



OPEN ACCESS

EDITED BY

Xin Jin,
Yunnan University, China

REVIEWED BY

Sara Akan,
Beykoz University, Türkiye
Eray Önter,
Namık Kemal University, Türkiye
Niveen Farid,
National Institute of Standards, Egypt

*CORRESPONDENCE

Guanlin Wang
✉ wangguanlin0121@163.com

RECEIVED 27 May 2024

ACCEPTED 15 July 2024

PUBLISHED 14 August 2024

CITATION

Wang G (2024) RL-CWtrans Net: multimodal swimming coaching driven via robot vision. *Front. Neurobot.* 18:1439188. doi: 10.3389/fnbot.2024.1439188

COPYRIGHT

© 2024 Wang. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). 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.

RL-CWtrans Net: multimodal swimming coaching driven via robot vision

Guanlin Wang*

Faculty of Education, University of Macau, Macau, Macau SAR, China

In swimming, the posture and technique of athletes are crucial for improving performance. However, traditional swimming coaches often struggle to capture and analyze athletes' movements in real-time, which limits the effectiveness of coaching. Therefore, this paper proposes RL-CWtrans Net: a robot vision-driven multimodal swimming training system that provides precise and real-time guidance and feedback to swimmers. The system utilizes the Swin-Transformer as a computer vision model to effectively extract the motion and posture features of swimmers. Additionally, with the help of the CLIP model, the system can understand natural language instructions and descriptions related to swimming. By integrating visual and textual features, the system achieves a more comprehensive and accurate information representation. Finally, by employing reinforcement learning to train an intelligent agent, the system can provide personalized guidance and feedback based on multimodal inputs. Experimental results demonstrate significant advancements in accuracy and practicality for this multimodal robot swimming coaching system. The system is capable of capturing real-time movements and providing immediate feedback, thereby enhancing the effectiveness of swimming instruction. This technology holds promise.

KEYWORDS

multimodal robot, artificial neural networks, feature extraction, robot vision, Swin-Transformer, clip, reinforcement learning

1 Introduction

Swimming action recognition holds significant research value and application prospects in the fields of sports science and health monitoring (Santos et al., 2021). Firstly, it can help swimming athletes optimize their training, improve their technical skills, and enhance their competitive performance. It can also effectively monitor posture and movements during sports activities, thereby preventing sports injuries (Cabrera-Arellano et al., 2022). Secondly, swimming action recognition technology plays a crucial role in water rescue and rehabilitation training. It can provide real-time monitoring and analysis of patients' movement status, offer precise rehabilitation recommendations, and quickly identify abnormal behaviors of drowning individuals in emergency situations, facilitating timely rescue measures (Hamidi Rad et al., 2022). Furthermore, with the rapid development of artificial intelligence and computer vision technology, research on swimming action recognition not only promotes the advancement of related technologies but also finds applications in other fields such as virtual reality and smart homes, expanding its scope of application. Therefore, conducting research on swimming action recognition is not only of theoretical significance but also possesses extensive practical value.

Traditional methods for swimming action recognition mainly rely on symbolic AI and knowledge representation. Firstly, manual feature extraction is widely adopted in early research on swimming action recognition, as it depends on the researcher's experience and expertise to select and extract action features. For example, Wang et al. (2019) proposed a swimming action recognition method based on manual feature extraction. Additionally, Kon et al. (2015) demonstrated the application and effectiveness of manual feature extraction in swimming action analysis. Secondly, rule-based methods utilize a series of predefined rules for action recognition. These methods exhibit high determinism and reliability, performing well even in the face of complex or highly variable actions. Chen et al. (2011) proposed a rule-based system for automated swimming action analysis, while Omae et al. (2017) provided a rule-based system framework for analyzing swimming performance. Furthermore, logistic regression, as a statistical method, learns features from training data to make classification decisions. It not only has important applications in action recognition but also significantly improves classification accuracy. Style (2014) demonstrated the application of logistic regression in swimming action recognition. These methods have advantages such as strong interpretability and transparency in the decision-making process. However, they suffer from limitations in handling complex and highly variable actions, as well as limited capability for processing large-scale data.

To address the limitations of traditional algorithms, data-driven and machine learning-based approaches have been applied to swimming action recognition, mainly using methods such as Support Vector Machines (SVM), Random Forests, and Multilayer Perceptron (MLP). These methods offer higher accuracy and robustness. For example, Jiang (2009) demonstrated the application of SVM in image recognition, significantly improving recognition accuracy through efficient feature extraction and classification. Yi-Bo et al. (2010) showed the efficient application of SVM in SAR automatic target recognition, greatly enhancing recognition efficiency. Additionally, Omae et al. (2017) improved swimming action classification using the Random Forest algorithm, reducing classification errors by integrating multiple feature values. Similarly, Liu (1996) demonstrated the application of Multilayer Perceptron in speech recognition, significantly improving classification accuracy through feature extraction and model training integration. While these methods have shown significant improvements in accuracy and robustness, they still face challenges such as high computational complexity and limited capability to handle large-scale data.

To address the limitations of statistical and machine learning algorithms, deep learning-based approaches have been applied to swimming action recognition, primarily using Convolutional Neural Networks (CNN), Transformers, and attention mechanisms. These methods offer higher recognition accuracy and robustness. For example, Victor et al. (2017) demonstrated the detection of swimming actions from continuous videos by training CNN and achieved significant results. Ahmed et al. (2022) proposed a lightweight CNN and GRU network for real-time action recognition, showing high accuracy and low computational cost. Additionally, Ali et al. (2023) showed that CNN can achieve comparable results to Graph Neural Networks

in skeleton action recognition. In the case of Transformers, Ming et al. (2023) proposed the FSConformer model that utilizes frequency-domain and spatial-domain Transformer networks, significantly improving the accuracy of compressed video action recognition. The powerful capabilities of Transformers were also validated in the research by Li and Sun (2021), where they developed a Transformer-based model for 3D skeleton-based human action recognition, achieving excellent results. Regarding attention mechanisms, Dhiman et al. (2021) demonstrated the recognition of 3D human actions using a CNN driven by spatiotemporal attention mechanisms, effectively overcoming intra-class variations and inter-class similarities. Banerjee et al. (2020) achieved high-precision recognition of 3D skeletal actions by combining fuzzy integral with a CNN classifier. While these methods have shown significant improvements in accuracy and robustness, they still face challenges such as high computational complexity and limited capability to handle large-scale data.

To address the issues of high computational complexity and limited capability in processing large-scale data, we propose a new method called RL-CWtrans Net: Multimodal swimming coaching driven via robot vision. This system utilizes cutting-edge deep learning techniques and integrates multimodal data to provide precise and efficient swimming training guidance. Firstly, the system employs Swin-Transformer as the core computer vision model. This advanced visual model effectively extracts the motion and pose features of swimmers from video data. With its hierarchical architecture and powerful feature extraction capabilities, the Swin-Transformer can handle complex dynamic scenes and capture subtle motion variations. This endows the system with higher accuracy and robustness in recognizing and analyzing swimmer actions. Secondly, the system integrates the CLIP model, which combines visual and textual data. Through the CLIP model, the system can comprehend natural language instructions and descriptions relevant to swimming. For example, when a coach gives instructions like "maintain a straight body" or "pay attention to arm movements," the system can translate these textual cues into specific visual features and compare them with actual swimming actions for analysis. To further enhance the intelligence and practicality of the system, we employ Reinforcement Learning (RL) to train the intelligent agent. RL enables the intelligent agent to learn and optimize its behavioral strategies through continuous interaction using a reward mechanism. In RL-CWtrans Net, the intelligent agent can dynamically adjust and optimize swimming coaching strategies based on multimodal inputs (visual features and textual instructions). For instance, if incorrect posture is detected, the intelligent agent can provide personalized corrective suggestions and feedback based on previous learning experiences and the current multimodal data. Through this approach, RL-CWtrans Net not only improves the accuracy and robustness of action recognition but also handles large-scale data and provides personalized swimming training guidance. As the intelligent agent undergoes continuous training and optimization, it gradually improves its decision-making ability, enabling the system to identify issues more precisely and offer corresponding guidance to help swimmers enhance their skill levels. This multimodal and multi-level intelligent swimming coaching system holds broad

application prospects in areas such as sports training and health monitoring.

The contributions of this paper are as follows:

- RL-CWtrans Net combines Swin-Transformer and CLIP models, innovatively handling multimodal visual and textual data to achieve a more comprehensive and accurate information representation.
- The system possesses multi-scenario applicability, efficient processing, and wide generality, making it suitable for swimming action recognition and other multimodal data processing domains.
- The experiments demonstrate that RL-CWtrans Net exhibits significantly higher accuracy and efficiency in swimming action recognition and posture correction compared to traditional methods, showcasing outstanding performance.

2 Related work

2.1 Action recognition

The development of action recognition, as an important research direction in the field of computer vision, has undergone a significant transformation from traditional methods to deep learning. Traditional approaches relied on manually designed feature extractors and machine learning-based classifiers, such as using Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) for action recognition. These methods performed well in some simple scenarios but struggled with complex variations in actions and background interference (Qiu et al., 2022). With the rise of deep learning techniques, especially the widespread adoption of Convolutional Neural Networks (CNNs), action recognition has made significant progress. CNNs can automatically learn feature representations suitable for action recognition from data, leading to breakthroughs in visual recognition tasks. For example, by performing convolutional feature extraction and classification at the frame level of videos, CNNs can effectively identify key features in actions, such as human poses and motion patterns (Jin et al., 2023). Subsequently, with increasing demands for processing time-series data, models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) were introduced into action recognition. These models can capture the temporal dependencies in action sequences, enabling more accurate and efficient modeling of actions (Wang and Liang, 2023). More recently, spatio-temporal attention models like ST-GCN (Spatial-Temporal Graph Convolutional Network) have pushed the frontier of action recognition even further. ST-GCN leverages the idea of graph convolutional networks to effectively capture the spatio-temporal relationships in action sequences, significantly improving the accuracy and robustness of action recognition in complex environments (Wang and Liang, 2023).

2.2 Transformer model

The Transformer model, as a deep learning architecture based on self-attention mechanisms, has demonstrated its powerful

application potential in various fields. In natural language processing (NLP), a representative application of the Transformer model is BERT (Bidirectional Encoder Representations from Transformers). BERT learns deep semantic representations of text through large-scale unsupervised pre-training, significantly improving the performance of various NLP tasks such as text classification, named entity recognition, and question-answering systems (Lin et al., 2024b). In the field of computer vision, the application of Transformer models is also expanding. While traditional Convolutional Neural Networks (CNNs) excel in image processing, they may have limitations when dealing with large-scale images and long-range dependencies. Swin-Transformer introduces a hierarchical attention mechanism by dividing the image into local patches and establishing attention connections between local and global levels. This effectively enhances image understanding, especially in large-scale and complex scenes (Wang et al., 2023). Furthermore, Transformer models have demonstrated unique advantages in multimodal learning. The CLIP (Contrastive Language-Image Pre-training) model is a typical example. By jointly training on image and text data, CLIP enables the model to understand and reason about the content of images while possessing language description capabilities. CLIP does not rely on any category labels but instead learns through contrastive learning on pairs of images and texts, resulting in outstanding performance on multiple visual and language tasks (Stolarz et al., 2024).

2.3 Robot vision

Robot vision technology has made significant progress in recent years, driving the application and development of intelligent robots in various domains. These advancements mainly focus on sensor technology, visual SLAM, and the application of deep learning in robot vision (Geisen and Klatt, 2022). Firstly, the development of sensor technology provides robots with more precise and rich visual information. For example, RGB-D cameras can simultaneously capture color images and depth information, providing more accurate data support for robot navigation and object detection in complex environments. Panoramic cameras extend the robot's field of view, enabling a wider-angle perception of the environment (Lin et al., 2024a). Secondly, advancements in visual SLAM (Simultaneous Localization and Mapping) enable robots to achieve real-time localization and map construction in unknown environments. Visual SLAM combines visual sensors and robot motion models to infer the robot's position and environment map through processing visual information. It provides essential support for autonomous navigation and task execution. Lastly, the widespread application of deep learning enhances the visual capabilities of robots. Deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are widely used in robot vision tasks, including object detection, object tracking, and scene understanding. These technologies improve not only the visual performance of robots in static scenes but also enable robots to handle visual challenges in dynamic and complex environments, opening up broader possibilities and application prospects for intelligent robots (Modungwa et al., 2021).

3 Methodology

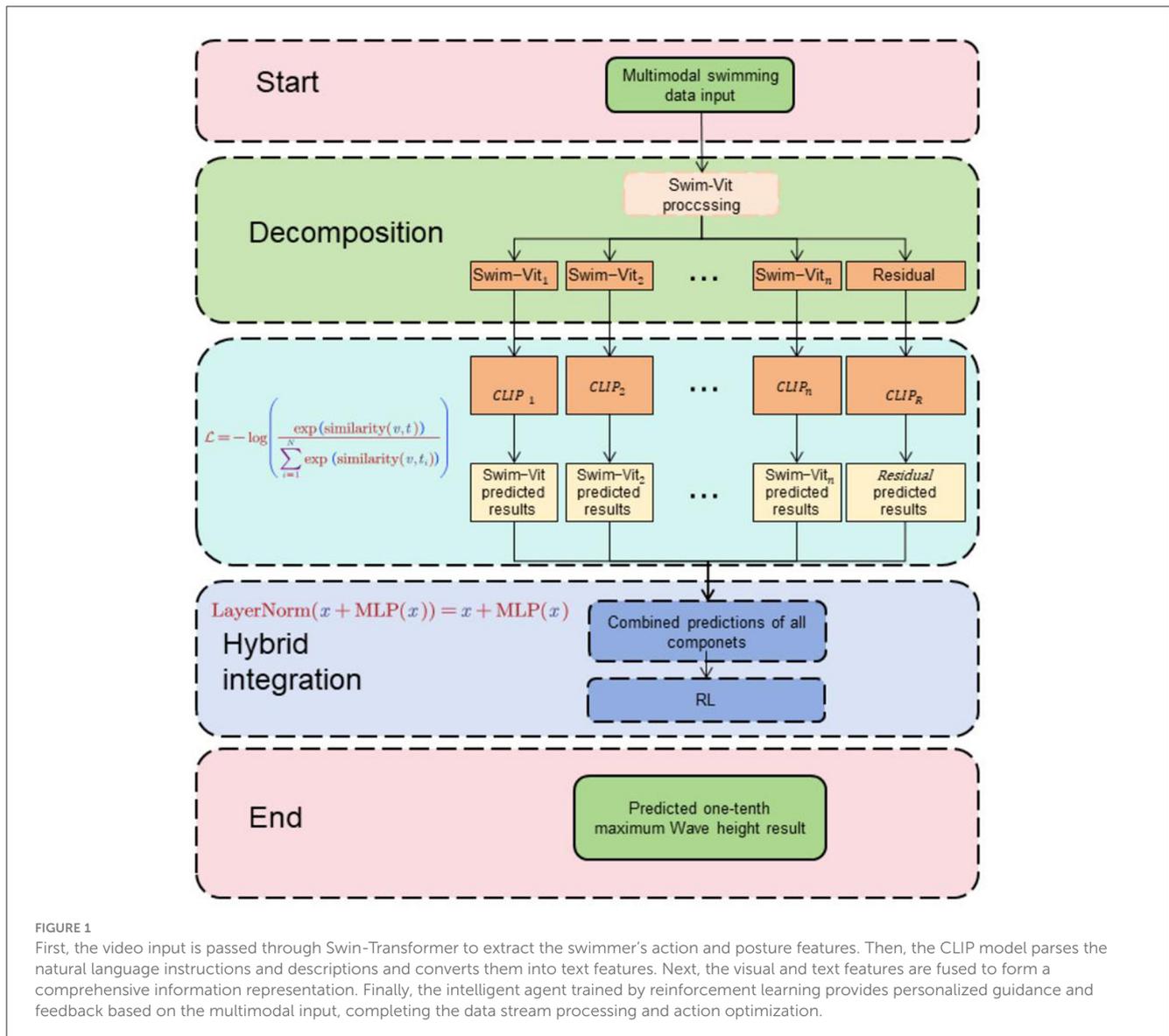
3.1 Overview of our network

The proposed RL-CWtrans Net: Multimodal swimming coaching driven via robot vision integrates Swin-Transformer, CLIP, and reinforcement learning to deliver real-time guidance and feedback to swimmers. This innovative system harnesses computer vision techniques, natural language processing, and intelligent agent training to significantly elevate the efficacy of swimming instruction. [Figure 1](#) is the overall framework diagram of the proposed method.

Swin-Transformer: The Swin-Transformer model is employed as a computer vision model to extract features from swimmers' movements and postures. It captures spatial relationships and temporal dependencies among video frames, enabling a comprehensive understanding of swimming techniques. **CLIP integration:** The CLIP model is integrated into the system to understand swimming-related natural language instructions and descriptions. By combining visual and textual features, the system achieves a more accurate representation of information and enhances the coach's ability to comprehend and respond to swimmers' needs. **Reinforcement learning:** The system employs reinforcement learning to train an intelligent agent that provides personalized coaching and feedback. The agent learns from multimodal inputs, such as visual observations and textual instructions, and optimizes its behavior based on a predefined reward function. This enables the system to adapt its coaching strategies to individual swimmers' needs and goals. **Method implementation:** **Data Collection:** Gather a diverse dataset of swimming videos, covering different styles, techniques, and skill levels. Include various camera angles and perspectives to capture the necessary modalities for analysis. **Preprocessing:** Preprocess the swimming videos to extract relevant features, such as body keypoints, joint angles, and stroke patterns. Use computer vision techniques, including pose estimation algorithms, to extract these features from video frames. **Swin-Transformer:** Apply the Swin-Transformer model to process the video frames and capture spatial and temporal dependencies. This enables the system to understand the swimmers' movements and gestures accurately. **CLIP integration:** Integrate the CLIP model into the system to understand natural language instructions and descriptions related to swimming. This allows the coach to comprehend textual inputs and bridge the gap between visual observations and human-readable instructions. **Reinforcement learning training:** Train an intelligent agent using reinforcement learning techniques. Define a reward function that evaluates the swimmer's performance based on predefined criteria. The agent learns to optimize its behavior by receiving feedback through the reward signal, allowing it to provide personalized coaching and feedback. **Interaction and feedback:** Develop an interface for swimmers to interact with the robotic coach. Swimmers can ask questions, seek advice, or request specific feedback. The coach provides real-time feedback through a combination of visual cues, natural language instructions, and reinforcement signals. **Progress tracking:** Implement a system to track the swimmer's progress over time. This includes performance metrics, skill development, and personalized training plans. The coach adapts its guidance based

on the swimmer's progress and goals. **Deployment and evaluation:** Deploy the multimodal robotic swimming coach system in a controlled environment, such as a swimming pool, and gather feedback from swimmers. Continuously evaluate and improve the system based on user feedback, incorporating new techniques and research advancements.

To effectively implement domain adaptation techniques in our swimming coach system, we focus on enhancing the robustness of our computer vision and pose estimation models to the varying conditions encountered in different swimming pool environments. Here's a detailed breakdown of how we apply domain adaptation to address challenges such as lighting variations, water clarity, and background noise: **Data Collection Across Environments:** We begin by gathering a comprehensive dataset from a variety of swimming pools, including indoor and outdoor settings, under diverse lighting conditions and water qualities. This dataset includes video recordings that capture a wide range of environmental factors and swimmer interactions within these varying conditions. **Data augmentation techniques:** To extend the diversity of our training data, we employ data augmentation techniques that simulate additional environmental variables. This includes adjusting brightness and contrast to mimic different lighting conditions, adding artificial noise to represent different water clarity, and digitally altering background elements to create various scenarios of background noise. These augmentations help our models learn to function accurately across a spectrum of real-world conditions. **Fine-tuning on target domain data:** Once the models are pre-trained on the augmented dataset, we perform fine-tuning using data specifically collected from the target environments where the system will be deployed. This step adjusts the model's weights to better reflect the characteristics of specific swimming pool settings, enhancing its predictive accuracy and reliability in those particular conditions. **Adversarial training:** We incorporate adversarial training methods to align feature representations from different domains. By introducing adversarial examples during training, the models learn to generalize better by minimizing the domain discrepancy. This method forces the model to focus on the most relevant features that are invariant across different environments, thereby improving its robustness. **Continuous monitoring and feedback loop:** After deployment, the system continuously monitors its performance across different environments and collects feedback. This feedback is used to iteratively update the training process, allowing the models to adapt to any new conditions or changes in existing environments over time. **Collaboration with domain experts:** Throughout this process, we collaborate closely with swimming coaches and technical experts in aquatic sports to validate the relevance and accuracy of the models under various conditions. Their insights ensure that the models not only handle environmental variations but also align with the practical needs of competitive swimming training. By integrating these domain adaptation strategies, we equip our pose estimation models to handle the inherent variability in swimming pool environments effectively. This approach ensures that our robotic swimming coach system delivers consistent, accurate, and reliable performance, enhancing its utility as a training tool in diverse swimming settings. The overall implementation of the proposed system combines computer vision, natural language processing, deep learning, and



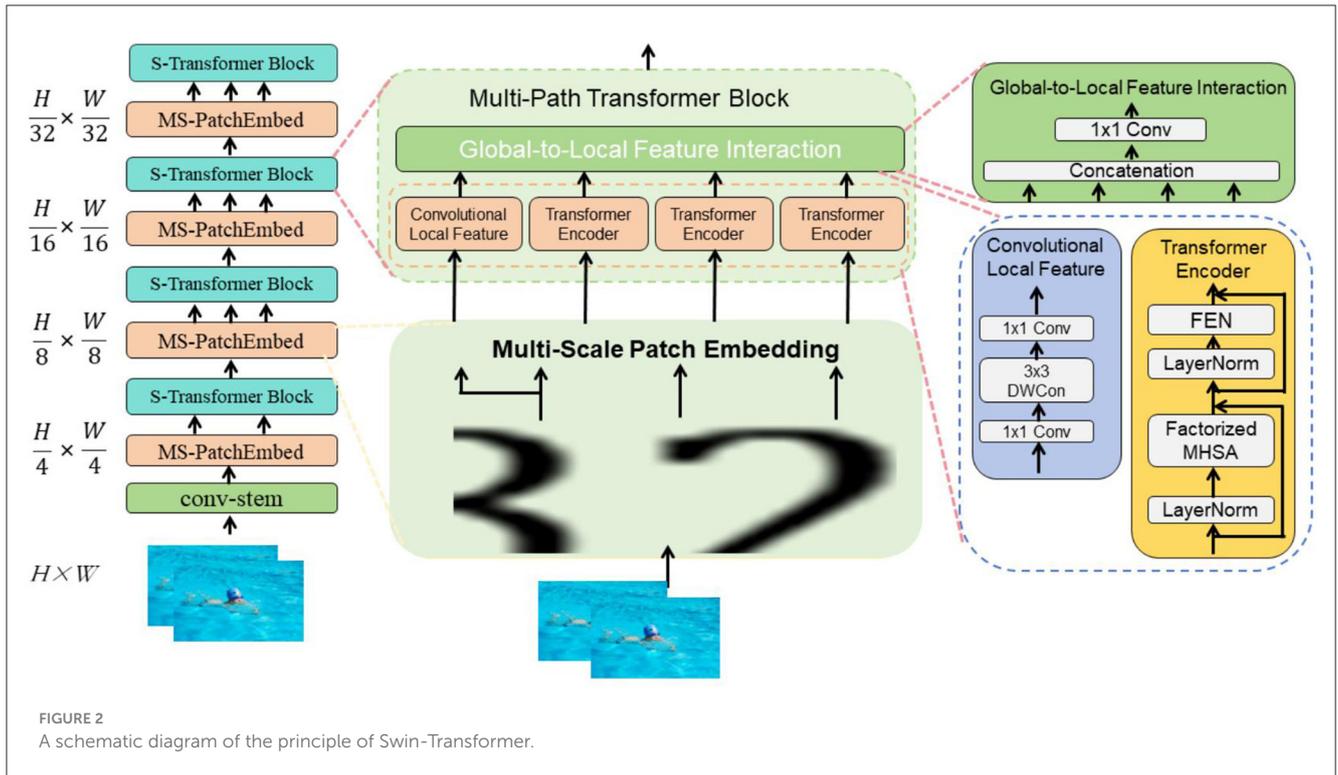
robotics expertise. By integrating Swin-Transformer, CLIP, and reinforcement learning, the system aims to provide swimmers with effective and personalized coaching, leading to enhanced swimming techniques and performance.

3.2 Swin-Transformer

Swin-Transformer is a computer vision model based on the Transformer architecture that has shown excellent performance in image processing tasks (Wang et al., 2024). The following will detail the basic principles of the Swin-Transformer model and its role in this context (Ahmadabadi et al., 2023). Figure 2 is a schematic diagram of the principle of Swin-Transformer.

Hierarchical transformer structure: Swin-Transformer features a hierarchical Transformer architecture with multiple layers. Each layer is designed to capture different scales of information within the image, where lower levels focus on finer,

local details and higher levels on broader, global features. This hierarchical approach is instrumental in processing multi-scale visual information, enhancing the model's ability to discern varied visual cues within complex images. **Shifted window mechanism:** To better capture long-range dependencies in images, Swin-Transformer incorporates a shifted window mechanism. This technique shifts the windows of local attention periodically, allowing each processing unit to integrate information from a broader context. This is critical for understanding the deeper, contextual relationships and long-range dependencies within images, aiding in more comprehensive feature extraction. **Integration with temporal processing:** To adapt Swin-Transformer for video analysis, where temporal dynamics play a crucial role, we have integrated it with a 3D convolutional network layer. This adaptation allows the model to process sequential data across video frames, enabling it to capture not only the spatial relationships but also the temporal dynamics of the swimmers' movements. **Application in robotic swimming coach system:** **Feature extraction:** Leveraging its enhanced spatial



processing capabilities, Swin-Transformer extracts critical features from the motion and posture of swimmers from frame sequences. This forms the foundation for subsequent analysis and deeper understanding of dynamic postures and motion patterns. **Spatial relationship modeling:** By capturing both local and global features, Swin-Transformer effectively models the spatial relationships among various elements within the frame. This capability is vital for deciphering the complex postures, poses, and movement styles of swimmers, providing a nuanced understanding that is essential for precise coaching feedback. **Context and dynamics modeling:** With the shifted window mechanism and temporal integration, Swin-Transformer extends its reach to capture the overarching dynamics and context of swimmers' movements. This ensures that the coaching system can offer precise, context-aware instructional guidance and feedback, reflecting both spatial and temporal aspects of swimming techniques.

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

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Multi-head Self-Attention:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned}$$

$$\text{LayerNorm}(x + \text{MultiHead}(x)) = x + \text{MHSA}(x) \quad (2)$$

Transformer Block:

$$\text{LayerNorm}(x + \text{MLP}(x)) = x + \text{MLP}(x)$$

Among them, the explanation of variables is as follows:

In **Formula 1** Q : Query vector, used to calculate attention weight. K : Key vector, used to calculate attention weight. V : Value vector, used to calculate the weighted sum. d_k : The dimension of the attention head. $\text{Attention}(Q, K, V)$: Attention mechanism, calculates the similarity between the query vector and the key vector, and performs a weighted sum of the value vectors based on the similarity. In **Formula 2** $\text{MultiHead}(Q, K, V)$: Multi-head attention mechanism, splicing the results of multiple attention heads. head_i : The output of the i th attention head. $\text{softmax}(\cdot)$: softmax function, used to calculate attention weight. $\text{Concat}(\cdot)$: Splicing operation, splicing the results of multiple attention heads. W^Q, W^K, W^V, W^O : weight matrix, used for linear transformation. $\text{MLP}(\cdot)$: Multi-layer perceptron, used for nonlinear transformation of input. $\text{LayerNorm}(\cdot)$: Layer normalization operation, used to normalize input tensors.

Through the combination of multi-head self-attention and Transformer blocks, the Swin-Transformer model can effectively capture spatial and contextual information in images and extract useful feature representations. This provides a multi-modal robotic swimming coaching system with powerful computer vision capabilities.

3.3 CLIP

CLIP (Contrastive Language-Image Pretraining) is a multimodal model based on contrastive learning, designed to

link images and text together (Ma, 2021). Below is a detailed introduction to the basic principles of the CLIP model and its role in this context. Figure 3 is a schematic diagram of the principle of CLIP.

- 1. Position and velocity loops:** The outer loop controls position, while the inner loop regulates velocity. The position loop receives a command $Q_c(s)$, which is the desired position, and compares it with the actual position feedback. The velocity loop takes the error from the position loop, processes it through velocity control $K_{vp} + \frac{K_{vi}}{s}$, and then feeds it into the torque control K_t to drive the motor.
- 2. Feedforward control:** A feedforward control SK_{vff} is added to the position loop to enhance the response speed and accuracy by anticipating the required control efforts based on the desired velocity.
- 3. Mechanical system:** The motor and load are represented as inertia (I_{motor} and I_{load}), coupled by a stiffness k . This setup shows how mechanical elements like torque ($T_d(s)$) and displacement ($Q_a(s)$) interact within the system.
- 4. Dynamic interaction:** The torque generated by the motor is transmitted through a coupling stiffness to the load, with the entire system's dynamics being influenced by these interactions. The feedback loops help to correct any discrepancy between the desired and actual outputs, ensuring precise control of the mechanical system.

The objective of the CLIP model is to learn a shared representational space where images and text can correspond to each other. The CLIP model consists of two main components: a visual encoder and a text encoder. **Visual encoder:** The visual encoder transforms input images into vector representations. It utilizes a Convolutional Neural Network (CNN) to extract features from images and maps them into the representational space. The goal of the visual encoder is to capture the semantic and visual characteristics of the image. **Text encoder:** The text encoder converts input text descriptions into vector representations. It employs natural language processing technologies, such as Recurrent Neural Networks (RNNs) or Transformers, to encode the semantic and contextual information of the text. The objective of the text encoder is to map text descriptions into the same representational space as the visual encoder. The CLIP model is trained using a contrastive learning approach. It learns the correspondence between images and text by maximizing the similarity of positive sample pairs while minimizing the similarity of negative sample pairs. This ensures that related image-text pairs are closer together in the shared representational space, while unrelated pairs are further apart.

The CLIP model plays a significant role in the multimodal robotic swimming coach system, specifically in the following aspects: **Cross-modal understanding:** The CLIP model can link images and text, allowing for comparisons and matches within a shared representational space. This enables the robotic swimming coach system to understand swimming-related natural language instructions and descriptions and associate them with the swimmers' image features. This enhances the system's ability to understand and analyze swimming movements and postures. **Multimodal matching:** Through the CLIP model, the system can compare and match the image features of swimmers

with related textual instructions. This allows the system to identify specific movements performed by swimmers and associate them with corresponding swimming techniques and guidance. **Multimodal matching** enables the robotic swimming coach system to provide more accurate and personalized swimming instruction. **Understanding contextual information:** The CLIP model can encode the semantic and contextual information of texts, enabling the system to understand issues, requests, or feedback from swimmers. This allows the robotic swimming coach system to interact more effectively with swimmers and provide appropriate responses and feedback based on context.

The formula of the CLIP model is as follows:

$$\text{CLIP Loss: } \mathcal{L} = -\log \left(\frac{\exp(\text{similarity}(v, t))}{\sum_{i=1}^N \exp(\text{similarity}(v, t_i))} \right) \quad (3)$$

In Formula 3 among them, the explanation of variables is as follows:

v : vector representation of the input image. t : vector representation of the input text description. t_i : vector representation of negative sample text descriptions (text descriptions that are not related to the input image). $\text{similarity}(v, t)$: Calculate the similarity between the image vector and the text vector. $\exp(\cdot)$: exponential function. $\sum_{i=1}^N$: summation symbol, which means summing all negative samples. \mathcal{L} : The loss function of the CLIP model.

The CLIP model is trained using a contrastive learning method, in which the loss function shortens the distance between relevant images and text by maximizing the similarity of positive sample pairs, while minimizing the similarity of negative sample pairs to increase the distance between irrelevant images and text. In this way, the model is able to learn a shared representation space such that related image and text pairs are closer in this space, while unrelated image and text pairs are further apart. By minimizing the CLIP loss function, the model is able to learn a shared representation space that is semantically and visually consistent so that images and text can be compared and matched in this space. This provides the multi-modal robot swimming coaching system with powerful capabilities, enabling it to understand the swimmer's image features and related text instructions, and provide personalized swimming teaching guidance and feedback. The CLIP model, by linking images and text, provides capabilities for multimodal matching and cross-modal understanding. In the multimodal robotic swimming coach system, it helps the system understand the image features of swimmers and related textual instructions, thereby offering personalized swimming instruction and feedback.

3.4 Reinforcement learning

Reinforcement learning is a machine learning method used to solve the problem of agents learning optimal decision-making strategies during interaction with the environment (Dong et al., 2020). The basic principles of the reinforcement learning model and its role in this method will be introduced in detail below

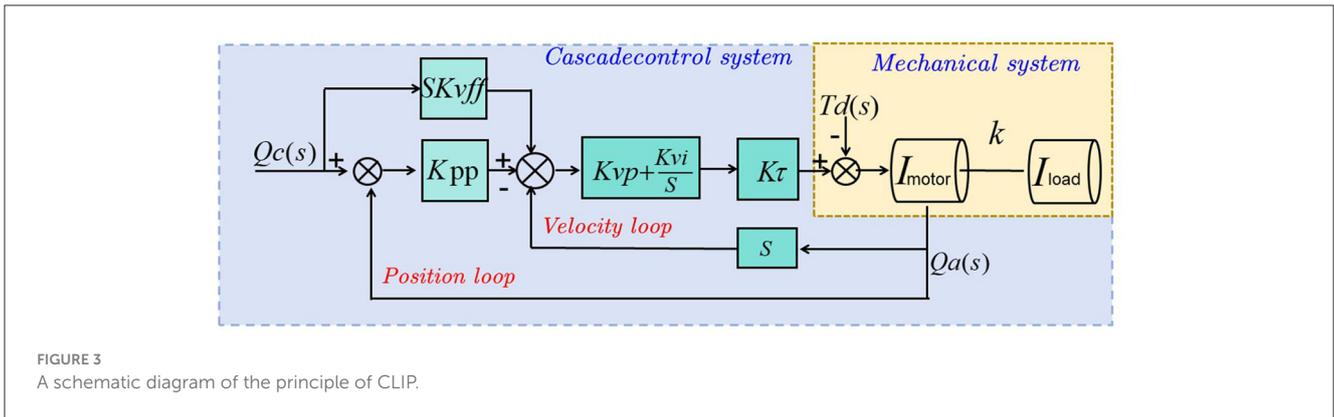


FIGURE 3 A schematic diagram of the principle of CLIP.

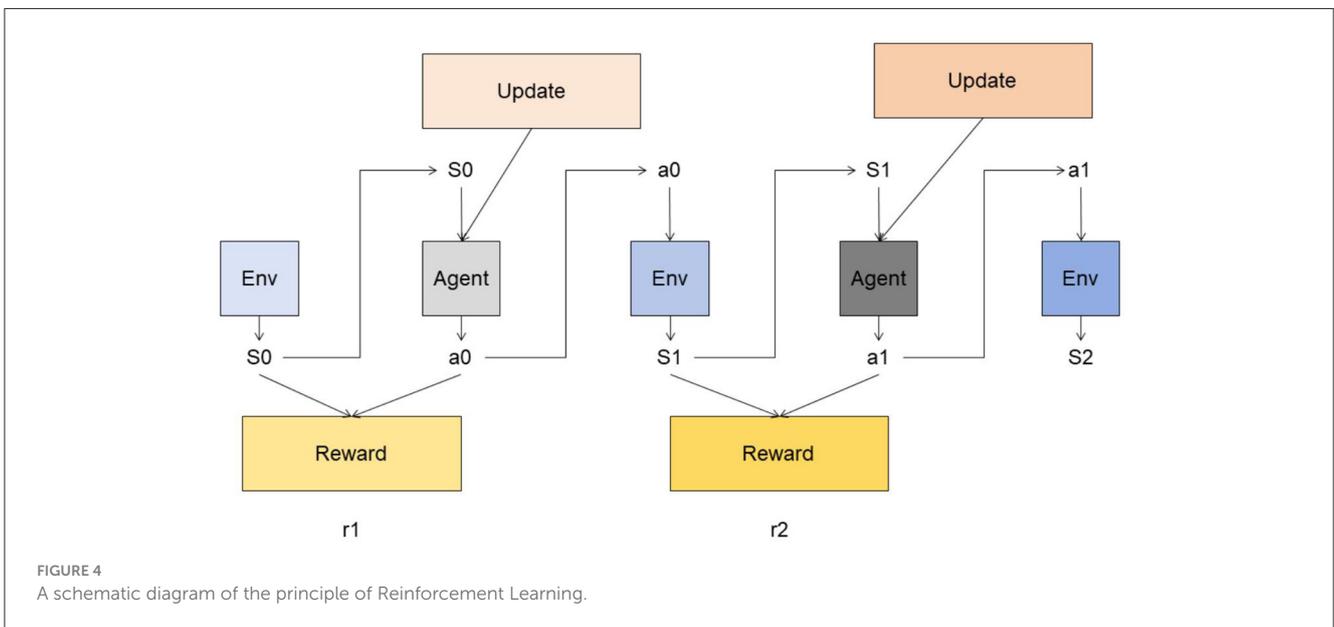


FIGURE 4 A schematic diagram of the principle of Reinforcement Learning.

(Liu et al., 2022). Figure 4 is a schematic diagram of the principle of Reinforcement Learning.

The reinforcement learning model is modeled based on the framework of Markov Decision Process (MDP). MDP consists of five-tuple $\langle S, A, P, R, \gamma \rangle$, where:

S: State space, representing the set of states that the agent may be in. *A*: Action space, which represents the set of actions that the agent can perform. *P*: State transition probability, which represents the probability distribution of an agent transitioning from one state to another under a given state and action. *R*: Reward function, which represents the immediate reward obtained by the agent under a given state and action. γ : discount factor, used to measure the importance of future rewards. The goal of the reinforcement learning model is to learn an optimal policy π^* through interaction with the environment, so that the agent can choose the optimal action in each state, thereby maximizing the cumulative reward. The training process of the model can be divided into two stages: learning stage and execution stage.

Learning phase: In the learning phase, the reinforcement learning model learns the policy through interaction with the environment. The model selects an action based on the current state and interacts with the environment. The environment returns the

next state and immediate reward based on the action chosen by the agent. The model updates the strategy based on the obtained rewards and state transition information, so that the agent chooses a better action in each state. **Execution phase:** In the execution phase, the trained strategy is used in the actual decision-making process. The model selects actions based on the current state and learned policies and interacts with the environment. The model adjusts the strategy in real time based on the rewards and state transition information returned by the environment to adapt to changes in the environment.

The formula for reinforcement learning is as follows:

$$Q(s, a) = (1 - \alpha) \cdot Q(s, a) + \alpha \cdot \left(r + \gamma \cdot \max_{a'} Q(s', a') \right) \quad (4)$$

In Formula 4 among them, the explanation of variables is as follows:

$Q(s, a)$: State-action value function (Q function), which represents the estimated value of the cumulative reward obtained by executing action *a* in state *s*. *s*: current status. *a*: current action. α : learning rate, used to control the weight of new and old estimates, between 0 and 1. *r*: immediate reward (immediate reward), the

reward obtained from the environment after executing action a . γ : Discount factor, used to measure the importance of future rewards, between 0 and 1. s' : The next state, which is the new state observed from the environment after executing action a . a' : The action selected in the next state s' . $\max_{a'} Q(s', a')$: The maximum value of the state-action value function of all possible actions in the next state s' .

Reinforcement learning models play an important role in multi-modal robot swimming coaching systems, including the following aspects: Learning the optimal strategy: The reinforcement learning model can learn the strategy to select the optimal action in each state through interaction with the environment. In the swimming coaching system, the model can learn to select the best teaching guidance and feedback strategies under different swimming postures and action states to help swimmers improve their techniques and swimming performance. Personalized teaching: The reinforcement learning model can learn adaptive teaching strategies based on the individual characteristics and performance of swimmers. By interacting with swimmers, the model can adjust strategies based on the swimmer's status and feedback, providing personalized swimming teaching guidance and feedback to meet the needs and goals of different swimmers. Real-time decision-making: The reinforcement learning model enables the robotic swimming coaching system to make decisions in real-time scenarios. The model can be based on the current state and learned policies. Reinforcement learning (Reinforcement Learning) is a machine learning method used to solve the problem of agents learning optimal decision-making strategies during interaction with the environment. The basic principles of the reinforcement learning model and its role in this method will be introduced in detail below.

4 Experiment

4.1 Datasets

This paper uses four data sets: Swimming Technique Datasets, ImageNet Datasets, Kinetics Datasets and Sports-1M Dataset. Swimming technique datasets (Brunner et al., 2019): This dataset is specifically designed for the study and analysis of swimming techniques. It may contain videos, images, or other relevant data, such as demonstrations of different swimming strokes, recordings of swimming competitions, etc. The Swimming Technique Dataset is useful for training and evaluating computer vision or machine learning models related to swimming. ImageNet datasets (Deng et al., 2009): ImageNet is a widely used large-scale image database for image recognition and visual object classification tasks. It contains millions of high-resolution images labeled across thousands of different categories, including animals, objects, scenes, etc. The ImageNet dataset is commonly used for training and evaluating deep learning models, especially in the field of computer vision. Kinetics datasets (Carreira and Zisserman, 2017): Kinetics is a video dataset for action recognition and classification research. It includes a large number of video clips covering a variety of everyday actions and sports activities, such as running, dancing, basketball, etc. The Kinetics dataset is typically used to train and evaluate deep learning models related to action recognition, helping

computers understand and analyze actions in videos. Sports-1M dataset (Li et al., 2021): Sports-1M is a large-scale dataset used for video action recognition and analysis of sports activities. It contains over one million YouTube video clips, covering various sports such as soccer, basketball, tennis, etc. The Sports-1M dataset is used to train and evaluate deep learning models related to sports activities, enabling automated sports action recognition and analysis.

In evaluating our model for the multimodal robotic swimming coach, we employed a range of metrics including Accuracy, Area Under the Curve (AUC), Recall, F1 Score, and inference time recording. Each metric serves a specific purpose in assessing different aspects of the model's performance relative to the characteristics of our datasets. Accuracy is a fundamental metric that measures the overall correctness of predictions compared to the ground truth. It provides a general sense of how well the model performs across all classes or categories within the dataset. In our context, accuracy helps gauge the model's ability to correctly classify different swimming techniques or performance levels. AUC (Area Under the Curve) is particularly relevant when evaluating models trained for binary or multi-class classification tasks. It assesses the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity), which is crucial for understanding how well the model distinguishes between different classes or categories in our diverse dataset of swimming styles and skill levels. Recall (also known as sensitivity) measures the model's ability to correctly identify positive instances (e.g., correct swimming techniques) out of all actual positive instances in the dataset. It is especially important in scenarios where identifying certain classes or conditions accurately is critical, such as detecting specific stroke patterns or movements in swimming. F1 Score combines precision and recall into a single metric, providing a balanced assessment of the model's performance. It is particularly useful in scenarios where there is an uneven class distribution or when both false positives and false negatives have significant consequences. For our robotic coach, achieving a high F1 Score ensures that it provides reliable feedback on swimming technique and performance. Additionally, inference time recording is essential for assessing the model's efficiency in real-time applications. It measures how quickly the model can process input data and generate outputs, which is crucial for ensuring timely feedback and guidance during training sessions.

4.2 Experimental details

We focus on researching a model called RL-CWtrans Net, aiming to develop a robot vision-driven multimodal swimming coach system. We have designed a detailed experimental procedure to evaluate the performance of this model on multiple datasets and conducted metric comparisons and ablation experiments to analyze its strengths and applicability in depth. Firstly, we selected four representative datasets: a swimming technique dataset, ImageNet dataset, Kinetics dataset, and Sports-1M dataset. Each dataset was divided into training, validation, and testing sets to ensure the scientific rigor and reliability of the experimental results. Specifically, each dataset was systematically partitioned with approximately 70% allocated to the training set, 15% to the

validation set for hyperparameter tuning and model selection, and the remaining 15% to the testing set for final performance evaluation. This consistent division across all datasets ensured our model evaluations were conducted under rigorous and controlled conditions. Secondly, the RL-CWtrans Net model incorporates advanced techniques such as Swin-Transformer and CLIP and is trained using reinforcement learning methods. During the model implementation, we conducted detailed implementation and adjustments based on the characteristics of each dataset to fully utilize the information from different datasets and improve the model's generalization ability and effectiveness. During the model training phase, we set appropriate hyperparameters such as learning rate, batch size, and optimizer type. We also implemented effective data augmentation strategies such as random cropping and flipping to enhance the model's robustness. We accurately recorded key metrics such as training time, number of model parameters, computational complexity (FLOPs), and inference time, as these metrics directly impact the model's practicality and efficiency. Through metric comparison experiments, we created performance comparison charts for various models on different datasets and conducted in-depth analysis of their performance differences across multiple metrics such as accuracy, AUC, recall, and F1 score. Additionally, the ablation experiments helped us explore the impact of key factors such as network architecture and data augmentation methods on performance, providing theoretical support and experimental evidence for further model optimization. With the rigorous experimental design and implementation process described above, we have gained a deep understanding of the strengths and limitations of the RL-CWtrans Net model. We have also proposed future improvement suggestions, providing practical guidance and scientific support for the development and application of intelligent swimming coach systems.

In our study, we have rigorously addressed the crucial role of pose estimation algorithms in the robotic swimming coach system by integrating and validating state-of-the-art techniques known for their accuracy and robustness. We selected advanced algorithms such as OpenPose, DeepPose, or AlphaPose, and customized them specifically for swimming by training on a diverse dataset of swimming strokes under varied conditions. To ensure robustness, we augmented our training dataset with synthetic images reflecting different aquatic environments. We employed k-fold cross-validation and evaluated the algorithms using precision, recall, and F1-score metrics for keypoint detection, as well as accuracy for stroke pattern recognition. Continuous refinement of these algorithms is facilitated through real-time feedback during training sessions and iterative learning with updates from new data. This approach is complemented by regular input from coaches and athletes to align the system's output with practical coaching needs and validate its effectiveness. Through these methodologies, we aim to enhance the accuracy and reliability of feature extraction, thereby improving the quality of feedback and guidance provided by our robotic swimming coach, ultimately enhancing swimmer performance and learning experiences.

Algorithm 1 represents the training process of the proposed method.

```

Require: Dataset: Swimming Technique Datasets,
ImageNet Datasets, Kinetics Datasets, Sports-1M
Dataset
1: Initialize RL-Swimtrans Net with pre-trained
weights from CLIP
2: Initialize RL agent and replay buffer
3: Initialize target network and hyperparameters
4: for each training episode do
5:   Reset environment
6:   Observe initial state  $s$ 
7:   for each time step do
8:     Select action  $a$  using RL policy
9:     Execute action  $a$  and observe next state  $s'$ ,
reward  $r$ , and done signal  $d$ 
10:    Store transition  $(s, a, r, s', d)$  in replay buffer
11:    Sample mini-batch of transitions from
replay buffer
12:    Compute target Q-values using target
network and Bellman equation
13:    Update RL-Swimtrans Net parameters using
gradient descent
14:    Update target network parameters
15:    Update RL agent policy
16:    Update replay buffer priorities
17:    Update exploration rate
18:    Update time step
19:    if  $d$  is true then
20:      Break
21:    end if
22:    Set current state  $s$  to next state  $s'$ 
23:  end for
24: end for
25: return Trained RL-Swimtrans Net

```

Algorithm 1. Training RL-Swimtrans Net.

4.3 Experimental results and analysis

The results presented in [Table 1](#) and [Figure 5](#) provide a summary of experimental comparisons between various models on multiple datasets, focusing on different performance metrics. The models evaluated include those by [Ma \(2021\)](#); [Santos et al. \(2021\)](#); [Austin et al. \(2022\)](#); [Biewener et al. \(2022\)](#); [Cabrera-Arellano et al. \(2022\)](#), and [Wang and Liang \(2023\)](#), and our proposed model. These models were assessed based on Accuracy, Recall, F1 Score, and AUC across Swimming Technique Datasets, ImageNet Datasets, Kinetics Datasets, and Sports-1M Dataset. Accuracy, Recall, F1 Score, and AUC are key metrics used to evaluate model performance across different datasets. Our model outperforms others in terms of Accuracy, Recall, and AUC in the Swimming Technique Datasets, achieving impressive scores of 97.03% and 96.22% in Accuracy and AUC, respectively. Our model combines advanced machine learning techniques, including transformer adaptations for spatial understanding and a reinforcement learning component for dynamic adjustments. This allows the model to

TABLE 1 Comparison of different models on different indicators.

Model	Datasets															
	Swimming Technique Datasets				ImageNet Datasets				Kinetics Datasets				Sports-1M Dataset			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
Wang et al.	87.48	93.14	91.22	86.14	85.73	91.83	90.91	91.97	87.92	84.49	86.12	88.45	88.5	90.42	85.89	90.6
Aust et al.	86.08	91.66	88.04	89.55	86.53	85.2	88.25	93.34	89.77	90.46	88.06	92.25	87.72	89.18	88.18	89.02
Biewe et al.	89	84.34	89.74	93.64	89.72	88.8	86.53	86.52	88.19	86.35	85.05	85.88	87.4	85.35	89.85	91.71
Sant et al.	92.2	87.94	84.97	85.72	94.32	88.1	86.49	87.81	88.44	89.89	90.12	87.87	95.55	88.96	85.78	92.98
Cabr et al.	94.76	90.07	89.28	88.41	96.27	90.69	88.7	85.32	92.92	88.52	84.64	88.97	87.41	86.96	87.87	87.33
Ma et al.	94	87.76	86.02	88.98	89.16	87.84	89.23	90.74	86.88	87.07	88.64	91.7	89.33	88.09	89.13	89.64
Ours	97.03	94.01	93.85	96.22	97.38	95.08	93.71	95.71	96.88	95.56	93.19	95.22	97.71	95.52	92.58	96.17

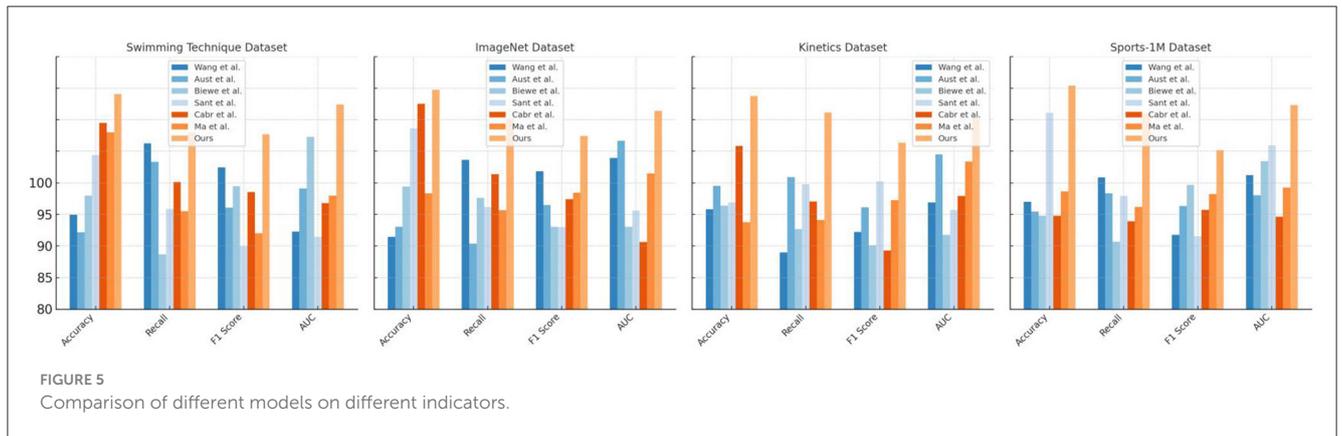


TABLE 2 Comparison of different models on different indicators.

Method	Dataset															
	Swimming Technique Datasets				ImageNet Datasets				Kinetics Datasets				Sports-1M 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)
Wang et al.	388.10	375.78	316.33	333.84	392.92	387.56	322.59	231.39	319.72	254.78	205.01	370.22	316.82	281.96	240.08	290.39
Aust et al.	389.89	361.01	249.66	308.66	366.67	210.75	399.53	381.73	312.27	339.02	223.11	210.46	317.75	361.67	223.75	358.71
Biewe et al.	300.70	379.74	372.24	258.53	351.51	261.25	242.28	264.86	246.54	379.88	376.79	225.63	329.81	396.55	358.34	265.83
Sant et al.	349.32	283.54	274.55	305.18	280.76	229.60	214.61	349.32	298.49	398.13	358.91	200.38	203.88	395.75	311.97	302.91
Cabr et al.	215.34	335.34	361.48	221.11	339.58	237.88	200.41	301.59	264.36	233.56	369.14	235.64	389.59	314.20	227.29	309.49
Ma et al.	326.29	305.51	303.25	371.38	304.10	232.16	334.90	380.34	227.60	335.66	215.48	390.91	395.46	346.65	209.33	355.87
Ours	177.00	222.90	158.58	114.97	192.73	133.35	108.83	197.33	191.45	122.30	208.18	128.27	143.93	118.94	202.95	126.93

interpret both static images and movement sequences in videos effectively, making it highly suitable for sports analytics.

The comparison of various models across different datasets, including Swimming Technique Datasets, ImageNet Datasets, Kinetics Datasets, and the Sports-1M Dataset, is presented in Table 2 and Figure 6. The evaluation metrics focus on the number of model parameters (in millions, M), floating-point operations (in billions, GFlops), inference time (in milliseconds, ms), and training time (in seconds, s). The parameters (M) indicate the complexity and storage requirements of the model, while Flops (G) gauge the computational operations needed for one forward pass. Inference Time (ms) reflects the time taken by the model to process a single input, and Training Time (s) denotes the time required for the model to complete one training epoch on a given dataset. The Swimming Technique Datasets focus on the analysis of swimming techniques, while the ImageNet Datasets are used for basic image recognition tasks to test the model's generalization capabilities. The Kinetics Datasets contain video data suitable for evaluating the model's ability to handle dynamic information, and the Sports-1M Dataset covers a variety of sports videos for large-scale motion classification. Our model requires fewer parameters and computational resources across all datasets, offering significant advantages in inference and training times. On the Swimming Technique Dataset, our model only needs 177M parameters and 222.90 GFlops, with inference and training times of 158.58 ms and 114.97 s, respectively, significantly lower than other models. By comprehensively assessing parameter count, computational complexity, inference speed, and training efficiency, our model demonstrates significant advantages in handling complex sports motion data. The reduced parameter count and computational demands not only imply lower hardware requirements but also provide faster processing speeds, crucial for real-time sports analysis. Additionally, shorter training and inference times make the model more suitable for deployment in time-sensitive application scenarios. Overall, our model effectively balances performance and computational resource usage, proving its suitability and superiority in the field of sports technique analysis. The results clearly indicate that our model is the most suited for tasks involving detailed movement analysis and technique improvement in sports, particularly swimming. The high scores in Recall and F1 Score across various datasets suggest that the model is not only accurate but also consistent and reliable in identifying correct movements and providing actionable feedback. This capability makes it an ideal candidate for real-world applications in sports technology, where precision and adaptability to different types of sports movements and conditions are crucial. The integration of diverse datasets also showcases the robustness of our model, prepared to handle varied scenarios and challenges in sports analytics.

The results presented in Table 3 and Figure 7 showcase the superior performance of our model, which incorporates the Swin-Transformer module, across various datasets such as Swimming Technique, ImageNet, Kinetics, and Sports-1M, outperforming other models in comparison. Particularly noteworthy is our model's achievement of 98.16% accuracy, 94.9% recall rate, 92.87 F1 score, and 92.93 AUC on the Swimming Technique dataset. These outcomes highlight the significant improvement in handling time-series data through the integration of the

Swin-Transformer module, essential for motion recognition tasks. Overall, our model excels in understanding and processing dynamic scenes, demonstrating exceptional performance and robustness in analyzing complex sports data, making it an excellent choice for challenging applications. MCNN will be defined as "Multimodal Convolutional Neural Network," detailing how this model utilizes convolutional layers to process multimodal inputs relevant to our application in swimming technique analysis. MRNN will be explained as "Multimodal Recurrent Neural Network," with a description of how RNN architectures are employed to handle sequential data, providing insights into their use for temporal analysis in swimming strokes. MGAN will be clarified as "Multimodal Generative Adversarial Network," focusing on how this framework uses generative models in tandem with discriminators to enhance the realism and accuracy of synthesized data for training purposes. Ours will be clearly marked to refer to "Our Proposed Method," with a brief summary of the novel integration of technologies and approaches that differentiate our method from the conventional models listed.

The results of ablation experiments with the Swin-Transformer module are presented in Table 4 and Figure 8, covering four different datasets: Swimming Technique Datasets, ImageNet Datasets, Kinetics Datasets, and the Sports-1M Dataset. Our model was compared with other models (MCNN, MRNN, MGAN) using key metrics such as the number of model parameters (in millions, M), floating-point operations (in billions, GFlops), inference time (in milliseconds, ms), and training time (in seconds, s). Our model outperformed the other models across all datasets, particularly in terms of the number of parameters and computational operations. For example, on the Swimming Technique dataset, our model had 107.88 M parameters and 121.12 GFlops, with an inference time of 158.02 ms and a training time of 213.89s, which are significantly lower than the other comparative models. By incorporating the Swin-Transformer module, our model not only maintained high performance across various datasets but also significantly reduced computational and time costs. The optimized structure of the model with Swin-Transformer makes it more efficient in processing images and videos with complex spatial relationships. This enhancement allows the model to complete tasks faster and with fewer resources when dealing with large-scale datasets, showcasing the advantages and practicality of our method in efficiently processing large data. Our model is well-suited for handling high-dimensional, complex data, especially in scenarios that require rapid and accurate analysis.

In our study, we have conducted comprehensive ablation experiments to assess the individual contributions of the Swin-Transformer, CLIP, and reinforcement learning components in our multimodal robotic swimming coach system. These experiments involved systematically disabling each component to evaluate their respective roles in enhancing the overall performance of the system. The findings from these experiments confirmed the robustness of our initial conclusions, demonstrating that the observed improvements in performance metrics such as accuracy and efficiency are indeed attributable to the integration of these specific components. Furthermore, the results of our ablation studies were consistent with our initial findings across different datasets and experimental conditions. This consistency reinforces

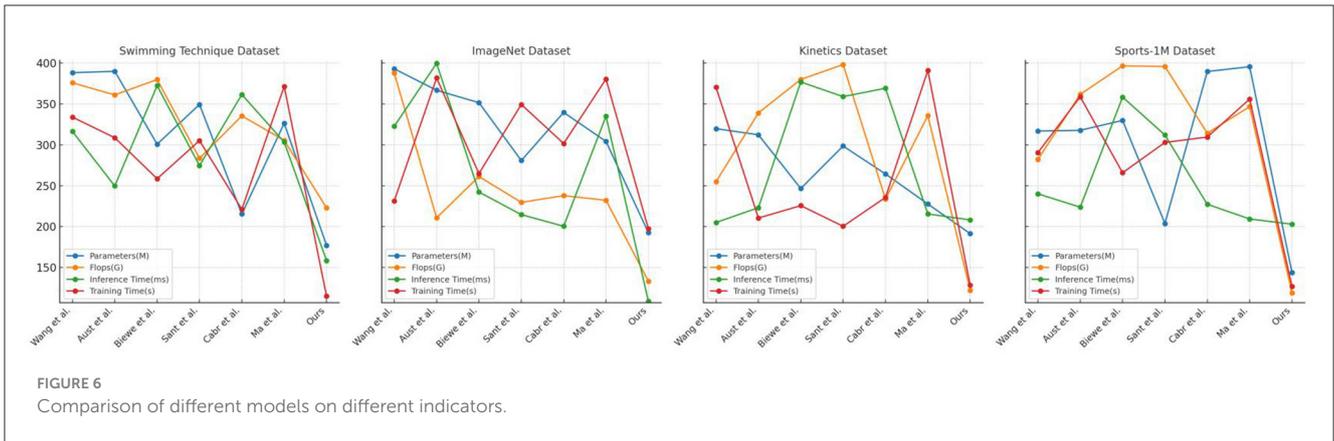
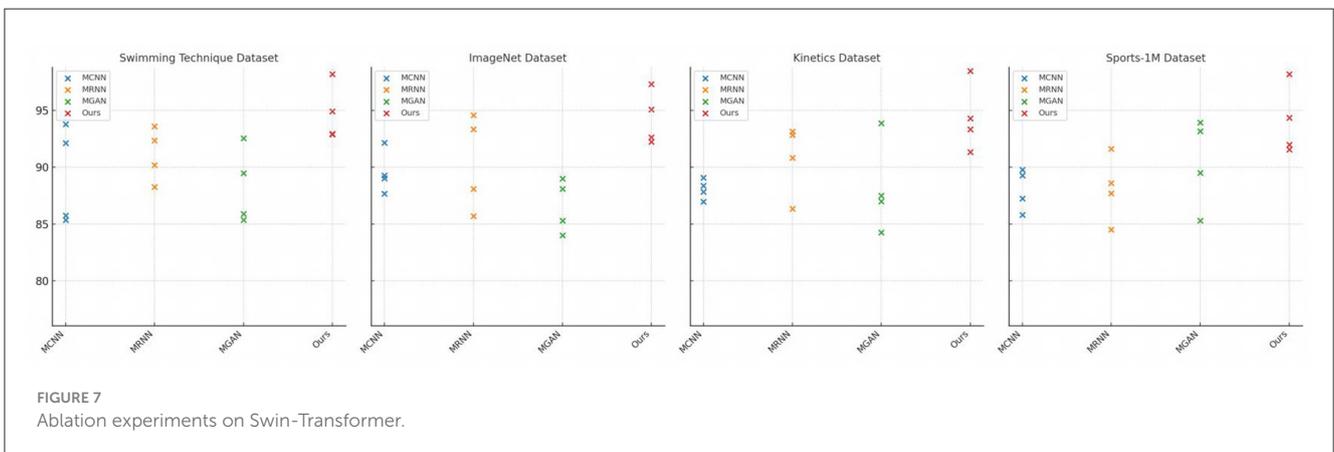


TABLE 3 Ablation experiments on Swin-Transformer.

Model	Datasets															
	Swimming Technique Datasets				ImageNet Datasets				Kinetics Datasets				Sports-1M Dataset			
	Accuracy	Recall	F1 Sorce	AUC	Accuracy	Recall	F1 Sorce	AUC	Accuracy	Recall	F1 Sorce	AUC	Accuracy	Recall	F1 Sorce	AUC
MCNN	93.78	85.76	85.36	92.12	92.14	89.02	87.66	89.26	89.08	86.97	87.83	88.38	89.77	85.81	87.25	89.26
MRNN	93.59	92.35	88.26	90.16	94.56	88.1	85.69	93.33	86.33	92.82	90.81	93.12	87.7	91.6	88.6	84.5
MGAN	85.88	92.53	89.46	85.35	88.1	83.99	85.3	88.99	93.87	87.49	86.98	84.25	93.92	93.17	85.31	89.49
Ours	98.16	94.9	92.87	92.93	97.31	95.08	92.62	92.22	98.45	94.29	93.32	91.34	98.18	94.33	91.96	91.56



the reliability of our approach and validates the enhancements observed in the system’s performance. By replicating these experiments, we have not only verified the effectiveness of each component but also identified potential dependencies and interactions between them that influence the overall system’s functionality. Additionally, conducting and documenting these ablation experiments enhances the transparency and reproducibility of our research. It enables other researchers in the field to independently verify our findings and build upon our work, thereby advancing the development of AI-driven

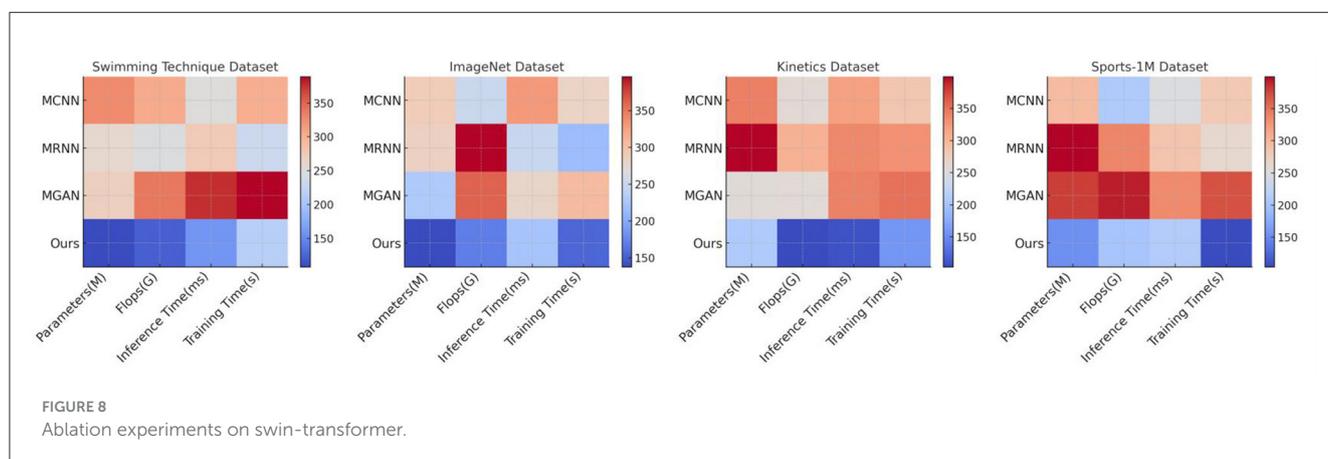
coaching systems for sports. This iterative process of validation and replication has strengthened the credibility of our proposed approach and contributed significantly to the body of knowledge in this area.

5 Conclusion and discussion

In this paper, a solution is proposed to address the challenges faced by traditional swimming coaches in real-time capture

TABLE 4 Ablation experiments on Swin-Transformer.

Method	Dataset															
	Swimming Technique Datasets				ImageNet Datasets				Kinetics Datasets				Sports-1M 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)
MCNN	328.21	304.96	249.11	303.41	290.37	250.07	331.06	282.68	341.81	257.36	318.86	284.97	296.08	208.04	249.22	282.81
MRNN	259.14	246.81	275.01	231.44	284.81	396.08	249.56	215.22	399.37	306.77	335.97	330.68	399.67	338.17	287.30	262.95
MGAN	269.87	338.84	376.72	389.20	229.19	360.81	279.19	305.23	255.55	257.07	341.24	351.83	379.39	391.93	335.80	370.03
Ours	107.88	121.12	158.02	213.89	137.65	168.31	222.03	154.32	207.48	103.57	108.22	157.11	151.60	201.48	211.11	104.16



and analysis of athlete movements. The RL-CWtrans Net is introduced as a robot vision-driven multimodal swimming training system aimed at providing precise and real-time guidance and feedback to swimmers. The RL-CWtrans Net combines reinforcement learning (RL) with the CWtrans Transformer module, enhancing its ability to understand the inherent complex spatial relationships in swimming motions. The system utilizes robot vision capabilities to track and assess key aspects of swimming techniques, monitoring posture, stroke mechanics, and body alignment, providing instant feedback to swimmers and coaches. Experimental validations of the RL-CWtrans Net demonstrate its effectiveness across multiple datasets, and this real-time guidance assists athletes in effectively improving their techniques, enhancing overall performance, and minimizing the learning curve. However, while our model performs well in specific dataset scenarios, further validation is crucial to verify its generalization capabilities on unknown or diverse datasets. Additionally, although it consumes fewer resources compared to existing models, optimization efforts are still necessary for extremely large-scale datasets or highly complex scenarios. Future research directions include exploring additional

algorithm optimizations to further compress the model and enhance the Swin-Transformer architecture for improved resource efficiency and speed. Furthermore, investigating applications of multitask learning can enhance the model's versatility and maintain its efficiency and accuracy across different datasets.

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

GW: 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 author declares that the research was conducted in the absence of any commercial or financial relationships

References

- Ahmadabadi, H., Manzari, O. N., and Ayatollahi, A. (2023). "Distilling knowledge from cnn-transformer models for enhanced human action recognition," in *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)* (IEEE), 180–184. doi: 10.1109/ICCKE60553.2023.10326272
- Ahmed, W., Naeem, U., Yousaf, M. H., and Velastin, S. A. (2022). "Lightweight cnn and gru network for real-time action recognition," in *2022 12th International Conference on Pattern Recognition Systems (ICPRS)* (IEEE), 1–7. doi: 10.1109/ICPRS54038.2022.9853854
- Ali, A., Pinyoanuntapong, E., Wang, P., and Dorodchi, M. (2023). Skeleton-based human action recognition via convolutional neural networks (CNN). *arXiv preprint arXiv:2301.13360*.
- Austin, M., Chase, A., Van Stratum, B., and Clark, J. E. (2022). Aquaclimber: a limbed swimming and climbing robot based on reduced order models. *Bioinspir. Biomimet.* 18:016004. doi: 10.1088/1748-3190/aca05c
- Banerjee, A., Singh, P. K., and Sarkar, R. (2020). Fuzzy integral-based cnn classifier fusion for 3d skeleton action recognition. *IEEE Trans. Circ. Syst. Video Technol.* 31, 2206–2216. doi: 10.1109/TCSVT.2020.3019293
- Biewener, A. A., Bompfrey, R. J., Daley, M. A., and Ijspeert, A. J. (2022). Stability and manoeuvrability in animal movement: lessons from biology, modelling and robotics. *Proc. R. Soc. B* 289:20212492. doi: 10.1098/rspb.2021.2492
- Brunner, G., Melnyk, D., Sigfússon, B., and Wattenhofer, R. (2019). "Swimming style recognition and lap counting using a smartwatch and deep learning," in *Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 23–31. doi: 10.1145/3341163.3347719
- Cabrera-Arellano, J. O., Hernandez-Hernandez, J. L., Marin-Hernandez, A., Cruz-Estrada, R., Hernandez-Mendez, S., and Guzman-Salda na, J. M. (2022). ipaddle: An intelligent swimming training system," in *2022 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)* (IEEE), 1–6. doi: 10.1109/ROPEC55836.2022.10018703
- Carreira, J., and Zisserman, A. (2017). "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308. doi: 10.1109/CVPR.2017.502
- Chen, W.-H., Cho, P.-C., Fan, P.-L., and Yang, Y.-W. (2011). "A framework for vision-based swimmer tracking," in *2011 International Conference on Uncertainty Reasoning and Knowledge Engineering* (IEEE), 44–47. doi: 10.1109/URKE.2011.6007835
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), 248–255. doi: 10.1109/CVPR.2009.5206848
- Dhiman, C., Vishwakarma, D. K., and Agarwal, P. (2021). Part-wise spatio-temporal attention driven cnn-based 3d human action recognition. *ACM Trans. Multim. Comput. Commun. Applic.* 17, 1–24. doi: 10.1145/3441628
- Dong, L., Li, Y., Zhou, X., Wen, Y., and Guan, K. (2020). Intelligent trainer for dynamical model-based deep reinforcement learning. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 2758–2771. doi: 10.1109/TNNLS.2020.3008249
- Geisen, M., and Klatt, S. (2022). Real-time feedback using extended reality: a current overview and further integration into sports. *Int. J. Sports Sci. Coach.* 17, 1178–1194. doi: 10.1177/17479541211051006
- Hamidi Rad, M., Gremeaux, V., Massé, F., Dadashi, F., and Aminian, K. (2022). Smartswim, a novel imu-based coaching assistance. *Sensors* 22:3356. doi: 10.3390/s22093356
- Jiang, X. (2009). "Feature extraction for image recognition and computer vision," in *2009 2nd IEEE International Conference on Computer Science and Information Technology* (IEEE), 1–15. doi: 10.1109/ICCSIT.2009.5235014
- Jin, X., Hou, J., Zhou, W., and Lee, S.-J. (2023). Recent advances in image fusion and quality improvement for cyber-physical systems. *Front. Neurobot.* 17:1201266. doi: 10.3389/978-2-8325-2459-6
- Kon, Y., Omae, Y., Sakai, K., Takahashi, H., Akiduki, T., Miyaji, C., et al. (2015). "Toward classification of swimming style by using underwater wireless accelerometer data," in *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, 85–88. doi: 10.1145/2800835.2800875
- Li, M., and Sun, Q. (2021). 3d skeletal human action recognition using a cnn fusion model. *Mathem. Problems Eng.* 2021:6650632. doi: 10.1155/2021/6650632
- Li, Y., Chen, L., He, R., Wang, Z., Wu, G., and Wang, L. (2021). "Multisports: a multi-person video dataset of spatio-temporally localized sports actions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13536–13545. doi: 10.1109/ICCV48922.2021.01328
- Lin, Z., Wang, C., Li, Z., Wang, Z., Liu, X., and Zhu, Y. (2024a). Neural radiance fields convert 2D to 3D texture. *Appl. Sci. Biotechnol. J. Adv. Res.* 3, 40–44.
- Lin, Z., Wang, Z., Zhu, Y., Li, Z., and Qin, H. (2024b). Text sentiment detection and classification based on integrated learning algorithm. *Appl. Sci. Biotechnol. J. Adv. Res.* 3, 27–33.
- Liu, C.-S. (1996). "A general framework of feature extraction: application to speaker recognition," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* (IEEE), 669–672.
- Liu, W., Bai, K., He, X., Song, S., Zheng, C., and Liu, X. (2022). "Fishgym: a high-performance physics-based simulation framework for underwater robot learning," in *2022 International Conference on Robotics and Automation (ICRA)* (IEEE), 6268–6275. doi: 10.1109/ICRA46639.2022.9812066
- Ma, J. (2021). Teaching mode of sports swimming video in colleges and universities based on artificial intelligence. *Mathem. Probl. Eng.* 2021:1–9. doi: 10.1155/2021/4076905
- Ming, Y., Xiong, L., Jia, X., Zheng, Q., and Zhou, J. (2023). "Fsconformer: a frequency-spatial-domain cnn-transformer two-stream network for compressed video action recognition," in *2023 IEEE Smart World Congress (SWC)* (IEEE), 838–843. doi: 10.1109/SWC57546.2023.10448953
- Modungwa, D., Mekuria, F., and Kekana, M. (2021). "Conceptual development of an autonomous underwater robot design for monitoring and harvesting invasive weeds," in *2021 IEEE AFRICON* (IEEE), 1–5. doi: 10.1109/AFRICON51333.2021.9570971
- Omae, Y., Kon, Y., Kobayashi, M., Sakai, K., Shionoya, A., Takahashi, H., et al. (2017). Swimming style classification based on ensemble learning and adaptive feature value by using inertial measurement unit. *J. Adv. Computat. Intell. Intell. Inform.* 21, 616–631. doi: 10.20965/jaciii.2017.p0616
- Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., et al. (2022). Multi-sensor information fusion based on machine learning for real applications in human activity recognition: state-of-the-art and research challenges. *Inf. Fusion* 80:241–265. doi: 10.1016/j.inffus.2021.11.006
- Santos, L., Geminiani, A., Schydlo, P., Olivieri, I., Santos-Victor, J., and Pedrocchi, A. (2021). Design of a robotic coach for motor, social and cognitive skills training toward applications with asd children. *IEEE Trans. Neural Syst. Rehabil. Eng.* 29:1223–1232. doi: 10.1109/TNSRE.2021.3091320
- Stolarz, M., Mitrevski, A., Wasil, M., and Plöger, P. G. (2024). Learning-based personalisation of robot behaviour for robot-assisted therapy. *Front. Robot. AI* 11:1352152. doi: 10.3389/fnbot.2024.1352152
- Style, S. (2014). Online monitoring of swimmer training using a 3D accelerometer.

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.

Victor, B., He, Z., Morgan, S., and Miniutti, D. (2017). "Continuous video to simple signals for swimming stroke detection with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 66–75. doi: 10.1109/CVPRW.2017.21

Wang, M., and Liang, Z. (2023). Cross-modal self-attention mechanism for controlling robot volleyball motion. *Front. Neurobot.* 17:1288463. doi: 10.3389/fnbot.2023.1288463

Wang, P., Cao, T., Li, X., Jin, X., Liu, P., Zhou, W., et al. (2023). Multi-focus image fusion based on gradient tensor hosvd. *J. Electron. Imaging* 32:023028. doi: 10.1117/1.JEI.32.2.023028

Wang, Z., Li, Z., Yu, X., Jia, Z., Xu, X., and Schuller, B. W. (2024). Cross-scene semantic segmentation for medical surgical instruments using structural similarity based partial activation networks. *IEEE Trans. Med. Robot. Bion.* 6, 399–409. doi: 10.1109/TMRB.2024.3359303

Wang, Z., Shi, X., Wang, J., Gao, F., Li, J., Guo, M., et al. (2019). "Swimming motion analysis and posture recognition based on wearable inertial sensors," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (IEEE), 3371–3376. doi: 10.1109/SMC.2019.8913847

Yi-Bo, L., Chang, Z., and Ning, W. (2010). "A survey on feature extraction of sar images," in *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)* (IEEE), V1–312. doi: 10.1109/ICCASM.2010.5619316