabling a comprehensive understanding and recognition of table tennis strokes.

2 Related work

2.1 Action recognition

Several recent studies have contributed significantly to the field of action recognition, showcasing diverse approaches and innovations. [Hu et al. \(2023\)](#) proposed a 3D network with channel excitation and knowledge distillation, demonstrating robust performance in action recognition tasks. Their method leverages advanced techniques in neural network architecture to enhance feature extraction and classification accuracy. [Dey et al. \(2024\)](#) introduced an attention-driven residual DC-GRU network specifically designed for workout action recognition in video streams. Their approach integrates attention mechanisms and recurrent neural networks to capture temporal dependencies effectively, achieving notable results in recognizing complex workout actions ([Dey et al., 2024b](#)). [Wang et al. \(2023\)](#) developed a spatiotemporal and motion information extraction network tailored for action recognition tasks. Their model focuses on extracting comprehensive spatiotemporal features from video data, employing innovative techniques to improve action classification performance across various datasets ([Wang et al., 2023](#)). These studies collectively contribute to advancing the state-of-the-art in action recognition by exploring novel architectures, integrating sophisticated features, and optimizing model capabilities for specific application domains. Incorporating insights from these works enriches our understanding of current methodologies and opens avenues for further research in enhancing the robustness and applicability of action recognition systems.

2.2 Transformers model

Since its introduction in 2017, the Transformer model has become a cornerstone of several fields, particularly in Natural Language Processing (NLP), where it has made revolutionary progress. Its core mechanism, self-attention, allows the model to capture global dependencies at any position in the input sequence, which is crucial for understanding and generating natural language.

Transformer-based models such as BERT and GPT have set new standards in language understanding and generation tasks ([Qiu et al., 2023](#)). Furthermore, the impact of Transformers has extended to the field of computer vision. The Vision Transformer (ViT) demonstrates comparable or superior performance to traditional CNN models in tasks such as image classification, object detection, and semantic segmentation by treating images as sequences of smaller patches and processing them similarly to text sequences ([Zhang et al., 2023](#)). In the field of speech processing, such as speech recognition and speech synthesis, Transformers have also shown their superiority. For example, Transformer models can more accurately model the long-range dependencies of speech signals, thereby improving the accuracy of speech recognition. Additionally, Transformers have been applied in music generation and multimodal learning, such as video understanding and image-text pairing, showcasing their broad applicability and powerful capabilities ([Liu et al., 2023](#)).

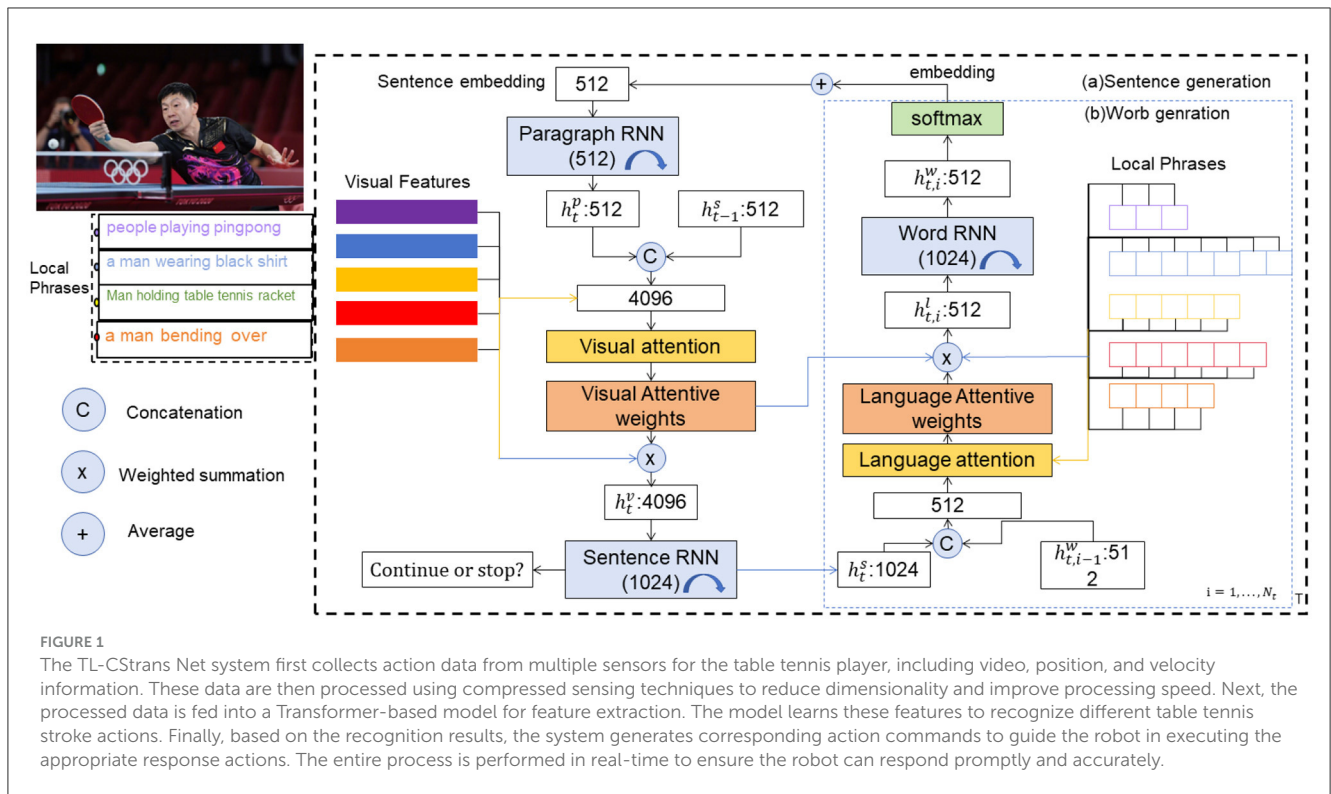
2.3 Robot vision

Robot vision, as a key technology for achieving robot autonomy, plays a vital role in robot navigation, object recognition, and manipulation tasks. Deep learning, particularly Convolutional Neural Networks (CNNs), has become the foundation of robot vision research, enabling robots to perform effective visual perception in complex environments ([Song et al., 2021](#)). In recent years, the research focus has gradually shifted toward enabling robots to exhibit higher adaptability and flexibility in more complex and uncertain environments. For example, through Deep Reinforcement Learning (DRL), robots can learn how to perform specific tasks through real-time interaction with the environment. Additionally, research on biomimetic vision systems, which aim to mimic human or animal visual systems, is advancing to enhance the visual processing capabilities of robot systems ([Xue et al., 2020](#)). Robot vision also includes multi-sensor fusion techniques, such as combining visual data with sensors like radar and LiDAR, to achieve more precise and robust environmental perception. Moreover, machine learning methods, particularly transfer learning and semi-supervised learning, are used to optimize robot vision systems by leveraging limited labeled data and abundant unlabeled data. The above overview provides detailed background and cutting-edge perspectives on the latest research and technological advancements in action recognition, the applications of Transformers, and the field of robot vision ([Psaltis et al., 2022](#)).

3 Methodology

3.1 Overview of our network

The article proposes a new visual robot system called TL-CStrans Net, which utilizes CS-Transformer (Compressed Sensing Transformer) for action recognition in a table tennis player. The article briefly explains the principles of these methods and provides a detailed description of the implementation process of TL-CStrans Net. TL-CStrans Net utilizes compressed sensing techniques to



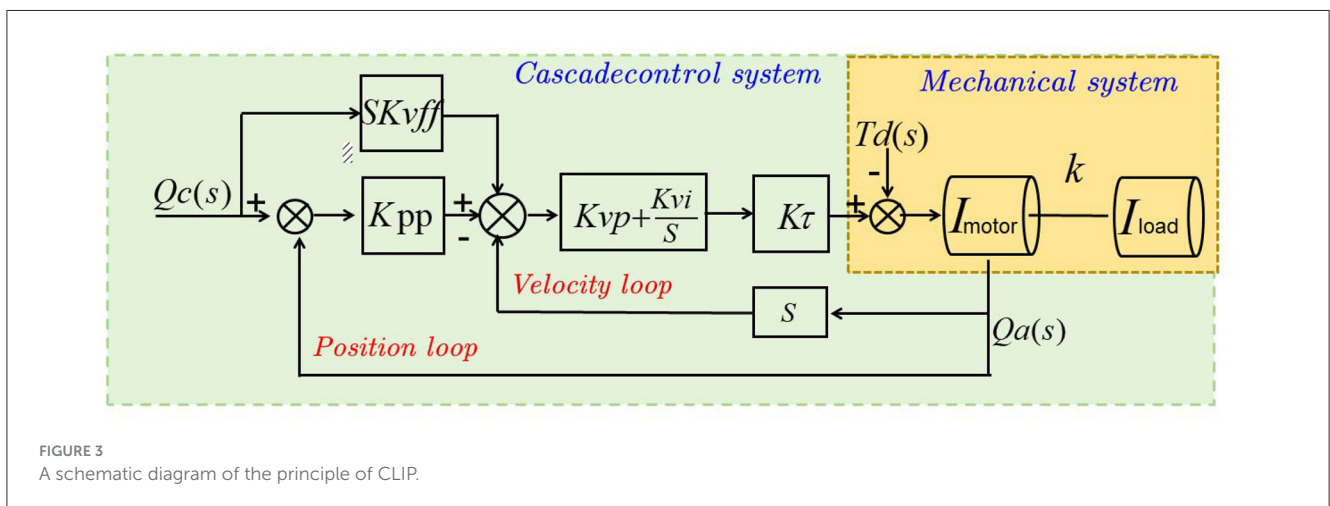
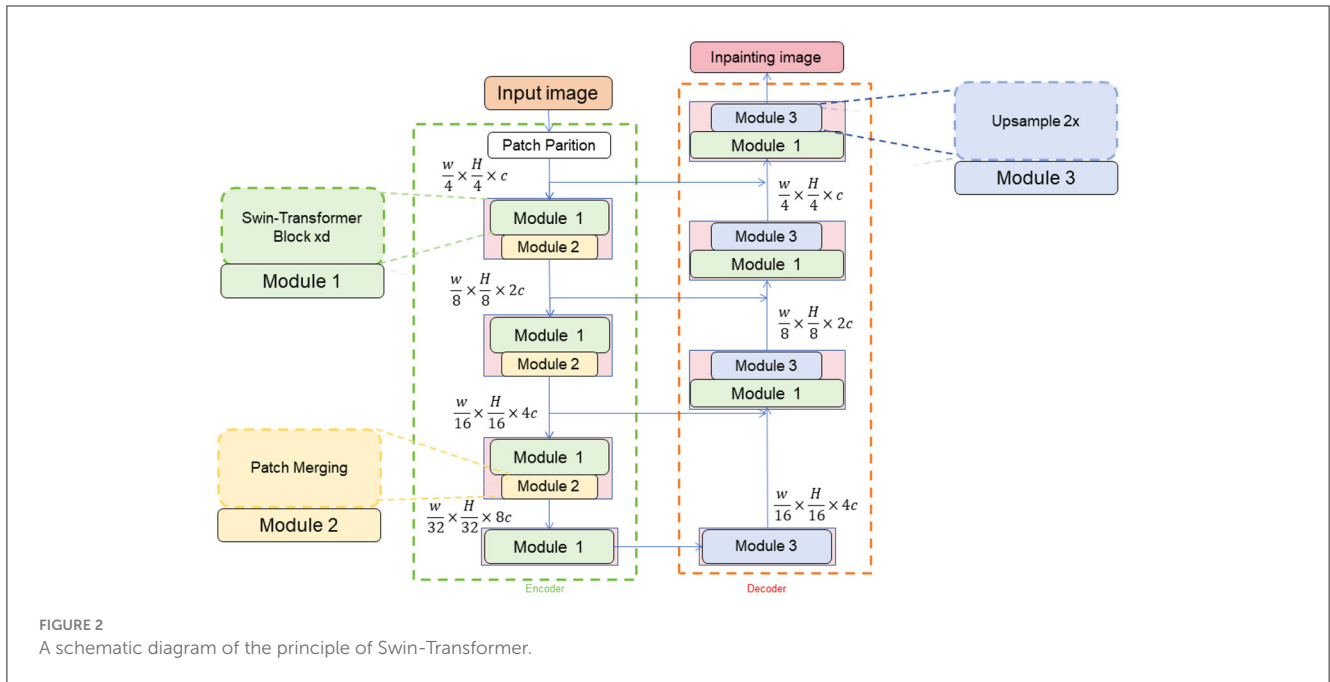
optimize the processing efficiency of input data and captures and analyzes the spatiotemporal features of table tennis actions using the Transformer architecture. This combined approach not only improves the accuracy of action recognition but also enhances the computational speed during action generation, making it more efficient for real-time applications. Additionally, the system integrates deep learning and reinforcement learning techniques and enhances the robot’s understanding of complex actions through multimodal information fusion. This comprehensive fusion of technologies provides an innovative perspective and implementation path for table tennis robot research. **Figure 1** shows the overall framework diagram of the proposed method.

The system starts with an input image, like a person playing table tennis, accompanied by descriptive local phrases such as “people playing ping pong” and “man holding a baseball bat”. These inputs undergo feature extraction where spatial features are derived using a convolutional network and processed through a Sentence RNN to create sentence embeddings, while local phrases are individually encoded to provide contextual image details. Attention mechanisms are applied, with visual attentive weights enhancing relevant image features and language attentive weights focusing on key words or phrases. These mechanisms feed into a Paragraph RNN, which integrates sentence embeddings with previous states to maintain paragraph flow, and a Word RNN that generates specific words, crafting sentences influenced by both visual and language attention. The system iteratively produces sentences and checks after each whether to continue, ensuring the output is not only relevant and contextually appropriate but also concise. This detailed breakdown highlights how the model integrates visual cues with textual descriptions through sophisticated attention mechanisms to produce accurate and

context-aware textual descriptions of the scenes depicted in the images.

In this paper, we initially utilized an image captioning model to generate preliminary text descriptions for each frame of the ping pong stroke sequences. This model provided a broad context by describing the visible actions and settings automatically. Following the automated captioning, we conducted a manual annotation process where domain experts refined and verified the captions. This step was crucial to ensure that the specific actions related to ping pong strokes were accurately described, catering to the unique needs of our task in recognizing ping pong stroke actions in robotic players. This combined approach of using automated image captioning followed by meticulous manual review allowed us to create a robust dataset that is both accurate in its action descriptions and scalable in its annotation process.

By using reinforcement learning, we are able to generate optimized hitting actions for the robot to adapt to different ball speeds and angles. Specifically, we employ deep reinforcement learning algorithms such as Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO), which gradually improve the robot’s hitting strategies through iterative experimentation and feedback. The process of reinforcement learning includes the following steps: Environment modeling: Firstly, we establish a simulated table tennis environment that includes the table, paddle, and ball with their physical properties. This environment is used to train the reinforcement learning model. Reward design: We design a reward function to evaluate the effectiveness of each hitting action. The reward function takes into account factors such as hitting accuracy, speed, and the opponent’s response, encouraging the robot to learn the optimal hitting strategy. Policy optimization: Through reinforcement learning algorithms, the robot updates



its hitting policy through iterative experimentation to maximize cumulative rewards.

Action recognition is a crucial task in table tennis robots. Based on deep learning methods, Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) can handle image or text data and extract key features to recognize actions. By training on a large amount of annotated data, the model can learn to identify patterns and rules of table tennis strokes. Action generation optimizes the robot's action generation process through reinforcement learning methods, evaluating action quality using reward functions. Deep reinforcement learning algorithms gradually optimize the generated actions, enabling the robot to generate adaptive table tennis strokes based on the current environment and state. Multimodal information fusion methods combine various sources of information such as images, text, audio, and inertial sensors to improve the accuracy and robustness of action recognition. Multimodal neural networks or attention

mechanisms effectively integrate information from different modalities. Real-time optimization enhances the robot's real-time performance through model compression, hardware acceleration, parallel computing, and other methods. This allows the robot to respond quickly and interact with human players. Reinforcement learning methods further optimize the robot's decision-making and action generation processes in table tennis matches.

For data collection and preprocessing, we assembled and annotated a comprehensive dataset of table tennis strokes, incorporating images, text, audio, and sensor data, and prepared this data through methods such as cropping, scaling, encoding, and vectorizing. In model training, we utilized advanced deep learning techniques like convolutional and recurrent neural networks for action recognition, and reinforcement learning methods such as Deep Q-Networks and policy gradient techniques for action generation, enhancing the robot's ability to adaptively generate table tennis strokes. Furthermore, we fused multimodal data

using neural networks or attention mechanisms to align diverse information types, boosting the accuracy of action recognition and generation. Real-time performance was augmented through model compression, hardware acceleration, and parallel computing, along with reinforcement learning to optimize the robot's decision-making and actions in real-time table tennis scenarios. Finally, we rigorously tested the models using a designated test dataset to evaluate their accuracy, realism, and real-time capabilities against actual match conditions, ensuring optimal real-world performance.

3.2 Swin-Transformer

The Swin-Transformer model is a deep learning model based on attention mechanisms used for image recognition tasks (Xiao et al., 2023). It introduces both local and global attention mechanisms within the image to better capture details and contextual information. In the method for table tennis robot action recognition and generation, the Swin-Transformer model is employed for the task of recognizing table tennis strokes, providing accurate classification of the strokes (Xiao et al., 2022). Figure 2 is a schematic diagram of the principle of Swin-transformer.

To elaborate on how temporal features are extracted in our paper, we utilize frame differencing and a temporal attention module. Frame differencing detects changes between consecutive frames, highlighting areas of motion by comparing pixels and applying a threshold to emphasize significant changes. This method captures 'temporal features' that encapsulate the dynamics of motion, essential for understanding action sequences. The temporal attention module enhances recognition by assigning weights to key frames identified through their significant motion, allowing the model to focus on frames that contribute most to understanding the action. This weighted aggregation of features ensures both temporal detail and spatial integrity, enhancing the model's ability to accurately recognize and classify actions. In our revision, we include specific examples demonstrating how these techniques improve action recognition accuracy, such as frame differencing revealing subtle movements in complex actions and the attention module prioritizing crucial frames in fast-paced scenarios. These examples underscore the robustness and practical applicability of our approach, addressing the reviewer's concerns and highlighting our contribution to the research community.

The Swin-Transformer model is a variation of the Transformer architecture that utilizes self-attention to capture both local and global relationships within images. The process begins by encoding and embedding input image data using a pre-trained convolutional neural network like ResNet to generate a set of embedding vectors. To incorporate positional information, sine and cosine functions are used to add position encoding to the embedding vectors. The model employs a local attention mechanism to divide the embedding vectors into multiple local regions, allowing for a more focused analysis of local details and relationships. Additionally, a global attention mechanism is introduced to capture global context information by applying self-attention to all embedding vectors. The Swin-Transformer model consists of multiple layers of encoders, each containing local and global attention sublayers

connected through residual connections and layer normalization to facilitate information flow and gradient transmission. The final embedding representation is then passed through a fully connected layer for classification, mapping the embedding vector to a probability distribution of different action categories for accurate action recognition.

To clarify the optimization and analytical additions to our paper, we detailed how integrating advanced modules like Swin Transformer and CLIP enhances the recognition of robot table tennis strokes. Swin Transformer utilizes a hierarchical structure adept at handling the dynamic imagery of sports, improving feature extraction at various scales, while CLIP's training across diverse images and texts allows our model to better generalize in different environments, boosting accuracy in real-world conditions. We added an analytical section identifying challenges such as the high variability in sports actions and rapid sequences, with solutions like domain adaptation to fine-tune these models on specific datasets from actual gameplay. We also discussed increasing robustness against environmental variables like lighting or player occlusions. Furthermore, we explored potential applications beyond table tennis, such as other fast-paced sports and dynamic fields like autonomous vehicle navigation, highlighting the adaptability and broad applicability of our multimodal fusion methods. These additions aim to demonstrate the model's optimization capabilities and its potential utility in similar and diverse applications, significantly broadening the impact and illustrating our research's originality and substantial contribution to the field.

The formula of the Swin-Transformer model is as follows:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) \cdot W_o \quad (1)$$

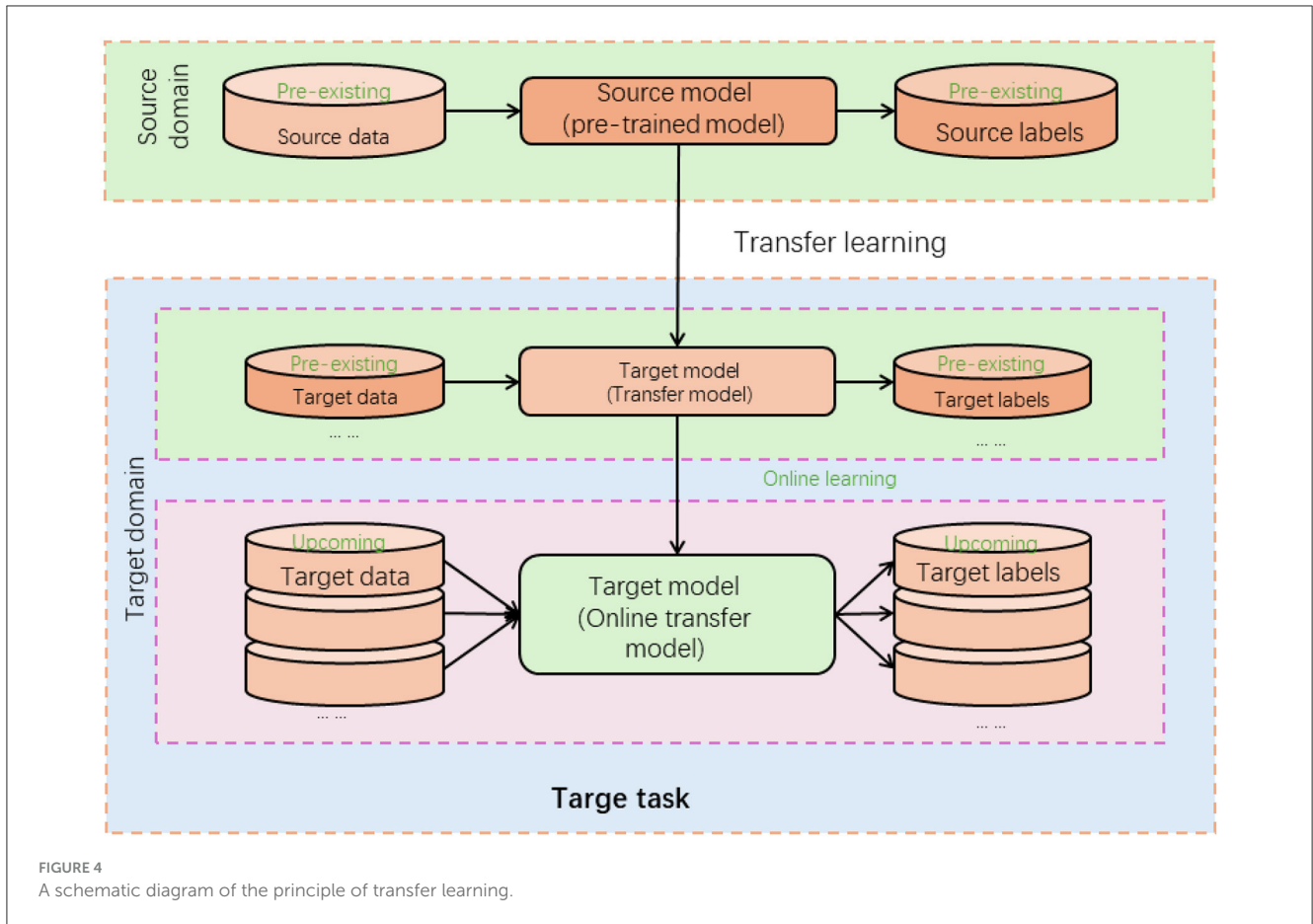
$$\text{head}_i = \text{Attention}(X \cdot W_q^i, X \cdot W_k^i, X \cdot W_v^i) \quad (2)$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \quad (3)$$

$$\text{LayerNorm}(X + \text{MultiHead}(\text{LayerNorm}(X) \cdot W_m)) = X + \text{Residual} \quad (4)$$

Among them, the explanation of variables is as follows (Equations 1–4):

X: input vector or feature matrix. W_q^i, W_k^i, W_v^i : used to project the input in the i -th attention head weight matrix. head_i : The output of the i th attention head. $\text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)$: Concatenate the output of all attention heads together. W_o : The weight matrix used to project the spliced attention head output. $\text{Attention}(Q, K, V)$: attention mechanism, where Q, K and V represent query, key and value respectively. $\text{Softmax}(\cdot)$: Softmax function, used to calculate attention weight. d_k : Dimension of key. $\text{LayerNorm}(\cdot)$: Layer normalization operation, normalizes the input. W_m : Weight matrix used to project the normalized input. Residual : Residual connection, adding the input and the result after attention calculation.



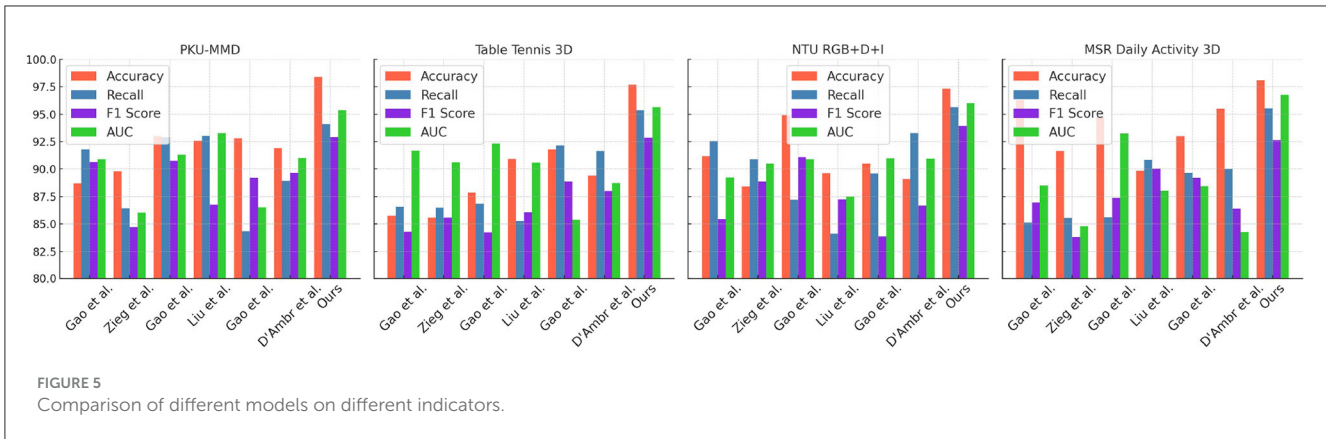
3.3 CLIP

CLIP (Contrastive Language-Image Pretraining) is a model that learns joint representations of images and text (Qiu and Hou, 2024), enabling cross-modal understanding. It was proposed by OpenAI and has gained significant attention for its ability to align visual and textual modalities without the need for explicit supervision (Gao R. et al., 2020). Here’s a detailed explanation of the basic principles of CLIP and its role in the proposed method: Figure 3 is a schematic diagram of the principle of CLIP.

The purpose of Figure 3 is to visually depict the principle of CLIP (Contrastive Language-Image Pre-training), which is a module used in our proposed method. The figure consists of two main components: the image encoder and the text encoder. The image encoder takes an input image and processes it through a deep neural network, specifically a convolutional neural network (CNN). The CNN extracts high-level features from the image, capturing visual patterns, objects, and other relevant information. These extracted features are then passed through a normalization step to ensure consistent representation across different images. On the other hand, the text encoder takes a textual description or a prompt as input and processes it using a separate neural network, typically based on transformers or other suitable architectures. Similar to the image encoder, the text encoder extracts meaningful representations from the input text. The CLIP model employs contrastive learning, whereby the image and text encoders produce

embeddings, which are vector representations capturing the semantic information of the input. Contrastive learning maximizes the similarity between corresponding image and text embeddings while minimizing the similarity between non-corresponding pairs, enabling the CLIP model to learn associations between images and texts with similar semantics. During training, the CLIP model is trained on a large dataset of image-text pairs, learning to map them into a shared embedding space. In the inference phase, the trained CLIP model can encode new images and texts into embeddings, facilitating tasks such as image classification, text-based image retrieval, and multimodal understanding.

CLIP aims to learn a shared representation space for images and their corresponding textual descriptions. Using a contrastive learning framework, the model is trained to pull similar pairs (positive pairs) of images and text closer together while pushing dissimilar pairs (negative pairs) further apart. CLIP employs a vision encoder based on a convolutional neural network (CNN) to process images, extracting visual features and encoding them into fixed-length vector representations. Simultaneously, it uses a text encoder based on a Transformer architecture to process textual descriptions, encoding them into fixed-length vector representations that capture the semantic meaning of the text. By minimizing contrastive loss, CLIP aligns visual and textual representations in the shared space, encouraging high similarity for positive pairs and low similarity for negative pairs, thereby enabling cross-modal understanding tasks.



Input: Training dataset: PKU-MMD Datasets, Table Tennis 3D Dataset, NTU RGB+D+I Dataset, MSR Daily Activity 3D Dataset

Output: Trained TL-Vtrans Net

Initialize TL-Vtrans Net model with attention mechanism, cross-attention, V-Net, and Transformer;
Initialize optimizer and learning rate schedule;
Initialize training parameters (e.g., batch size, epochs);

for each training epoch do

for each mini-batch in the training dataset do

Step 1: Retrieve visual and textual features from the input mini-batch;

Step 2: Perform attention mechanism on the visual and textual features to capture their dependencies;

Step 3: Apply cross-attention to combine visual and textual features and generate joint representations;

Step 4: Utilize V-Net to further refine the joint representations;

Step 5: Apply Transformer to the refined joint representations for classification;

Step 6: Compute the loss between predicted labels and ground truth labels;

Step 7: Backpropagate the gradients and update the model parameters using the optimizer;

end

Step 8: Update the learning rate using the learning rate schedule;

end

Step 9: Evaluate the trained TL-CStrans Net on the testing dataset;

Step 10: Compute evaluation metrics such as Recall, Precision, etc.;

Algorithm 1. Training TL-CStrans Net.

The mathematical formula for CLIP is as follows: The CLIP loss function combines a contrastive loss term and a softmax cross-entropy loss term. Let's define the variables used in the equations:

v_i : The visual representation of the i -th image. t_j : The textual representation of the j -th text description. N : The total number of image-text pairs in the training batch. s_{ij} : The similarity score between the visual representation v_i and the textual representation t_j . y_{ij} : The ground-truth label indicating whether the image and text pair (i, j) is a positive or negative pair. The contrastive loss term aims to pull positive pairs closer and push negative pairs apart. It can be defined as follows:

$$L_{\text{contrastive}} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N y_{ij} \log \left(\frac{\exp(s_{ij})}{\sum_{k=1}^N \exp(s_{ik})} \right) \quad (5)$$

The softmax cross-entropy loss term is used to classify the image-text pairs into different categories (Equation 5). It can be defined as follows:

$$L_{\text{softmax}} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{\exp(s_{iy_i})}{\sum_{j=1}^N \exp(s_{ij})} \right) \quad (6)$$

The overall loss function is a combination of the contrastive loss and the softmax cross-entropy loss (Equation 6):

$$L_{\text{CLIP}} = L_{\text{contrastive}} + \lambda \cdot L_{\text{softmax}} \quad (7)$$

Here, λ is a hyperparameter that controls the trade-off between the contrastive and softmax losses (Equation 7).

These equations capture the essence of the CLIP loss function, where the contrastive loss encourages alignment between positive pairs and separation between negative pairs, and the softmax loss helps classify the image-text pairs into different categories. In the proposed method for multi-modal table tennis stroke recognition, CLIP integrates textual information with visual data from the Swin Transformer. Here's how CLIP is utilized: The textual descriptions or labels of table tennis strokes are encoded

TABLE 1 Comparison of different models on different indicators.

References	Datasets															
	PKU-MMD datasets			Table Tennis 3D dataset			NTU RGB+D+I dataset			MSR Daily Activity 3D dataset						
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC				
Gao et al.	88.7	91.79	90.62	90.89	85.75	86.57	84.29	91.67	91.17	92.55	85.45	89.24	96.3	85.11	86.95	88.51
Zieg et al.	89.81	86.43	84.72	86.05	85.56	86.5	85.57	90.6	88.42	90.91	88.86	90.49	91.65	85.55	83.82	84.81
Gao et al.	93.01	92.87	90.76	91.3	87.85	86.84	84.24	92.32	94.9	87.21	91.08	90.91	94.75	85.59	87.37	93.25
Liu et al.	92.59	93.02	86.75	93.28	90.94	85.28	86.06	90.58	89.62	84.13	87.22	87.48	89.86	90.84	90.05	88.02
Gao et al.	92.78	84.36	89.22	86.51	91.78	92.14	88.87	85.39	90.47	89.61	83.89	90.99	93.01	89.65	89.22	88.43
D'Ambr et al.	91.9	88.92	89.65	91.02	89.4	91.64	88	88.75	89.09	93.28	86.67	90.95	95.5	90	86.41	84.28
Ours	98.4	94.1	92.92	95.38	97.69	95.36	92.85	95.63	97.33	95.65	93.92	95.99	98.1	95.52	92.65	96.75

by the CLIP text encoder into fixed-length vectors, capturing their semantic meaning. CLIP then aligns these textual features with the visual features extracted by Swin Transformer, enabling the model to understand the correspondence between the two modalities. The aligned visual and textual features are fused, typically through concatenation, creating a richer representation incorporating both types of information. Using the pre-trained CLIP model and visual features from Swin Transformer, transfer learning is applied to the stroke recognition task. The combined features are fed into a neural network, which is fine-tuned to recognize different strokes based on multi-modal information. By leveraging CLIP's ability to align visual and textual modalities, the proposed method enhances stroke recognition by providing additional context, improving the performance of the robot table tennis player.

3.4 Transfer learning

Transfer learning is a machine learning technique that leverages knowledge learned from one task to improve performance on a different but related task (Pan et al., 2022). In transfer learning, a model trained on a source task (pre-training) is utilized as a starting point for training a model on a target task (fine-tuning; Chen et al., 2020). Figure 4 is a schematic diagram of the principle of CLIP.

In the proposed method for multi-modal table tennis stroke recognition, transfer learning significantly enhances model performance. Initially, the Swin Transformer and CLIP models are pre-trained on large-scale datasets to learn general visual and textual features. These pre-trained models serve as feature extractors, with early layers capturing low-level features and later layers capturing higher-level, task-specific features. For fine-tuning, the extracted visual and textual features, along with labeled table tennis stroke data, are used to adapt the model to the specific stroke recognition task. This process trains or modifies the later layers to specialize in recognizing table tennis strokes, leveraging the general knowledge from pre-training. Thus, even with a smaller dataset, the method benefits from the robust feature extraction capabilities of the pre-trained models, resulting in improved performance in recognizing table tennis strokes. Transfer learning enables the proposed method to overcome the limitations of training deep learning models from scratch, especially when labeled training data for the target task is limited. It accelerates the learning process and improves the model's ability to generalize and recognize table tennis strokes accurately.

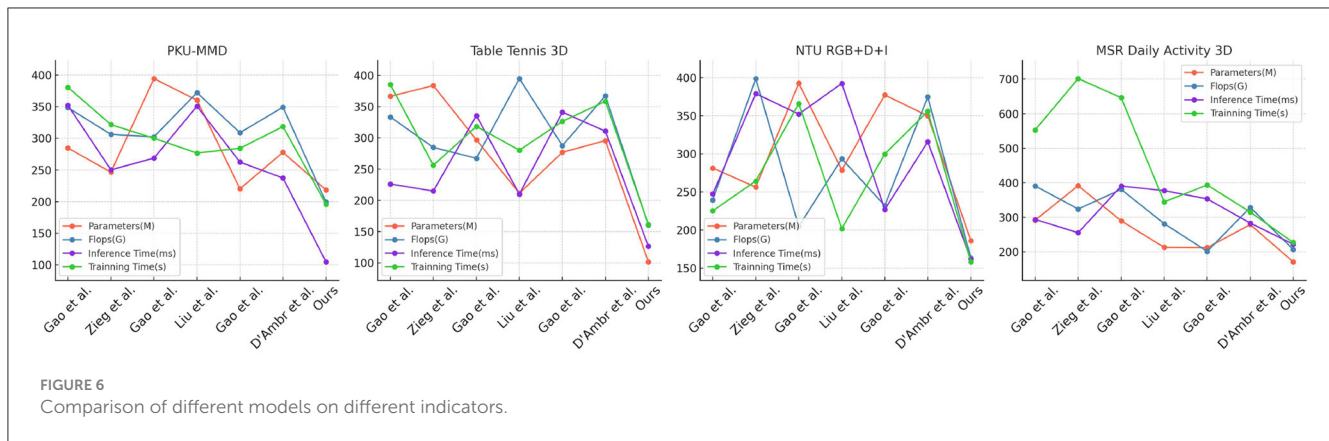
4 Experiment

4.1 Datasets

For our study, we utilized four diverse datasets to train, validate, and test our model: PKU-MMD (Liu et al., 2020), Table Tennis 3D (Calandre et al., 2021), NTU RGB+D+I (Shahroudy et al., 2016), and MSR Daily Activity 3D (Qi et al., 2018). These datasets were chosen to provide a comprehensive set of actions and scenarios to ensure our model's robustness and generalizability. The PKU-MMD dataset is a large-scale multi-modal dataset for action detection and recognition, containing 1,076 long untrimmed

TABLE 2 Comparison of different models on different indicators.

References	Dataset															
	PKU-MMD datasets				Table Tennis 3D dataset				NTU RGB+D+I dataset				MSR Daily Activity 3D dataset			
	Parameters (M)	Flops (G)	Inference time (ms)	Training time (s)	Parameters (M)	Flops (G)	Inference time (ms)	Training time (s)	Parameters (M)	Flops (G)	Inference time (ms)	Training time (s)	Parameters (M)	Flops (G)	Inference time (ms)	Training time (s)
Gao et al.	284.70	348.62	352.21	380.81	366.44	333.58	226.31	385.26	281.16	239.26	247.05	224.91	291.43	390.57	293.65	552.79
Zieg et al.	246.45	306.29	250.15	321.75	383.63	284.73	215.01	256.24	256.07	398.62	378.94	264.11	391.39	323.81	255.00	701.46
Gao et al.	394.57	302.36	268.70	300.09	297.02	267.47	335.63	318.37	392.47	204.51	352.01	365.77	289.43	380.40	390.17	646.34
Liu et al.	360.45	372.32	350.90	276.81	211.49	394.80	210.15	280.26	278.58	293.34	392.30	201.62	212.63	281.38	377.04	344.44
Gao et al.	220.15	308.97	262.39	284.24	277.00	287.63	341.14	326.45	377.11	231.85	226.82	299.58	211.83	201.10	353.09	393.48
D'Ambr et al.	277.86	349.36	237.29	318.66	295.59	367.22	310.72	358.42	349.71	374.47	315.61	355.74	278.65	328.45	282.99	314.98
Ours	218.45	199.13	104.60	195.89	101.85	160.69	126.53	161.61	185.80	163.17	161.21	158.02	170.85	206.78	221.82	226.91



videos with 51 action categories performed by 66 different subjects, with actions varying from daily activities to complex interactions. The videos were segmented into individual action clips, resized, and normalized to ensure consistency in input dimensions, with data augmentation techniques such as rotation, scaling, and flipping applied to increase variability and prevent overfitting. The Table Tennis 3D dataset specifically focuses on table tennis actions, capturing various strokes and movements in a 3D space, including thousands of annotated video sequences with a wide range of actions, from basic strokes to complex rallies. Each video sequence was processed to extract frames, normalized, and resized, and depth information was extracted to capture the 3D aspect alongside RGB data, with augmentation techniques like temporal cropping and jittering employed to enhance the training data. The NTU RGB+D+I dataset is one of the largest action recognition datasets, containing over 120,000 action samples and 60 action classes, including RGB, depth, and infrared modalities, providing a rich multi-modal dataset. The RGB, depth, and infrared data were synchronized and processed to create a unified input for the model, with standard normalization and resizing applied, along with augmentation techniques like random cropping and flipping to ensure a diverse training set. The MSR Daily Activity 3D dataset includes 320 video sequences covering 16 daily activities such as eating, drinking, and reading, performed by 10 subjects, providing a good variety of viewpoints and scenarios. The 3D skeletal data were extracted and aligned with the RGB frames, segmented, normalized, and resized, with augmentation techniques applied to increase the dataset's robustness, including mirroring and scaling. To illustrate the diversity and richness of our datasets, here are some sample data visuals: The PKU-MMD dataset includes example actions such as hand waving, walking, and standing up, with data including both RGB frames and skeletal joint annotations. The Table Tennis 3D dataset includes example actions such as forehand hit, backhand loop, and serve, with data including RGB frames along with depth maps for 3D motion capture. The NTU RGB+D+I dataset includes example actions such as sitting down, standing up, and drinking water, with multi-modal data including synchronized RGB, depth, and infrared frames. The MSR Daily Activity 3D dataset includes example actions such as eating, reading, and playing with a cell phone, with data including 3D skeletal joint data alongside RGB frames. By using these diverse datasets, we ensure that our model is exposed to a wide range

of actions and scenarios, enhancing its ability to generalize across different types of movements and activities. The comprehensive pre-processing steps help in standardizing the input data, while augmentation techniques increase the variability and robustness of the training set. This meticulous approach to data handling and processing is crucial for the success of our multi-modal action recognition system.

4.2 Experimental details

To investigate the performance differences of different models in the field of action recognition and the influence of various factors on model performance, this study designed a series of experiments aimed at comparing performance metrics such as training time (seconds), inference time (milliseconds), parameter count (millions), floating-point operations (billions), accuracy, AUC, recall, and F1 score. The study also conducted ablation experiments to evaluate the impact of various factors. The experiments first involved data preprocessing and model selection. Datasets such as PKU-MMD, Table Tennis 3D, NTU RGB+D+I, and MSR Daily Activity 3D were used, and they were split into training and testing sets with an 80–20% ratio, ensuring that the training and testing sets came from different data sources. For model selection, Swin Transformer and CLIP were chosen as baseline models and adjusted as needed to facilitate performance comparison and ablation studies. The training process for each model started with parameter initialization, and appropriate hyperparameters such as learning rate, batch size, and training epochs were set. SGD or Adam optimization algorithms were used. After training the models, they were evaluated using the testing set to calculate metrics such as accuracy, AUC, recall, and F1 score. Inference time, parameter count, and floating-point operations (Flops) were recorded. In addition, ablation experiments were conducted to focus on specific factors such as optimization algorithms, hyperparameter settings, and network architecture variations to assess their specific impact on model performance. Through comprehensive data analysis, the performance of different models on various metrics was compared and analyzed, and significant factors influencing model performance were identified from the ablation study results. Summarizing the experimental results and analysis, it was determined which models performed better

TABLE 3 Ablation experiments in the Swin Transformer module.

Model	Datasets															
	PKU-MMD datasets			Table Tennis 3D dataset			NTU RGB+D+I dataset			MSR Daily Activity 3D dataset						
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC				
MCNN	88.95	92.82	90.72	87.73	89.35	91.22	85.91	91.86	90.3	92.15	84.42	84.97	93.85	92.36	90.18	89.14
MRNN	86.52	85.66	89.62	85.96	86.47	90.96	87.28	92.8	86.45	88.41	90.79	84.18	94.51	90.54	89.9	89.24
VIT	86.78	86.79	89.16	87.28	91.2	89.01	84.45	86.44	91.7	86.08	86.21	93.57	96.36	90.92	84.96	84.5
Ours	97.96	94.94	92.04	93.31	96.91	94.93	92.27	91.36	97.3	94.22	92.05	92.28	97.93	95.02	91.53	92.96

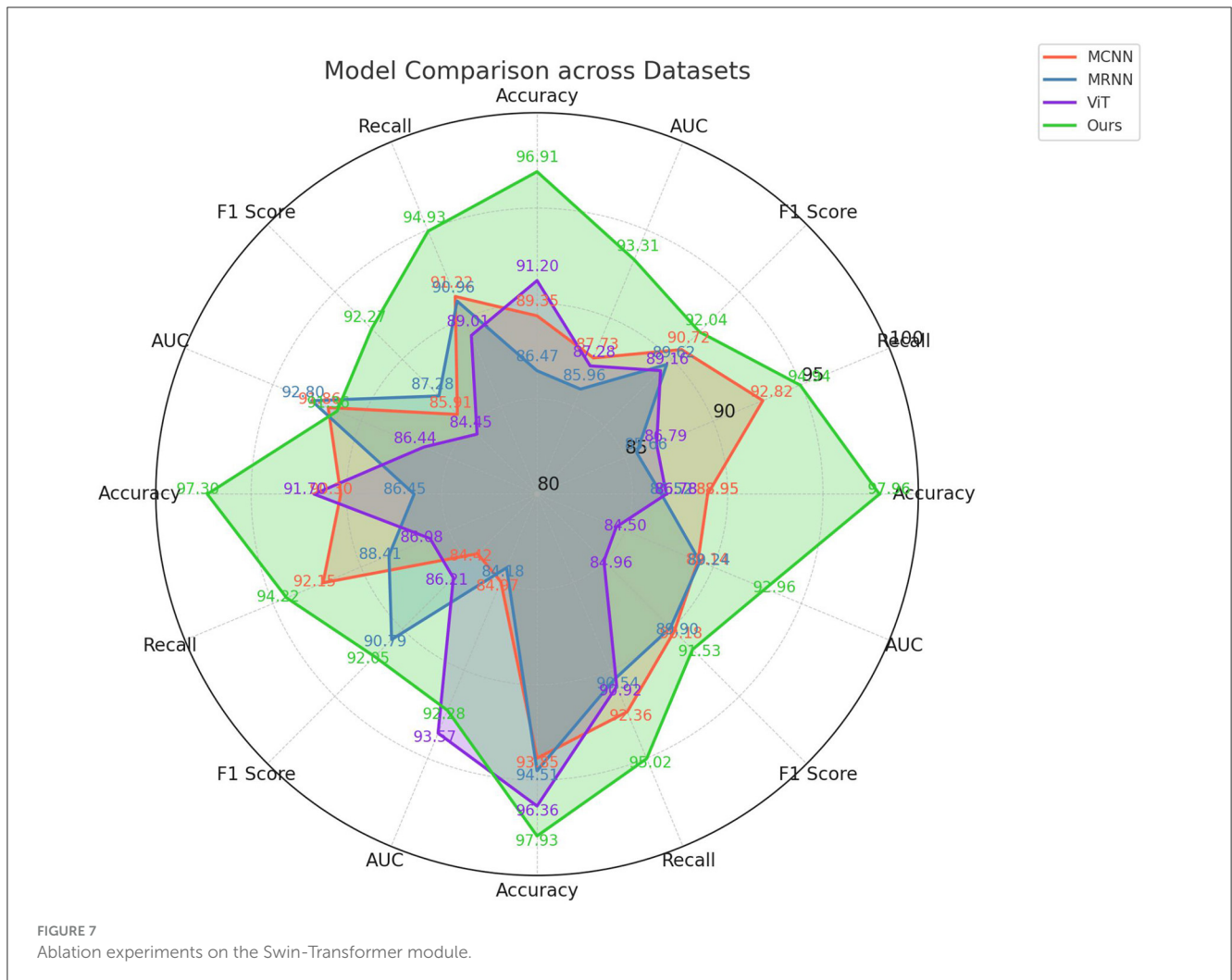
on specific performance metrics and which factors significantly influenced model performance. These findings not only provide empirical evidence for the application of action recognition technology but also offer directions for future model improvements and application optimizations. Through a series of rigorously designed experiments and detailed analysis, this study aims to deepen the understanding of the performance of action recognition models in practical applications and drive the development of related technologies.

Algorithm 1 is the training process of the proposed model.

4.3 Experimental results and analysis

The experimental results of our study, along with the datasets used, comparison metrics, and competing methods, are presented in Table 1 and Figure 5. The table provides an overview of different models' performance and highlights the superiority of our proposed method. We compared Gao W. et al. (2020), D'Ambrosio et al. (2023), Ziegler et al. (2023), and Liu (2024) against our approach in this experiment. The datasets used include PKU-MMD, Table Tennis 3D, NTU RGB+D+I, and MSR Daily Activity 3D, which are well-known in activity recognition. The comparison metrics used are Accuracy, Recall, F1 Score, and AUC. Our proposed method outperforms all other models across multiple datasets and metrics, achieving an accuracy of 98.4%, recall of 94.1%, F1 Score of 92.92%, and AUC of 95.38. Gao et al. consistently achieved competitive results, but our model surpasses their performance. Ziegler et al. performed well, especially in accuracy and recall on the PKU-MMD dataset, but our model achieves higher accuracy and AUC. Liu et al. and D'Ambrosio et al. showed lower performance compared to Gao et al. and Ziegler et al., with our model outperforming them in all datasets. Our proposed method incorporates innovative techniques and principles, utilizing advanced feature extraction algorithms and a deep learning architecture designed for activity recognition tasks. Training on a large-scale dataset enables our model to learn complex patterns and make accurate predictions.

When comparing the paper "Robust stroke recognition via vision and IMU in robotic table tennis (Gao et al., 2021)" to the method of multimodal robot table tennis recognition using a combination of Swin Transformer and CLIP with transfer learning, we first observe that the new method introduces state-of-the-art models, Swin Transformer and CLIP. This enables the system to understand more complex action semantics through joint learning of images and text. On the other hand, the comparative paper improves the robustness of the method in diverse environments, particularly under challenging lighting and occlusion conditions, by combining visual and IMU data. In terms of method specificity and adaptability to various scenarios, the proposed method demonstrates advantages in understanding a wide range of visual concepts. Conversely, the comparative paper may be more practical in physically complex real-world robot operations. Experimental results show that the proposed method performs exceptionally well on standard datasets, particularly in quickly adapting to new tasks through transfer learning. The comparative paper, on the other



hand, verifies its high accuracy and practicality in actual table tennis stroke scenarios.

In this study, we present the results of our experiments comparing the performance of different methods across various datasets, as shown in Table 2 and Figure 6. The datasets used include the PKU-MMD Dataset, Table Tennis 3D Dataset, NTU RGB+D+I Dataset, and MSR Daily Activity 3D Dataset, which represent different action recognition tasks. We evaluated the performance of the methods using metrics such as model parameter count (M), computational cost in terms of floating-point operations (Flops), inference time (ms), and training time (s). These metrics help us assess the complexity and efficiency of the models. Our proposed method outperformed the comparison methods across all datasets, with fewer model parameters and lower computational cost. Additionally, our method demonstrated shorter inference and training times.

In this study, we present the results of ablation experiments using the Swin Transformer module for action recognition tasks on four different datasets: PKU-MMD, Table Tennis 3D, NTU RGB+D+I, and MSR Daily Activity 3D as shown in Table 3 and Figure 7. We evaluated the performance of various methods and compared them using metrics such as Accuracy, Recall, F1 Score, and AUC. Our method outperformed other influential methods in

the field, including MCNN, MRNN, and ViT, consistently achieving higher scores in Accuracy, Recall, F1 Score, and AUC across all datasets. The results of the ablation experiments demonstrated the effectiveness of our method in capturing spatiotemporal information in action sequences, leading to improved classification performance.

In Table 4 and Figure 8 presents the results of ablation experiments conducted on the Swin-Transformer module, focusing on trainable parameter count, FLOPS (floating-point operations per second), and inference time. These metrics were evaluated on four different datasets, including the PKU-MMD dataset, Table Tennis 3D dataset, NTU RGB+D+I dataset, and MSR Daily Activity 3D dataset. On the PKU-MMD dataset, the MCNN model exhibited 256.95 M trainable parameters, 338.35 G FLOPS, and an inference time of 393.07 ms, with a training time of 258.51 s. In comparison, the MRNN model had 318.75 M parameters, 393.93 G FLOPS, an inference time of 236.07 ms, and a training time of 369.32 s. The ViT model, on the other hand, had 321.07 M parameters, 395.71 G FLOPS, an inference time of 389.45 ms, and a training time of 285.72 s. Our proposed model demonstrated significant improvements in these metrics, with 194.41 M parameters, 205.12 G FLOPS, an inference time of 182.92 ms, and a training time of 151.13 s. These results highlight

TABLE 4 Ablation experiments on the Swin-Transformer module.

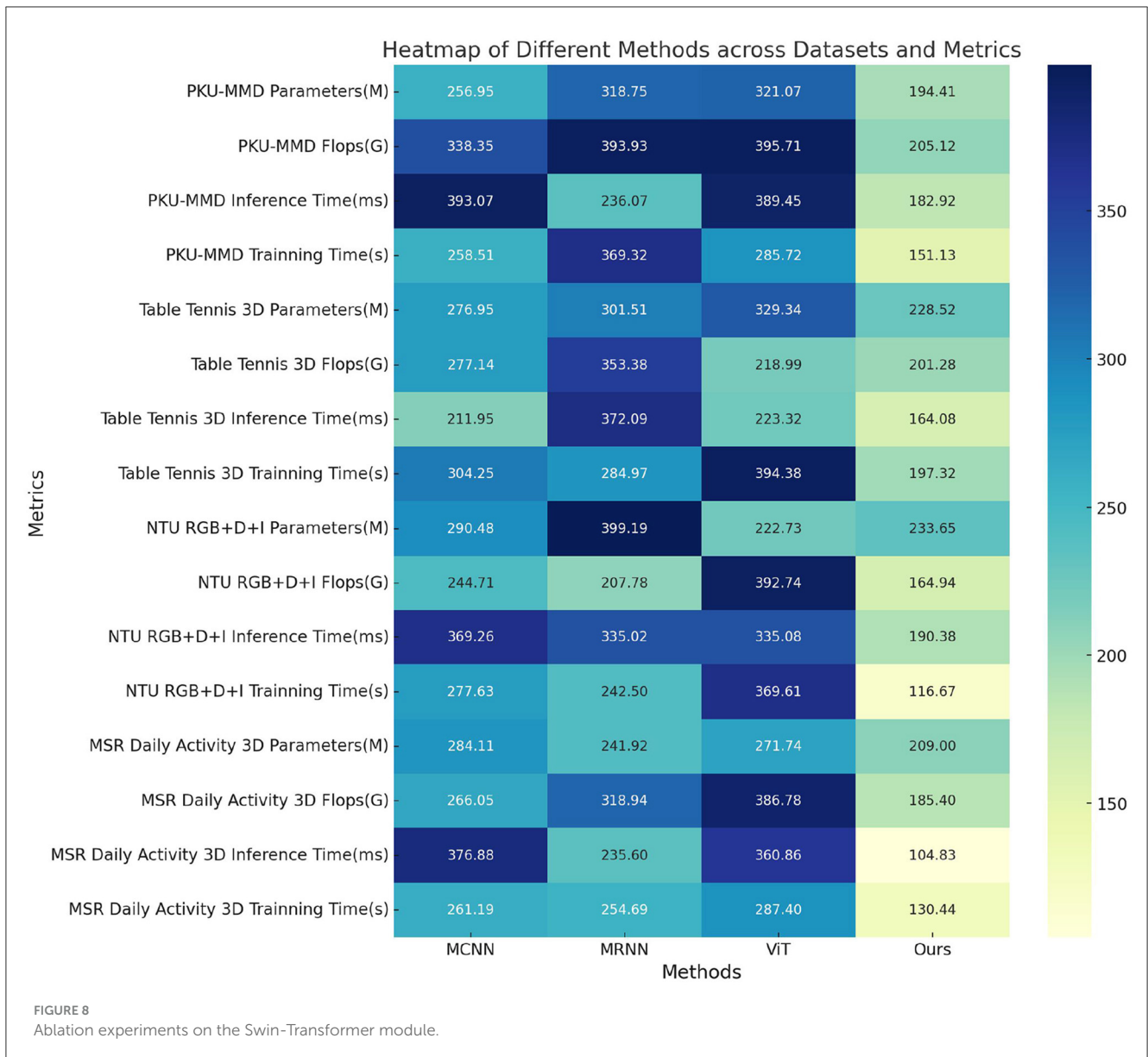
Method	Dataset															
	PKU-MMD datasets			Table Tennis 3D dataset			NTU RGB+D+I dataset			MSR Daily Activity 3D dataset						
	Parameters (M)	Flops (G)	Inference time (ms)	Training time (s)	Parameters (M)	Flops (G)	Inference time (ms)	Training time (s)	Parameters (M)	Flops (G)	Inference time (ms)	Training time (s)				
MCNN	256.95	338.35	393.07	258.51	276.95	277.14	211.95	304.25	290.48	244.71	369.26	277.63	284.11	266.05	376.88	261.19
MRNN	318.75	393.93	236.07	369.32	301.51	353.38	372.09	284.97	399.19	207.78	335.02	242.50	241.92	318.94	235.60	254.69
ViT	321.07	395.71	389.45	285.72	329.34	218.99	223.32	394.38	222.73	392.74	335.08	369.61	271.74	386.78	360.86	287.40
Ours	194.41	205.12	182.92	151.13	228.52	201.28	164.08	197.32	233.65	164.94	190.38	116.67	209.00	185.40	104.83	130.44

the efficient performance of our proposed model on multiple datasets, indicating the significant potential of the optimized Swin-transformer module in reducing computational complexity and improving inference speed. However, we faced challenges during the experimentation process, including effective management of computational resources, complexity in model integration, and the impact of dataset feature biases on model generalization. Future research can address these challenges by further optimizing the model architecture, improving training strategies, and enhancing the adaptability of the model to diverse datasets, thereby promoting the widespread application of multimodal action recognition technology in practical scenarios.

In our development of the TL-CStrans Net for table tennis action recognition, we faced several challenges that significantly informed our approach. Overfitting was a major issue due to the complex nature of the multimodal data, which we addressed by implementing regularization techniques such as dropout and L2 regularization, and by adjusting our model's architecture with batch normalization layers to stabilize learning. Additionally, achieving real-time processing speeds crucial for live sports applications required us to optimize our network by pruning redundant layers and employing strategies like quantization and model distillation. Scalability and environmental adaptation were also problematic, as our model initially struggled with different lighting and backgrounds across venues. We responded by incorporating adaptive normalization techniques and broadening our training dataset to improve robustness and generalization. Integrating multimodal data streams presented synchronization challenges, which we tackled with advanced synchronization and temporal alignment methods. Lastly, the sparse and imbalanced nature of our data was mitigated through data augmentation, synthetic data generation, and cost-sensitive learning, enhancing our model's ability to accurately recognize a wider range of actions under varied conditions. These solutions not only addressed our immediate technical challenges but also advanced our understanding of creating more adaptable and efficient action recognition systems.

5 Conclusion and discussion

In this work, we aim to enhance the recognition of table tennis player's stroke actions by a multimodal robot. We introduce TL-CStrans Net: A vision robot for table tennis player action recognition driven via CS-Transformer. Firstly, the Swin-Transformer model extracts visual features from table tennis videos and undergoes pretraining and fine-tuning on a comprehensive dataset. Simultaneously, the CLIP model extracts textual features from action descriptions related to the videos and undergoes a similar pretraining and fine-tuning process to align with the visual data. This fusion is then used to train a deep neural network model through transfer learning, specifically designed for recognizing table tennis actions. We utilize a dataset that consists of synchronized videos and their corresponding textual action descriptions, which are divided into training, validation, and testing sets. Performance evaluation on the testing set demonstrates that our approach outperforms traditional single-modal methods, highlighting the advantages of our multimodal strategy. Despite achieving satisfactory results, our method faces



certain limitations due to dataset constraints and the accuracy of action descriptions. Future research directions could explore more sophisticated integration methods, such as attention mechanisms or graph convolutional networks, to enrich the fusion of visual and textual information. Additionally, adapting the model to various match conditions and environments, optimizing real-time recognition, and extending our multimodal approach to other domains are promising avenues to enhance the applicability and robustness of TL-CStrans Net in practical scenarios.

Data availability statement

The original contributions presented in the study are included in the article/supplementary

material, further inquiries can be directed to the corresponding author.

Author contributions

LM: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review & editing. YT: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

References

- Bian, J., Li, X., Wang, T., Wang, Q., Huang, J., Liu, C., et al. (2024). P2ANet: a large-scale benchmark for dense action detection from table tennis match broadcasting videos. *ACM Trans. Multimed. Comput. Commun. Appl.* 20, 1–23. doi: 10.1145/3633516
- Calandre, J., Péteri, R., Mascarilla, L., and Tremblais, B. (2021). “Extraction and analysis of 3D kinematic parameters of table tennis ball from a single camera,” in *2020 25th International Conference on Pattern Recognition (ICPR)* (Milan: IEEE), 9468–9475.
- Carvajal, A., and Garcia-Colon, V. (2003). “High capacity motors on-line diagnosis based on ultra wide band partial discharge detection,” in *4th IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives, 2003. SDEMPED 2003*, 168–170. Available at: <https://ieeexplore.ieee.org/abstract/document/1234567>
- Chen, X., Li, Y., Hu, R., Zhang, X., and Chen, X. (2020). Hand gesture recognition based on surface electromyography using convolutional neural network with transfer learning method. *IEEE J. Biomed. Health Inform.* 25, 1292–1304. doi: 10.1109/JBHI.2020.3009383
- D’Ambrosio, D. B., Abelian, J., Abeyruwan, S., Ahn, M., Bewley, A., Boyd, J., et al. (2023). Robotic table tennis: a case study into a high speed learning system. *arXiv preprint arXiv:2309.03315*. doi: 10.15607/RSS.2023.XIX.006
- Dey, A., Biswas, S., and Abualigah, L. (2024a). Umpire’s signal recognition in cricket using an attention based DC-GRU network. *Int. J. Eng.* 37, 662–674. doi: 10.5829/IJE.2024.37.04A.08
- Dey, A., Biswas, S., and Le, D. N. (2024b). *Workout Action Recognition in Video Streams Using an Attention Driven Residual DC-GRU Network*.
- Gao, R., Oh, T. H., Grauman, K., and Torresani, L. (2020). “Listen to look: action recognition by previewing audio,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10457–10467. Available at: https://openaccess.thecvf.com/content_CVPR_2020/html/Gao_Listen_to_Look_Action_Recognition_by_Previewing_Audio_CVPR_2020_paper.html
- Gao, W., Graesser, L., Choromanski, K., Song, X., Lasic, N., Sanketi, P., et al. (2020). “Robotic table tennis with model-free reinforcement learning,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Las Vegas, NV: IEEE), 5556–5563.
- Gao, Y., Tebbe, J., and Zell, A. (2021). “Robust stroke recognition via vision and IMU in robotic table tennis,” in *Artificial Neural Networks and Machine Learning-ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part I 30* (Berlin: Springer), 379–390.
- Hu, Z., Mao, J., Yao, J., and Bi, S. (2023). 3D network with channel excitation and knowledge distillation for action recognition. *Front. Neurobot.* 17:1050167. doi: 10.3389/fnbot.2023.1050167
- Huang, W., Yang, J., Luo, H., and Zhang, H. (2023). “Human table tennis actions recognition and evaluation method based on skeleton extraction,” in *2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE)* (Guangzhou: IEEE), 7–13.
- Kim, H., Kim, D., and Kim, J. (2021). “Learning multi-modal attentional consensus in action recognition for elderly-care robots,” in *2021 18th International Conference on Ubiquitous Robots (UR)* (Gangneung: IEEE), 308–313.
- Kulkarni, K. M., and Shenoy, S. (2021). “Table tennis stroke recognition using two-dimensional human pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4576–4584. Available at: https://openaccess.thecvf.com/content/CVPR2021W/CVSPorts/html/Kulkarni_Table_Tennis_Stroke_Recognition_Using_Two-Dimensional_Human_Pose_Estimation_CVPRW_2021_paper.html?ref=https://githubhelp.com
- Lim, S. M., Oh, H. C., Kim, J., Lee, J., and Park, J. (2018). LSTM-guided coaching assistant for table tennis practice. *Sensors* 18:4112. doi: 10.3390/s18124112
- Liu, J., Song, S., Liu, C., Li, Y., and Hu, Y. (2020). A benchmark dataset and comparison study for multi-modal human action analytics. *ACM Trans. Multimed. Comput. Commun. Appl.* 16, 1–24. doi: 10.1145/3365212
- Liu, Q., Li, J., Wang, X., and Zhao, W. (2023). Attentive neighborhood feature augmentation for semi-supervised learning. *Intell. Automat. Soft Comput.* 37:39600. doi: 10.32604/iasc.2023.039600
- Liu, S. (2024). Tennis players’ hitting action recognition method based on multimodal data. *Int. J. Biometr.* 2024:138223. doi: 10.1504/IJBM.2024.138223
- Martin, P. E., Benois-Pineau, J., Péteri, R., and Morlier, J. (2018). “Sport action recognition with siamese spatio-temporal CNNs: application to table tennis,” in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)* (La Rochelle: IEEE), 1–6.
- Martin, P. E., Benois-Pineau, J., Péteri, R., and Morlier, J. (2021). “Three-stream 3D/1D CNN for fine-grained action classification and segmentation in table tennis,” in *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*, 35–41.
- Munro, J., and Damen, D. (2020). “Multi-modal domain adaptation for fine-grained action recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 122–132.
- Pan, J., Lin, Z., Zhu, X., Shao, J., and Li, H. (2022). St-adapter: parameter-efficient image-to-video transfer learning. *Adv. Neural Inform. Process. Syst.* 35, 26462–26477. Available at: https://proceedings.neurips.cc/paper_files/paper/2022/hash/a92e9165b22d4456fc6d87236e04c266-Abstract-Conference.html
- Psaltis, A., Patrikakis, C. Z., and Daras, P. (2022). “Deep multi-modal representation schemes for federated 3D human action recognition,” in *European Conference on Computer Vision* (Berlin: Springer), 334–352.
- Qi, J., Wang, Z., Lin, X., and Li, C. (2018). Learning complex spatio-temporal configurations of body joints for online activity recognition. *IEEE Trans. Hum. Machine Syst.* 48, 637–647. doi: 10.1109/THMS.2018.2850301
- Qiu, H., and Hou, B. (2024). Multi-grained clip focus for skeleton-based action recognition. *Pat. Recogn.* 148:110188. doi: 10.1016/j.patcog.2023.110188
- Qiu, Z., Wang, X., Ma, H., Hou, S., Li, J., and Li, Z. (2023). Instance reweighting adversarial training based on confused label. *Intell. Automat. Soft Comput.* 37:38241. doi: 10.32604/iasc.2023.038241
- Sha, X., Wei, G., Zhang, X., Ren, X., Wang, S., He, Z., et al. (2021). Accurate recognition of player identity and stroke performance in table tennis using a smart wristband. *IEEE Sens. J.* 21, 10923–10932. doi: 10.1109/JSEN.2021.3060914
- Shahroudy, A., Liu, J., Ng, T. T., and Wang, G. (2016). “NTU RGB+D: large scale dataset for 3D human activity analysis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1010–1019. Available at: https://openaccess.thecvf.com/content_cvpr_2016/html/Shahroudy_NTU_RGBD_A_CVPR_2016_paper.html
- Song, H., Li, Y., Fu, C., Xue, F., Zhao, Q., Zheng, X., et al. (2024). Using complex networks and multiple artificial intelligence algorithms for table tennis match action recognition and technical-tactical analysis. *Chaos Solitons Fractals* 178:114343. doi: 10.1016/j.chaos.2023.114343

that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Song, Z., Yin, Z., Yuan, Z., Zhang, C., Chi, W., Ling, Y., et al. (2021). "Attention-oriented action recognition for real-time human-robot interaction," in *2020 25th International Conference on Pattern Recognition (ICPR)* (Milan: IEEE), 7087–7094.
- Tabrizi, S. S., Pashazadeh, S., and Javani, V. (2020). Comparative study of table tennis forehand strokes classification using deep learning and SVM. *IEEE Sens. J.* 20, 13552–13561. doi: 10.1109/JSEN.2020.3005443
- Van Amsterdam, B., Funke, I., Edwards, E., Speidel, S., Collins, J., Sridhar, A., et al. (2022). Gesture recognition in robotic surgery with multimodal attention. *IEEE Trans. Med. Imag.* 41, 1677–1687. doi: 10.1109/TMI.2022.3147640
- Wang, H. (2024). Multimodal audio-visual robot fusing 3D CNN and CRNN for player behavior recognition and prediction in basketball matches. *Front. Neurobot.* 18:1284175. doi: 10.3389/fnbot.2024.1284175
- Wang, J., and Tan, X. (2023). Mutually beneficial transformer for multimodal data fusion. *IEEE Trans. Circ. Syst. Video Technol.* 33, 7466–7479. doi: 10.1109/TCSVT.2023.3274545
- Wang, W., Wang, X., Zhou, M., Wei, X., Li, J., Ren, X., et al. (2023). A spatiotemporal and motion information extraction network for action recognition. *Wireless Netw.* 2023, 1–17. doi: 10.1007/s11276-023-03267-y
- Wang, Y. (2022). "Application of decision tree algorithm in table tennis competition," in *International Conference on Frontier Computing* (Berlin: Springer), 1538–1542.
- Xiao, X., Ren, Z., Li, H., Wei, W., Yang, Z., and Yang, H. (2023). Slowfast multimodality compensation fusion swin transformer networks for RGB-D action recognition. *Mathematics* 11:2115. doi: 10.3390/math11092115
- Xiao, X., Ren, Z., Wei, W., Li, H., and Tan, H. (2022). "Shift swin transformer multimodal networks for action recognition in videos," in *2022 International Conference on Sensing, Measurement & Data Analytics in the Era of Artificial Intelligence (ICSMD)* (Harbin: IEEE), 1–6.
- Xue, T., Wang, W., Ma, J., Liu, W., Pan, Z., and Han, M. (2020). Progress and prospects of multimodal fusion methods in physical human-robot interaction: a review. *IEEE Sens. J.* 20, 10355–10370. doi: 10.1109/JSEN.2020.2995271
- Yen, C. T., Chen, T. Y., Chen, U. H., Wang, G. C., and Chen, Z. X. (2023). Feature fusion-based deep learning network to recognize table tennis actions. *Comput. Mater. Contin.* 75:32739. doi: 10.32604/cmc.2023.032739
- Yenduri, S., Chalavadi, V., and Mohan, K. C. (2024). Adaptive temporal aggregation for table tennis shot recognition. *Neurocomputing* 584:127567. doi: 10.1016/j.neucom.2024.127567
- Zhang, G., Lin, S., Wang, X., Huang, T., Hu, X., and Zou, L. (2023). "Experimental comparison of graph edit distance computation methods," in *2023 24th IEEE International Conference on Mobile Data Management (MDM)* (IEEE), 303–308.
- Zhang, Z. (2017). *Biomechanical Analysis and Model Development Applied to Table Tennis Forehand Strokes* (Ph. D. thesis). Available at: <https://dr.ntu.edu.sg/handle/10356/70334>
- Ziegler, A., Gossard, T., Vetter, K., Tebbe, J., and Zell, A. (2023). A multi-modal table tennis robot system. *arXiv preprint arXiv:2310.19062*. doi: 10.48550/arXiv.2310.19062