

OPEN ACCESS

EDITED BY
Long Jin,
Lanzhou University, China

REVIEWED BY
Zhongbo Sun,
Changchun University of Technology, China
Lin Chen,
Chinese Academy of Sciences (CAS), China
Mingchuan Zhou,
Zhejiang University, China

*CORRESPONDENCE
Zhanghe Liang
✉ liangzhanghe520@163.com

RECEIVED 04 September 2023
ACCEPTED 30 October 2023
PUBLISHED 10 November 2023

CITATION
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

COPYRIGHT
© 2023 Wang and Liang. 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.

Cross-modal self-attention mechanism for controlling robot volleyball motion

Meifang Wang¹ and Zhanghe Liang^{2*}

¹Sports Department, Anhui Agricultural University, Hefei, China, ²School of Sports Science, Hefei Normal University, Hefei, Anhui, China

Introduction: The emergence of cross-modal perception and deep learning technologies has had a profound impact on modern robotics. This study focuses on the application of these technologies in the field of robot control, specifically in the context of volleyball tasks. The primary objective is to achieve precise control of robots in volleyball tasks by effectively integrating information from different sensors using a cross-modal self-attention mechanism.

Methods: Our approach involves the utilization of a cross-modal self-attention mechanism to integrate information from various sensors, providing robots with a more comprehensive scene perception in volleyball scenarios. To enhance the diversity and practicality of robot training, we employ Generative Adversarial Networks (GANs) to synthesize realistic volleyball scenarios. Furthermore, we leverage transfer learning to incorporate knowledge from other sports datasets, enriching the process of skill acquisition for robots.

Results: To validate the feasibility of our approach, we conducted experiments where we simulated robot volleyball scenarios using multiple volleyball-related datasets. We measured various quantitative metrics, including accuracy, recall, precision, and F1 score. The experimental results indicate a significant enhancement in the performance of our approach in robot volleyball tasks.

Discussion: The outcomes of this study offer valuable insights into the application of multi-modal perception and deep learning in the field of sports robotics. By effectively integrating information from different sensors and incorporating synthetic data through GANs and transfer learning, our approach demonstrates improved robot performance in volleyball tasks. These findings not only advance the field of robotics but also open up new possibilities for human-robot collaboration in sports and athletic performance improvement. This research paves the way for further exploration of advanced technologies in sports robotics, benefiting both the scientific community and athletes seeking performance enhancement through robotic assistance.

KEYWORDS

multimodal perception, volleyball robot, spiking skill, cross-modal self-attention mechanism, adversarial network, transfer learning

1. Introduction

With the rapid advancement of technology, robotics is gradually permeating various fields, including sports. This study aims to enhance robotic skills in volleyball through deep learning and multimodal sensing technology, injecting innovation, and vitality into the realm of sports (Hong et al., 2021).

High-level sports demand athletes to possess outstanding perceptual, reaction speed, and motor control abilities. The development of modern technology has created opportunities for the application of robotics (Siedentop and Van der Mars, 2022). Robots can serve as ideal practice partners for athletes, enriching

the levels and enjoyment of competitions, and offering audiences novel viewing experiences (Siegel and Morris, 2020).

This research is focused on volleyball, a sport characterized by intense teamwork, demanding athletes to make precise decisions and immediate reactions in rapidly changing game scenarios (Oliveira et al., 2020; Weiss et al., 2021). Despite the increasing utilization of robots in sports, there is still room for improvement in robot spiking skills in volleyball. Therefore, this study focuses on improving the skill level of robots in volleyball matches by integrating multimodal perception and deep learning methods, aiming to enable their practical use in real competitions.

In recent years, there has been significant interest and research in the application of robotics technology in the field of sports (Thuruthel et al., 2019; Chen et al., 2020; Oliff et al., 2020). However, despite the extensive research in various sports disciplines, the exploration and study of robotics in volleyball have been relatively limited. Current research primarily focuses on aspects such as robot design, perception, and interaction (Ji et al., 2022; Hu et al., 2023). Nevertheless, there is still a need for further investigation into the application of multimodal perception and deep learning in this context (Olaniyan et al., 2022).

In recent years, there has been a growing interest and research focus on the application of robotics technology in the field of sports (Thuruthel et al., 2019; Chen et al., 2020; Oliff et al., 2020). However, despite extensive research across various disciplines in sports, exploration and research in volleyball robot technology have remained relatively limited. Current research primarily centers around aspects such as robot design, perception, and interaction (Ji et al., 2022; Hu et al., 2023). Jinho So and his colleagues (So et al., 2021) investigated the precise estimation of soft manipulator shape using stretchable shape sensors, while Li and Peng (2022) introduced a monocular visual-tactile sensor to enhance the robustness of robot manipulation. Nevertheless, there is still a need for further research on the application of multimodal sensing and deep learning in this domain (Olaniyan et al., 2022).

The contributions of this paper can be summarized in the following three aspects:

1. This study introduces a cross-modal self-attention mechanism designed to holistically address the amalgamation of diverse multimodal data collected by disparate sensors, including images and action sequences. Leveraging self-attention, we seamlessly integrate information from distinct modalities, enabling robots to comprehensively perceive cyclic motion scenarios. This innovative approach empowers robots to execute various operations with heightened accuracy in repetitive tasks, such as assessing ball velocity, trajectory, and opponent position in volleyball spiking, thus significantly elevating spiking proficiency.
2. The successful application of generative adversarial networks (GANs) to synthesize immersive cyclic motion scenarios is showcased. Through the generative and discriminative processes of GANs, we fabricate authentically textured virtual environments, imbuing robot skill training with heightened challenge and practicality. This augmentation not only fosters skill adaptability but also furnishes an expanded pool of training data, further propelling the prowess of robots.
3. The study maximizes the philosophy of transfer learning, funneling insights gleaned from alternate cyclic motion

datasets into the enhancement of robotic skills. This knowledge infusion expedites the robot's mastery of cyclic motion domains, facilitating swift adaptation to competitive settings and accelerated skill growth. This method not only introduces fresh paradigms for robot training but also widens the horizons of transfer learning's applicability in the realm of robotics.

The logical structure of this article is as follows. In Section 2, methods, the technical methods used in this study are introduced in detail, including cross-modal self-attention mechanism, adversarial network, and transfer learning. In Section 3, experiments, the experimental environment and data are described, and the evaluation indicators are introduced. At the same time, the experimental results were analyzed in detail, the performance of different methods and data sets were compared, and the effectiveness of the technical method was verified. In Section 4, discussion and conclusion, the research results are summarized, the significance and contribution of the research are evaluated, the limitations of the research are pointed out, and prospects for future work are proposed.

2. Methodology

In the method part of Chapter 3, we will introduce the overall algorithm flow of this research in detail, and show how to improve the spiking skills of volleyball robots through key technologies such as cross-modal self-attention mechanism, adversarial network, and transfer learning. This comprehensive algorithm process will provide the basis for subsequent experiments and comparative analysis, and also present the overall framework of this study for readers. The overall algorithm flow chart is shown in Figure 1.

2.1. Cross-modal self-attention mechanism

When dealing with multimodal data, attention mechanisms are powerful tools that allow the model to focus on the most relevant information from different modalities. We leverage a cross-modal self-attention mechanism to effectively integrate data from various sensors for enhancing the skills of our volleyball robot (Wang et al., 2021). Attention mechanisms are widely used in deep learning, enabling models to selectively attend to important parts of the data while disregarding irrelevant portions. There are two types of attention mechanisms: self-attention and cross-attention (Niu et al., 2021). Self-attention involves interactions and fusion of information within the same modality. For example, in a language model, self-attention allows each word to adjust its representation based on the context. Cross-attention involves interactions and fusion of information between different modalities. For instance, in visual question-answering tasks, cross-attention can establish correspondences between questions and images. The cross-modal self-attention mechanism is illustrated in Figure 2.

The key to the self-attention mechanism is to calculate the attention weight. One of the classic methods is to use Scaled Dot-Product Attention. Given a set of query vectors (Q), key vectors (K), and value vectors (V), it can compute attention weights

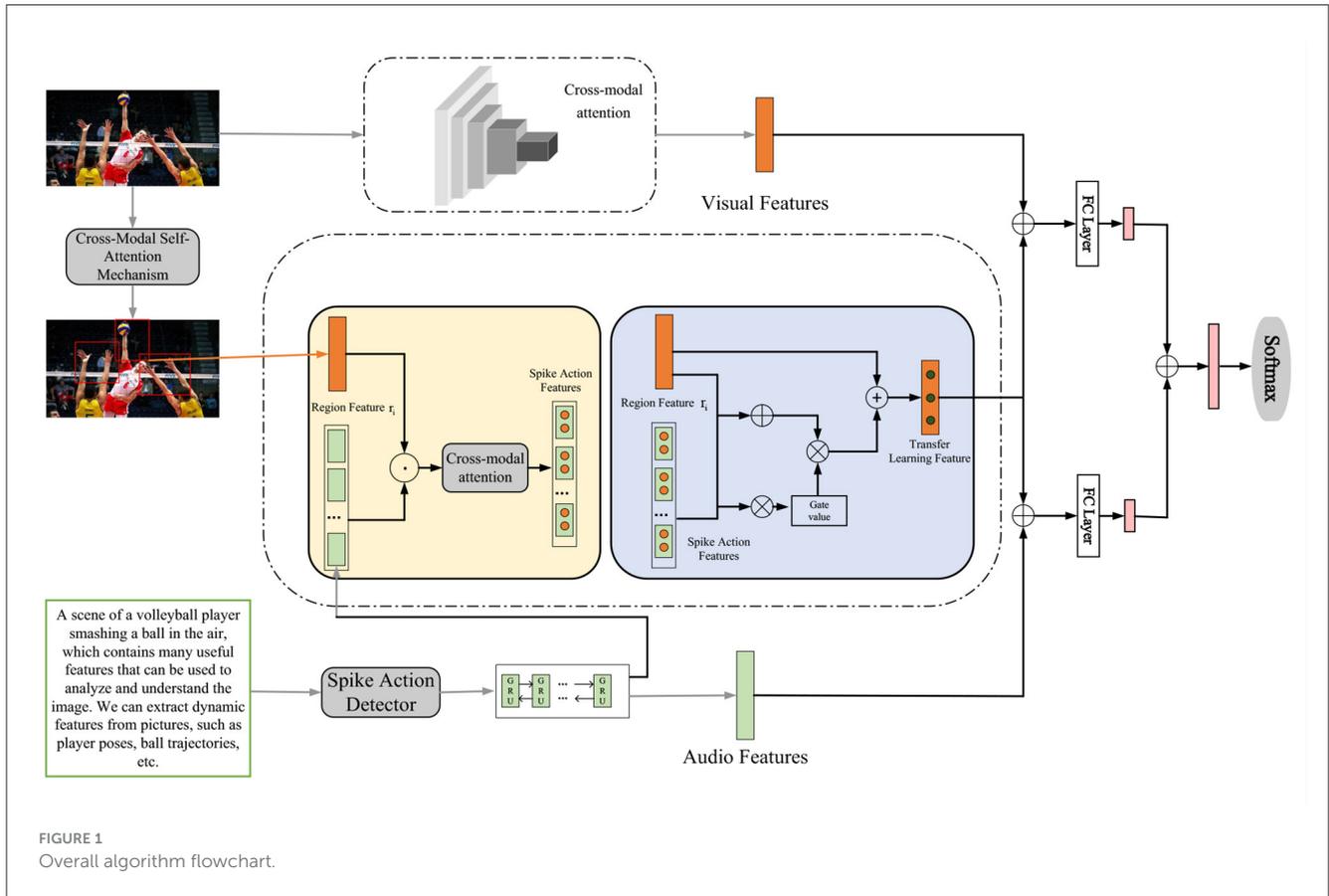


FIGURE 1 Overall algorithm flowchart.

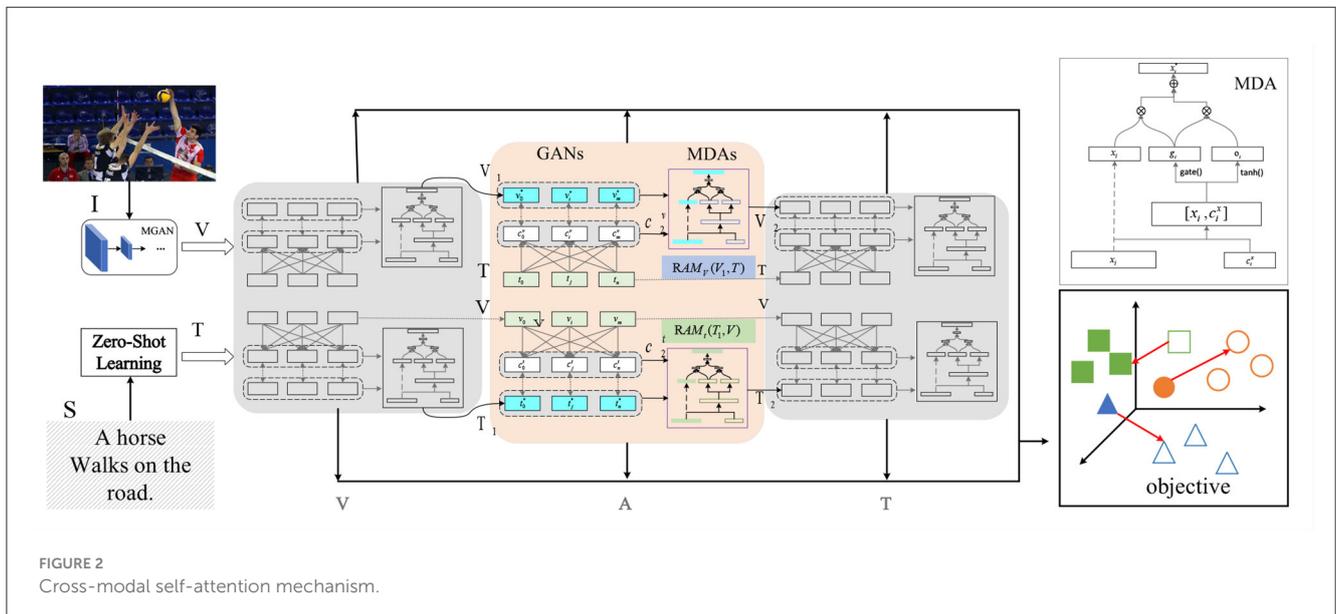


FIGURE 2 Cross-modal self-attention mechanism.

by the following formula:

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

where d_k is the dimension of the query and key vectors. The dot product operation in this formula expresses the similarity between the query and the key, and then normalizes using the *softmax* function to get the attention weights. Finally, the weighted

values are obtained by multiplying the attention weights with the value vector.

In our study, we further apply the attention mechanism to multimodal data. To synthesize information from different sensors, we introduce a cross-modal self-attention mechanism. In this approach, we take the feature representations of different modalities as queries, keys and values, so that the model can automatically learn the correlation between different modalities.

Formally, suppose we have two modalities M_1 and M_2 with corresponding queries, keys, and values Q_1, K_1, V_1 and Q_2, K_2, V_2 , respectively. We can compute the cross-modal self-attention weights as follows:

$$\text{Cross-Modal Attention}(Q_1, K_2, V_2) = \text{softmax}\left(\frac{Q_1 K_2^T}{\sqrt{d_k}}\right) V_2 \quad (2)$$

Similarly, we can calculate the attention weight of modality M_2 to modality M_1 .

In practical applications, we also need to consider optimization methods such as loss function and gradient descent to train our model. A commonly used optimization function is the cross-entropy loss function, which has good results in multi-classification tasks. For neural network training, we usually use the backpropagation algorithm to calculate gradients and perform parameter updates. Its formula is as follows:

$$\text{CrossEntropy}(p, q) = -\sum_i p_i \log(q_i) \quad (3)$$

where p is the actual probability distribution, q is the probability distribution predicted by the model, and i represents the index of the category. By minimizing the cross-entropy loss, the model can better fit the training data, thus improving the accuracy of predictions.

During the training process of the neural network, we use the backpropagation algorithm to calculate the gradient (Zhang, 2019), and use optimization methods such as gradient descent to update the model parameters. Backpropagation calculates the gradient of each parameter to the loss function through the chain rule, and then uses gradient descent to update the parameters to gradually optimize the model.

Through the cross-modal self-attention mechanism, we can extract key information from different sensor data, realizing the organic fusion and collaboration of multi-modal data. This provides a more solid foundation for our subsequent Generative Adversarial Network and transfer learning. Next, we will detail how to further improve the skills of volleyball robots with the help of Generative Adversarial Network.

2.2. Generative adversarial networks

Generative Adversarial Networks (GANs) are a deep learning framework that consists of two neural networks called a generator and a discriminator (Mi et al., 2020). The goal of a generator is to generate data, such as images, audio, etc., from a random noise vector that has a distribution similar to real data. The goal of the discriminator is to distinguish the data generated by the generator from the real data and give a probability value indicating its authenticity. There is an adversarial relationship between the generator and the discriminator, that is, the generator tries to deceive the discriminator, and the discriminator tries to see through the generator. By alternately training the two networks, the generator is finally able to generate high-quality data, while

the discriminator cannot distinguish between real and fake. The confrontation network is shown in Figure 3.

The basic objective function of GAN can be expressed as:

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (4)$$

In this study, we use this method to enhance the spiking skills of a volleyball robot. Our method can effectively utilize the idea of adversarial learning, enabling the generator to learn useful knowledge from other sports game data and transfer it to volleyball games. Our method consists of the following three steps:

1. Data preprocessing. We used a video feature extraction tool to extract the features of each frame in the volleyball video and save it as a feature vector. This tool can use a variety of pre-trained models (such as I3D, I3D-non-local, SlowFast, etc.) to extract powerful video features. We divide each video into segments and label each segment indicating whether the segment contains a smashing action. We regard the clips containing the smashing action as positive samples and the clips not containing the smashing action as negative samples.
2. GANs are trained. We used a Conditional Generative Adversarial Networks to train our model (Xu et al., 2019). Conditional Generative Adversarial Networks is a method of introducing additional information into GANs, such as category labels, text descriptions, etc. The objective function of Conditional Generative Adversarial Networks can be expressed as:

$$\text{minmax}_G V(D, G) = E_{x \sim p_{data}(x), y \sim p_{data}(y)} [\log D(x|y)] + E_{z \sim p_z(z), y \sim p_{data}(y)} [\log(1 - D(G(z|y)))] \quad (5)$$

Among them, $V(D, G)$ is the objective function of GANs, $D(x)$ is the probability that the discriminator gives the input x is real data, $G(z)$ is the data generated by the generator from the noise vector z , $p_{data}(x)$ is the real data distribution, and $p_z(z)$ is the noise vector distribution. y is extra information, such as category labels. In our method, we use a textual description as additional information, indicating the requirement of the smashing action, such as "the smashing angle is 45 degrees, and the force is 80%". The training process of GANs can be regarded as a zero-sum game, that is, the discriminator and the generator compete with each other so that the objective function reaches the Nash equilibrium, namely:

$$G^* = \underset{G}{\text{argminmax}} V(D, G) \quad (6)$$

3. GANs application. We use a decoder to restore the sequence of feature vectors of the video clips produced by the generator to a sequence of images, which are stitched into a single video. We compare the generated videos with real volleyball match videos to evaluate their quality and authenticity. We also use the generated videos as training data for the volleyball robot to improve its spiking skills. The output of the decoder can be expressed as:

$$\hat{x}_t = f_{dec}(h_t) \quad (7)$$

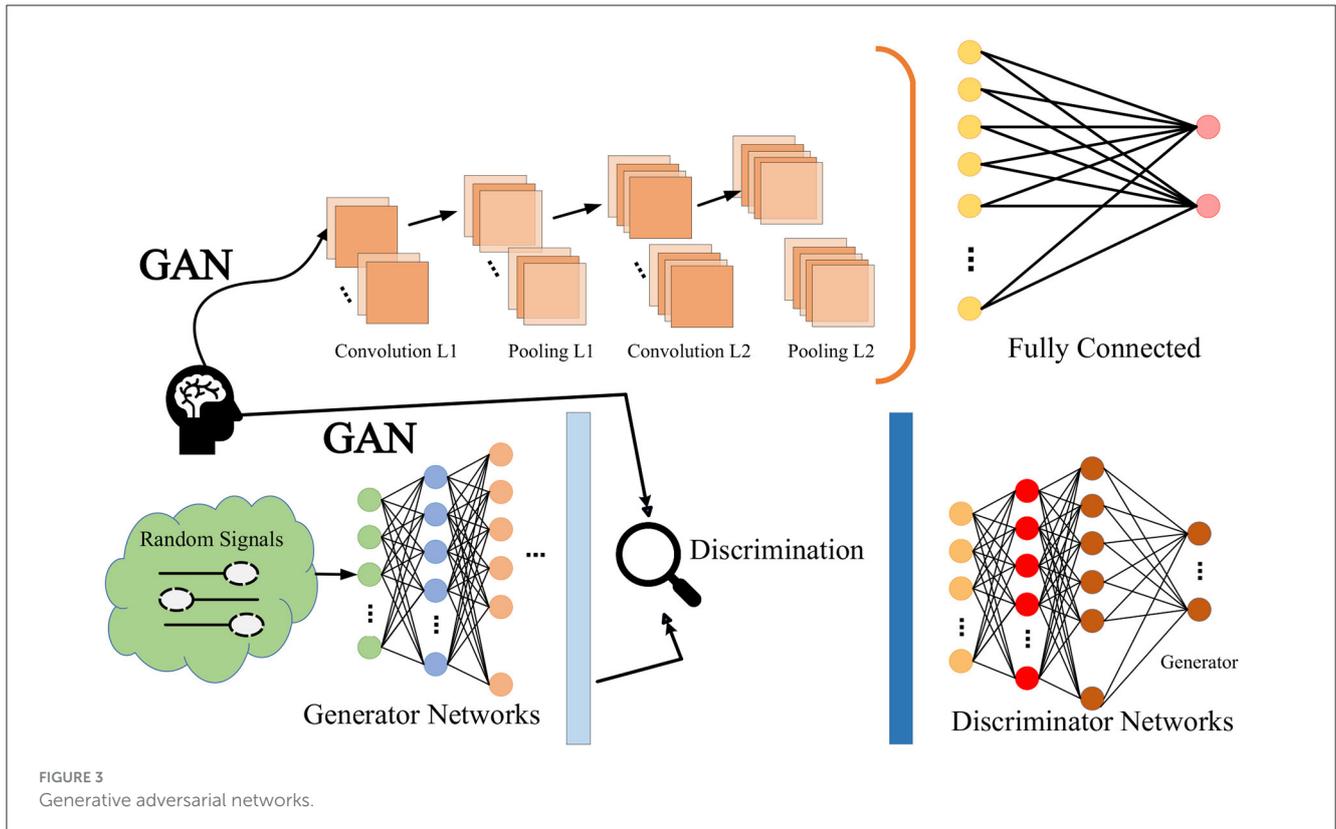


FIGURE 3
Generative adversarial networks.

Among them, \hat{x}_t is the image generated at the t-th moment, f_{dec} is the decoder function, and h_t is the feature vector output by the generator at the t-th moment.

During training, the generator and discriminator are optimized through adversarial learning, specifically, the generator tries to minimize $V(D, G)$, while the discriminator tries to maximize $V(D, G)$. This leads to a dynamic balancing process at which the samples generated by the generator are realistic enough that the discriminator cannot effectively distinguish real samples from generated samples. In terms of optimization functions, for the generator G, we can use the following optimization functions to update the parameters of the generator:

$$\min_G V(D, G) = \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (8)$$

In practical applications, through methods such as backpropagation and gradient descent, the parameters of the generator and discriminator can be gradually optimized to achieve the goal of training GANs.

By introducing GANs, we can further improve the skills of volleyball robots and generate more realistic and diverse game scenes, thus laying a more solid foundation for the improvement of robot skills. Next, we explore how transfer learning can be applied to skill improvement for volleyball robots.

2.3. Transfer learning

We use a transfer learning method called domain adaptation (Zhuang et al., 2020). In this approach, we improve the

generalization of the model on the target domain by minimizing the domain difference between the source and target domains. Transfer learning is shown in the Figure 4.

Assuming we have source domain data D_{source} and target domain data D_{target} , our goal is to transfer the knowledge on the source domain to the target domain. We can achieve this by minimizing the distribution difference between the source and target domains. A common method is Maximum Mean Difference (MMD):

$$MMD(D_{source}, D_{target}) = \left\| \frac{1}{n_{source}} \sum_{i=1}^{n_{source}} \phi(x_{source}^i) - \frac{1}{n_{target}} \sum_{j=1}^{n_{target}} \phi(x_{target}^j) \right\|^2 \quad (9)$$

Among them, x_{source}^i and x_{target}^j denote samples in the source domain and target domain, respectively, and $\phi(\cdot)$ is a mapping function that maps samples into a latent space. By minimizing MMD, we can reduce the distribution difference between source and target domains, thus enabling transfer learning.

Another common approach is Domain Adversarial Neural Network (DANN) (Ajakan et al., 2014). In DANN, we introduce a domain classifier whose goal is to distinguish samples in the source and target domains. At the same time, we train a feature extractor to generate features that are indistinguishable to domain classifiers. This can be achieved by minimizing the loss function of the domain classifier:

$$\mathcal{L}_{domain} = -\frac{1}{n} \sum_{i=1}^n \log D(f(x_i)) \quad (10)$$

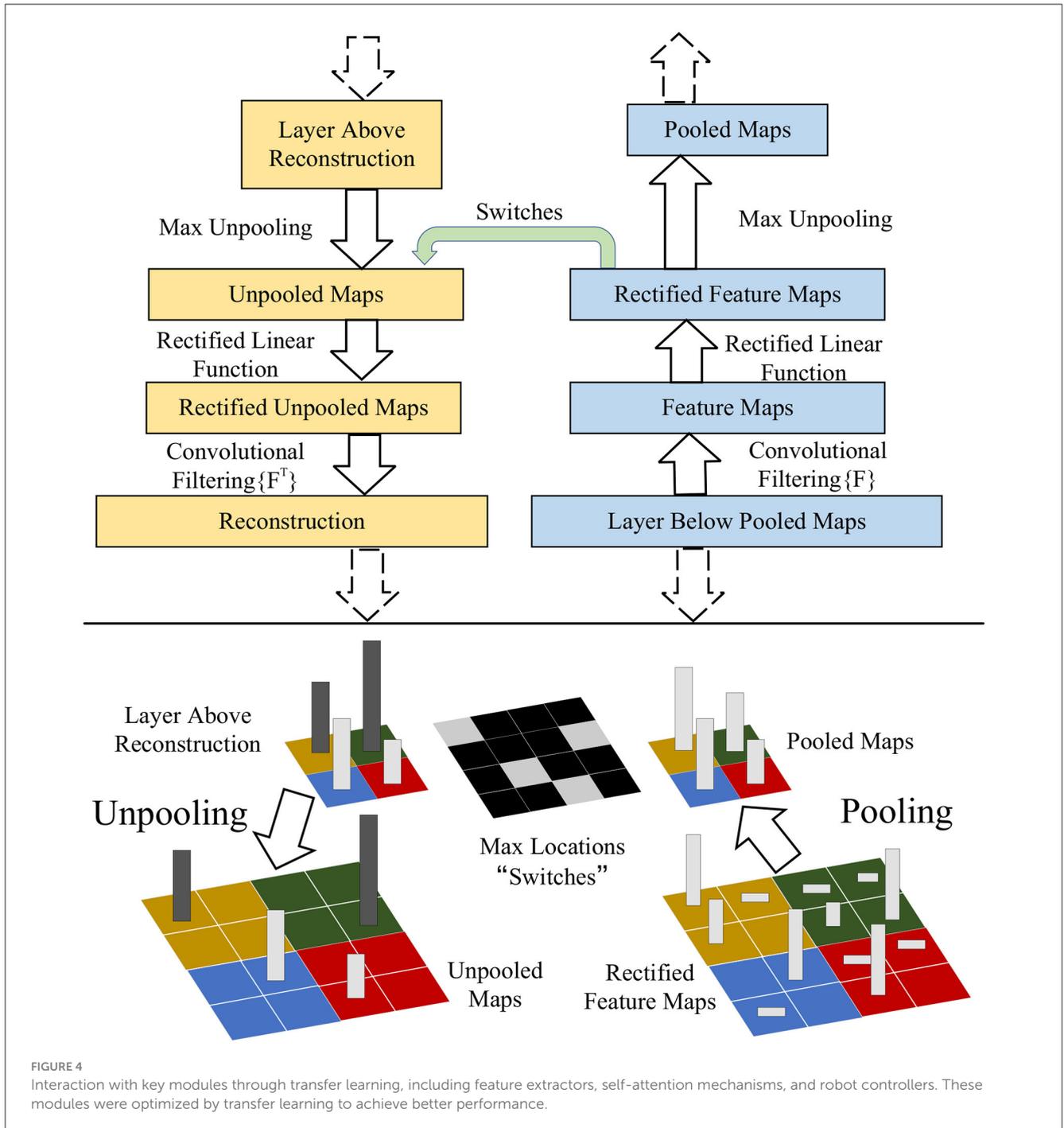


FIGURE 4 Interaction with key modules through transfer learning, including feature extractors, self-attention mechanisms, and robot controllers. These modules were optimized by transfer learning to achieve better performance.

Among them, $D(\cdot)$ is the domain classifier, $f(\cdot)$ is the feature extractor, and E is the sample. By minimizing L_{domain} , we can make features more consistent across domains, enabling transfer learning.

In addition, there is a common method of transfer learning by training an initial model on the source domain, then using the parameters of this model as the initial parameters of the target domain model, and then fine-tuning the model parameters on the target domain. This can be achieved by minimizing a loss function over the target domain:

$$\mathcal{L}_{target}(f_{target}, \mathcal{D}_{target}) = \mathbb{E}_{(x,y) \sim \mathcal{D}_{target}} [\ell(f_{target}(x), y)] \quad (11)$$

Among them, f_{target} is the model on the target domain, \mathcal{D}_{target} is the data distribution of the target domain, (x, y) is the sample of the target domain, and ℓ represents the loss function.

Optimization methods for transfer learning usually consist of two steps: feature extraction and fine-tuning. In the feature extraction stage, we can extract general feature representations from the source domain through pre-trained models. Then, in the

fine-tuning stage, we train the extracted feature representations together with data from the target domain to further adapt to the target domain. Specifically, the fine-tuning optimization function can be expressed as:

$$\mathcal{L}_{\text{target}}(f_{\text{target}}, D_{\text{target}}) + \lambda \cdot \mathcal{L}_{\text{source}}(f_{\text{target}}, D_{\text{source}}) \quad (12)$$

Among them, L_{source} represents the loss function on the source domain, and λ is a hyperparameter that weighs the two losses.

Through transfer learning, we can make full use of the knowledge of the existing modality in the task of volleyball robot and improve the performance of the model in the new modality.

In the Section 2 of this chapter, we propose a method that comprehensively applies attention mechanisms, GANs, and transfer learning to improve the skills of volleyball robots. First, we introduce a cross-modal self-attention mechanism, which effectively integrates multi-modal sensor data, enabling the model to automatically learn the correlation between different modalities. By calculating attention weights, we are able to extract key information from different sensor data, laying a solid foundation for the subsequent steps. Then, we introduced the application of GAN. Through domain adaptation and domain confrontation neural network, the knowledge transfer between the source domain and the target domain is realized, thereby improving the generalization ability of the model in the target domain. Finally, we explore how to train the initial model on the source domain and fine-tune the parameters on the target domain to fit the data distribution of the target domain through transfer learning. The comprehensive application of these methods provides strong support for our experimental part. In the next chapter, we will introduce the experimental design and result analysis in detail to verify the effectiveness and performance improvement of our proposed method in improving the skills of volleyball robots.

3. Experiment

The experimental process of this paper is shown in [Figure 5](#).

3.1. Experimental environment

- **Hardware Environment**

In this research, we rely on an advanced computing platform as the hardware environment, which is equipped with a high-performance AMD Ryzen 7 5800X processor, equipped with 64GB ultra-high-speed DDR4 memory, and configured with 2 NVIDIA GeForce RTX 3080 10GB graphics card. This excellent hardware configuration endows us with powerful computing and storage capabilities, especially suitable for training and inference of deep learning tasks. In addition, we also use multi-channel SSD hard disk to ensure the high efficiency of data reading and storage. Such a hardware environment provides strong support for the

smooth progress of the experiment, making the training process of the model more efficient, stable, and reliable.

- **Software Environment**

In this study, we used Python and PyTorch to implement a method for improving the spiking skills of volleyball robots based on deep learning. As the main deep learning framework, PyTorch provides us with powerful model building and training tools, allowing us to flexibly design and optimize our spiking skill model. In the experiment, we made full use of PyTorch's efficient computing power and automatic differentiation function to speed up the model training process, so that our model can converge faster and achieve better results.

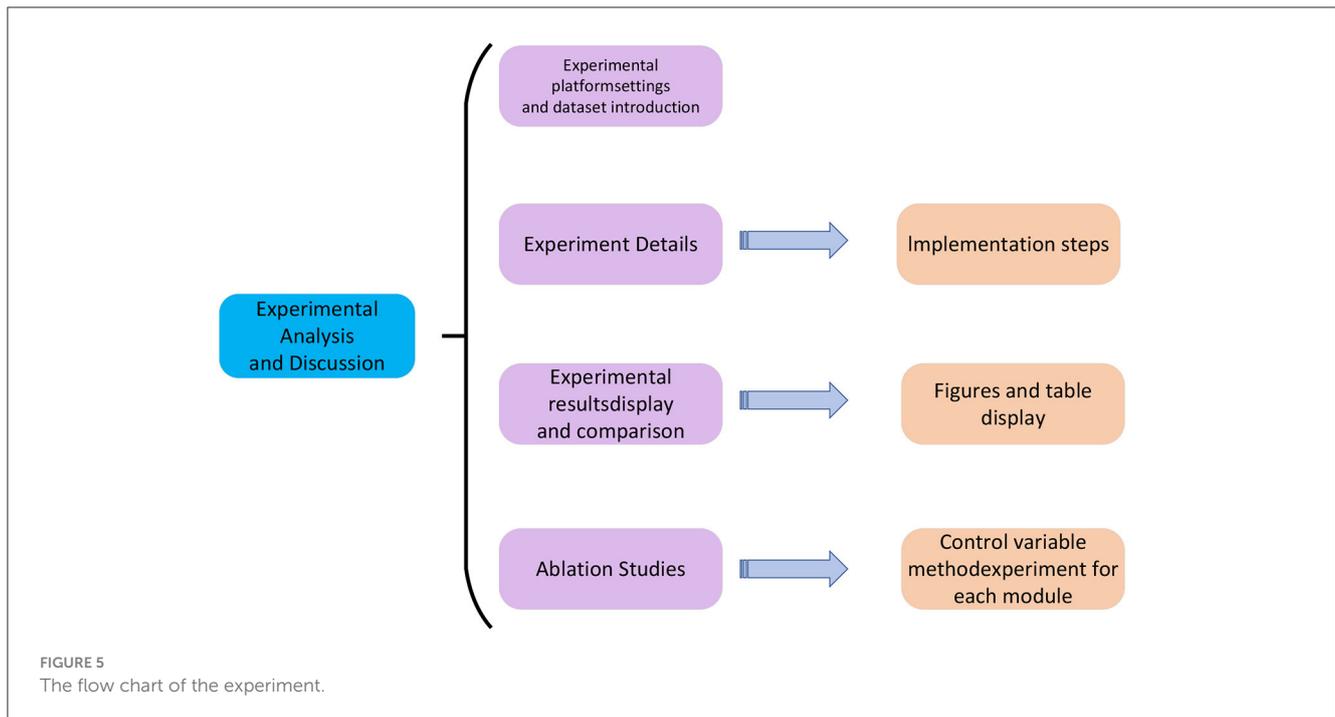
3.2. Experimental data

- **Volleyball Dataset**

Volleyball Dataset is a video action recognition dataset proposed by Ibrahim et al. of Simon Fraser University in Canada in 2016. The data set consists of 55 volleyball game videos, in which 4830 key frames mark the player's position, individual action and group behavior. Single action includes 9 categories, such as smash, block, pass, etc. Group behavior includes 8 categories, such as passing the ball to the left, scoring from the right, and both sides scrimmage. This dataset aims to provide a challenging scenario for studying the recognition and understanding of human actions and group activities in videos. It can be used for a variety of video analysis tasks, such as action recognition, group activity recognition, person tracking, etc. This dataset has been used and cited by several research papers, demonstrating its value and influence in the field of video analysis.

- **VREN: Volleyball Rally Dataset with Expression Notation Language**

VREN is a video volleyball game dataset proposed by Xia et al. at the University of California, Santa Barbara in 2022. This dataset contains video clips from professional and NCAA Div-I indoor volleyball matches, where each round is annotated with a volleyball description language. This language can completely describe the player's action, position, and volleyball trajectory in the volleyball game. This dataset aims to provide a rich and high-level benchmark for studying the skills of robots in volleyball games. Based on the language, this dataset proposes three tasks for automated volleyball action and tactical analysis: (1) volleyball round prediction, which aims to predict the outcome of rounds and help players and coaches improve decision-making in practice; (2) setter type and Smash type prediction, helping athletes, and coaches to prepare for the game more effectively; (3) Volleyball tactics and offensive zone statistics, providing advanced volleyball statistics to help coaches better understand the game and opponent's tactics. The authors conduct a case study showing how experimental results can provide insight to the volleyball analysis community. Furthermore, experimental evaluations on real data establish a baseline for future research and applications. The research bridges



the gap between the field of indoor volleyball and computer science.

- UCF101

The UCF101 dataset is a video action recognition dataset proposed by Soomro et al. at the University of Central Florida in 2012. The dataset consists of 13,320 real action videos from YouTube, covering 101 action categories. These action categories can be divided into five types: human-object interaction, body movement, human-human interaction, playing musical instruments, and sports, some of which are related to volleyball, such as smashing, blocking, passing, etc. This dataset is an extension of the UCF50 dataset, which has only 50 action categories. The UCF101 dataset is highly diverse and challenging because there are a large number of changes in camera motion, object appearance and pose, object scale, viewing angle, background clutter, and lighting conditions in the video. This dataset aims to facilitate further research in the field of action recognition by learning and exploring new categories of real actions.

- MultiSports dataset

The MultiSports dataset is a video multiplayer sports action detection dataset, which was proposed by Li et al. of Nanjing University in 2021. The dataset consists of 3200 video clips of sports games from YouTube, covering 4 sports categories: aerobics, basketball, football, and volleyball. The dataset annotates 37,701 action instances and 902k bounding boxes, and each action instance has a fine-grained action category label, such as smashing, blocking, passing, etc. This dataset aims to provide a rich and challenging benchmark for studying multi-person video action detection. The dataset has the characteristics of high diversity, high density, and high quality, and can reflect real sports competition scenes.

3.3. Evaluation index

In the assessment process of this research, in order to comprehensively and objectively measure the effectiveness and performance of the proposed sports teaching method, a series of key evaluation metrics were employed. These metrics not only facilitate a quantitative evaluation of the model's performance across various tasks but also provide us with in-depth insights to better comprehend the strengths and limitations of the method. In the following section, we will provide a detailed introduction and analysis of the following key metrics: accuracy, recall, precision, and F1 score. These metrics will assist us in objectively evaluating the efficacy of the proposed method in the context of sports teaching, thereby providing robust support for the reliability of the research and the feasibility of its practical application.

- Hit rate

In the skill improvement task of the volleyball robot, the hitting rate is a critical evaluation metric used to measure the performance of the proposed method. The hitting rate is defined as the ratio between the number of events correctly predicted by the model and the total number of samples. It provides an intuitive reflection of the model's prediction accuracy, aiding in the assessment of its performance. The hitting rate can be calculated using the following formula:

$$\text{Hit Rate} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (13)$$

In the context of skill enhancement tasks for the volleyball robot, the hitting rate is a pivotal evaluation metric used to gauge the performance of the proposed method. The hitting rate is defined as the ratio between the number of correctly predicted positive samples (True Positives, TP), indicating the

number of instances where skill improvement was accurately identified, and the total number of samples. Similarly, the correctly predicted negative samples (True Negatives, TN) represent instances where the absence of skill improvement was accurately recognized. Conversely, the false positives (FP) correspond to instances where the model erroneously predicted positive samples when they were, in fact, negative. The false negatives (FN) denote cases where the model inaccurately predicted negative samples as positive.

By calculating the hitting rate, we gain insights into the model's accuracy in predicting skill levels, thereby evaluating the practicality and effectiveness of this approach in real-world sports teaching scenarios. In our research, the hitting rate will serve as a critical evaluation metric, assisting us in conducting a thorough analysis of model performance and providing robust support for subsequent experimental findings.

- Recall

In the skill enhancement task of the volleyball robot, recall is a critical evaluation metric used to assess the effectiveness of the proposed attention-based mechanism in capturing the skill level of volleyball players. Recall measures the model's ability to correctly identify actual positive samples, i.e., the proportion of samples that the model correctly predicts out of all actual positive samples. This is of significant importance for evaluating the model's overall performance in sports education. Recall can be calculated using the following formula:

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (14)$$

In the context of skill enhancement tasks for the volleyball robot, recall is a crucial evaluation metric used to assess the model's ability to correctly identify positive samples. Specifically, it measures the proportion of samples that the model accurately predicts as skill level improvements out of all actual positive samples in the volleyball skill enhancement task. Conversely, false negatives (FN) represent the positive samples that the model fails to predict accurately, indicating instances where skill level improvements were missed.

By computing the recall rate, we gain insights into the model's capacity to recognize positive samples, thus evaluating the effectiveness of the attention-based mechanism in enhancing the volleyball robot's skills. In our research, recall will serve as a vital evaluation metric, enabling us to conduct an in-depth analysis of model performance, providing a comprehensive assessment, and supporting subsequent experimental results.

- Precision

In the context of skill enhancement tasks for the volleyball robot, precision is a critical evaluation metric used to measure the accuracy of the attention-based method in predicting the skill level of volleyball players. Precision assesses the proportion of samples that the model predicts as positive samples, which are indeed positive samples in reality. This is of paramount importance for evaluating the reliability and accuracy of the model in sports education. Precision can be

calculated using the following formula:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \quad (15)$$

TP (True Positives): The number of positive samples correctly predicted by the model, indicating the instances where skill level improvement was accurately identified. FP (False Positives): The number of positive samples incorrectly predicted by the model, signifying the instances where the model erroneously predicted negative samples as positive.

By calculating precision, we gain insights into the model's accuracy when predicting positive samples, thereby evaluating the effectiveness of the attention-based method in the volleyball robot's skill enhancement task. In our research, precision will serve as a crucial evaluation metric, aiding us in analyzing model performance, providing a dependable assessment, and supporting our experimental results.

- F1 Score

In our study of skill enhancement in volleyball robots, the F1 score serves as a critical evaluation metric employed for the comprehensive assessment of the method's performance in skill improvement. This score takes into account both precision and recall, thus facilitating the equilibrium between the model's accuracy and comprehensiveness in identifying skill improvement instances. Consequently, it provides a more comprehensive performance measurement metric. The formula for calculating the F1 score is as follows:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \quad (16)$$

In this formula, we introduce previously discussed precision and recall as parameters. Precision measures the accuracy of the model in identifying positive samples as positive samples, while recall gauges the model's comprehensive recognition capability of positive samples.

The F1 score combines the accuracy and comprehensiveness of the model in skill improvement cases, making it a crucial evaluation metric in the volleyball skill enhancement study. By calculating the F1 score, we can gain a more comprehensive understanding of the method's performance, ensuring that the model achieves accurate and comprehensive results in skill improvement.

Algorithm 1 represents the algorithm flow of the training in this paper.

3.4. Experimental comparison and analysis

In the preceding sections, we provided a comprehensive introduction to the design and implementation of the multimodal perception-based deep learning approach for enhancing volleyball robot spiking skills. In this chapter, our focus shifts toward a comparative analysis of experimental results, aiming to comprehensively evaluate the effectiveness and superiority of the proposed methods. By conducting experiments on multiple datasets, our goal is to delve into the contributions of different

```

1: Input: Volleyball Dataset, VREN Dataset, UCF101
   Dataset, MultiSports Dataset
2: Initialize Trans-GAN Net parameters:  $\theta$ 
3: Initialize discriminator parameters:  $D$ 
4: Define cross-modal self-attention mechanism:
    $Attention(X)$ 
5: for each epoch do
6:   for each dataset in [Volleyball, VREN, UCF101,
   MultiSports] do
7:     Load batch of data:  $X$ 
8:     Compute attention weights:  $W = Attention(X)$ 
9:     Compute transformed features:  $X' = W \times X$ 
10:    Generate fake data:  $X_{fake} = TransGAN(X')$ 
11:    Compute discriminator loss:  $L_D = -\log(D(X_{fake})) - \log(1 - D(X))$ 
12:    Compute generator loss:  $L_G = -\log(D(X_{fake}))$ 
13:    Backpropagate and update  $\theta$  and  $D$  using  $L_D$ 
   and  $L_G$ 
14:   end for
15:   if epoch % transfer interval == 0 then
16:     Perform transfer learning by copying
     features to next layers
17:   end if
18:   Compute evaluation metrics on validation set:
19:   Hit Rate:  $\frac{\sum_{i=1}^N 1(y_i = \hat{y}_i)}{N}$ 
20:   Recall:  $\frac{\sum_{i=1}^N 1(y_i = \hat{y}_i \text{ and } y_i = 1)}{\sum_{i=1}^N 1(y_i = 1)}$ 
21:   Precision:  $\frac{\sum_{i=1}^N 1(y_i = \hat{y}_i \text{ and } y_i = 1)}{\sum_{i=1}^N 1(\hat{y}_i = 1)}$ 
22:   F1 Score:  $2 \times \frac{Precision \times Recall}{Precision + Recall}$ 
23:   if F1 Score > best score then
24:     Save best model
25:   end if
26: end for

```

Algorithm 1. Training of Trans-GAN Net.

models and their combinations in enhancing volleyball robot skills, as well as to validate the applicability of our approach across various scenarios. This process of experimental comparison and analysis not only directly showcases the practical effectiveness of our approach but also provides deeper insights, guiding us toward optimizing and advancing the technological trajectory of sports robots.

Through comparing experimental results across different datasets, we will uncover the performance of the multimodal perception-based deep learning approach in varying contexts. Simultaneously, we will integrate the evaluation metrics introduced earlier, such as hit rate, recall rate, precision, and F1 score, to conduct a comprehensive assessment of the overall model performance. We will also analyze the introduction of different modules, exploring the specific roles of cross-modal self-attention mechanisms, GANs, transfer learning, and other methods in enhancing volleyball robot skills. In-depth analysis of the experimental results will allow us to understand the interplay between different modules and their impact on enhancing robot skills.

Furthermore, we will compare the experimental results with those of the baseline models to quantify and illustrate the

superiority of our approach. Through comparative analysis, we can accurately evaluate the performance improvement brought about by the multimodal perception-based deep learning approach in enhancing volleyball robot skills. These comparative and analytical results will further validate the feasibility and practicality of our approach, providing robust support and references for research and applications in the field of sports robotics.

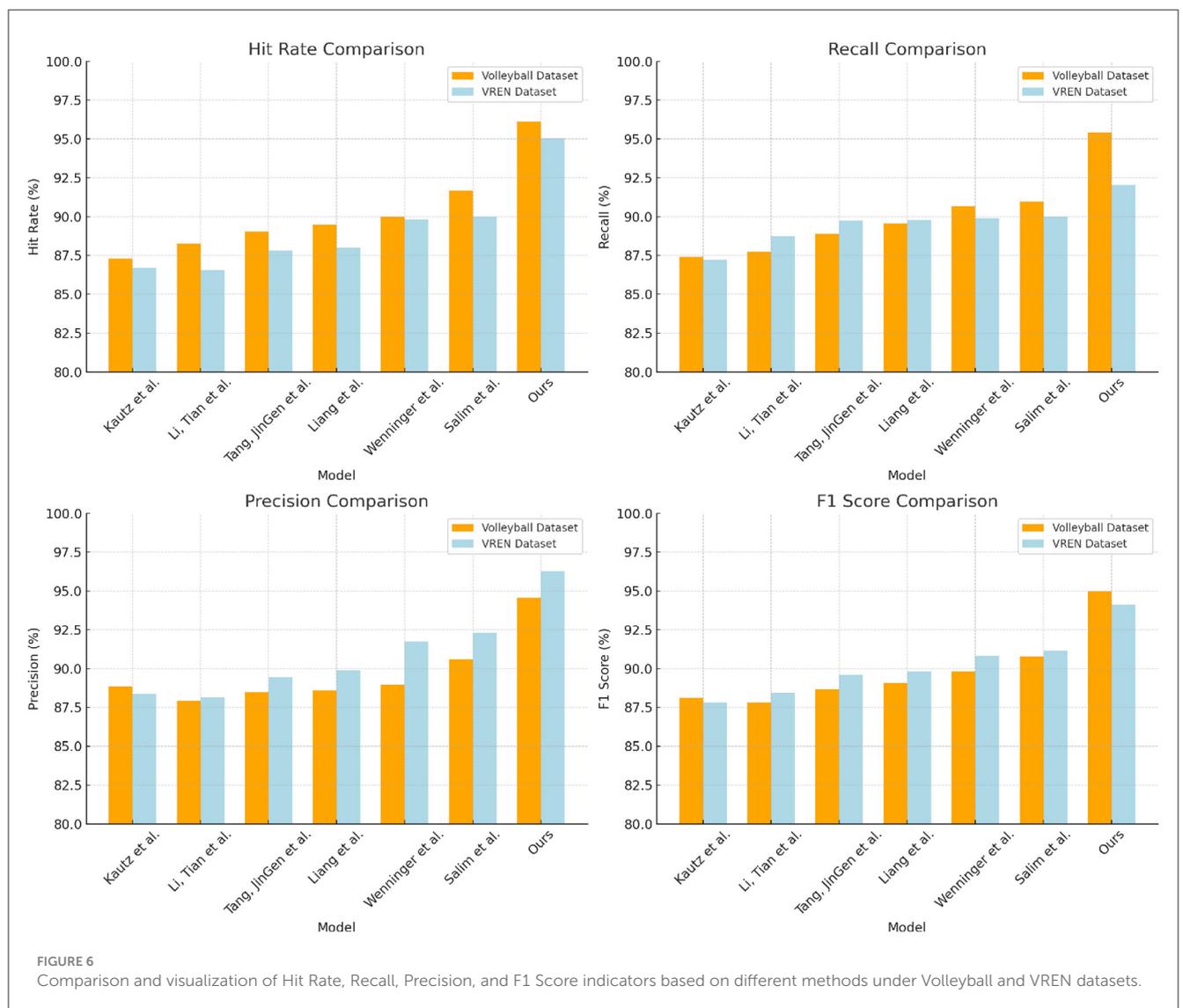
Next, we will meticulously dissect the experimental results, comprehensively showcasing the performance of our model across different datasets and metrics, providing readers with a comprehensive understanding of the model's capabilities and its potential value in real-world applications.

From the data in Table 1 above, it can be seen that our method outperforms other research works on both the Volleyball dataset and the VREN dataset. Specifically, on the Volleyball data set, after removing our method, compared with the research method of Salim et al., who achieved the highest hit rate of 91.66% and the F1 score of 90.77%, our hit rate It has increased by 4.45%, and the F1 score has also increased by 3.98%. At the same time, our precision and recall rate have also reached the optimal value of all methods, reaching 95.41 and 94.57%, respectively, and the performance on the VREN data set is also better than other methods, our hit rate and The F1 score is 8.33 and 6.31% higher than the research method of Kautz et al., and 7.03 and 4.28% higher than the method of Liang et al. In general, from the evaluation results on these two classic volleyball datasets, it can be seen that our new deep learning method with multi-modal information learning and deep generative network as the backbone is effective in identifying and predicting volleyball. There is a significant advantage in action. It can better learn and mine the visual and motion features in volleyball, so it has higher precision and recall. This shows that the method has great potential in improving the motion control skills of volleyball robots. Finally, we compared and visualized the results in Table 1, as shown in Figure 6.

According to the comparative data of Hit Rate, Recall, Precision, and F1 Score of different methods on the two datasets in Table 2 above, it can be seen that our method has significant advantages over other methods. On UCF101 dataset and MultiSports dataset, compared with the work of Kautz et al. using the same dataset, our proposed method achieves 9.57% higher hit rate and 7.79% higher recall rate on UCF101 dataset, F1 The score is 8.48% higher; the hit rate is 7.85% higher in the MultiSports dataset, the recall rate is 6.78% higher, and the F1 score is 7.46% higher. At the same time, excluding our method, compared with Salim et al.'s study on UCF101 which obtained the highest recall rate of 90.81%, our recall rate improved by 3.86%. Compared with Tang et al. who obtained F1 score of 88.20% in the MultiSports dataset, our F1 score increased by 6.88%. Furthermore, we exceed the main evaluation metrics of other methods such as Liang et al. and Wenninger et al. on these two action datasets. This shows that the method shows stronger generalization ability in learning joint motion and action features, and can better identify and classify different types of sports actions. Overall, its excellent performance on two large-scale general-purpose motion datasets once again confirms the advantages of this method in the field of action recognition. We compared and visualized the results in Table 2, as shown in Figure 7.

TABLE 1 Comparison of Hit Rate, Recall, Precision, and F1 Score indicators based on different methods under Volleyball and VREN datasets.

Model	Datasets							
	Volleyball dataset (Ibrahim et al., 2016)				VREN dataset (Ma et al., 2022)			
	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)
Kautz et al. (2017)	87.28	87.41	88.87	88.13	86.69	87.23	88.37	87.8
Li and Tian (2023)	88.24	87.75	87.93	87.84	86.54	88.73	88.15	88.44
Tang (2021)	89.02	88.88	88.47	88.67	87.82	89.74	89.46	89.6
Liang and Liang (2022)	89.47	89.57	88.59	89.08	87.99	89.79	89.88	89.83
Wenninger et al. (2020)	89.98	90.68	88.96	89.81	89.81	89.89	91.76	90.82
Salim et al. (2019)	91.66	90.95	90.59	90.77	90.02	89.99	92.31	91.14
Ours	96.11	95.41	94.57	94.99	95.02	92.02	96.29	94.11

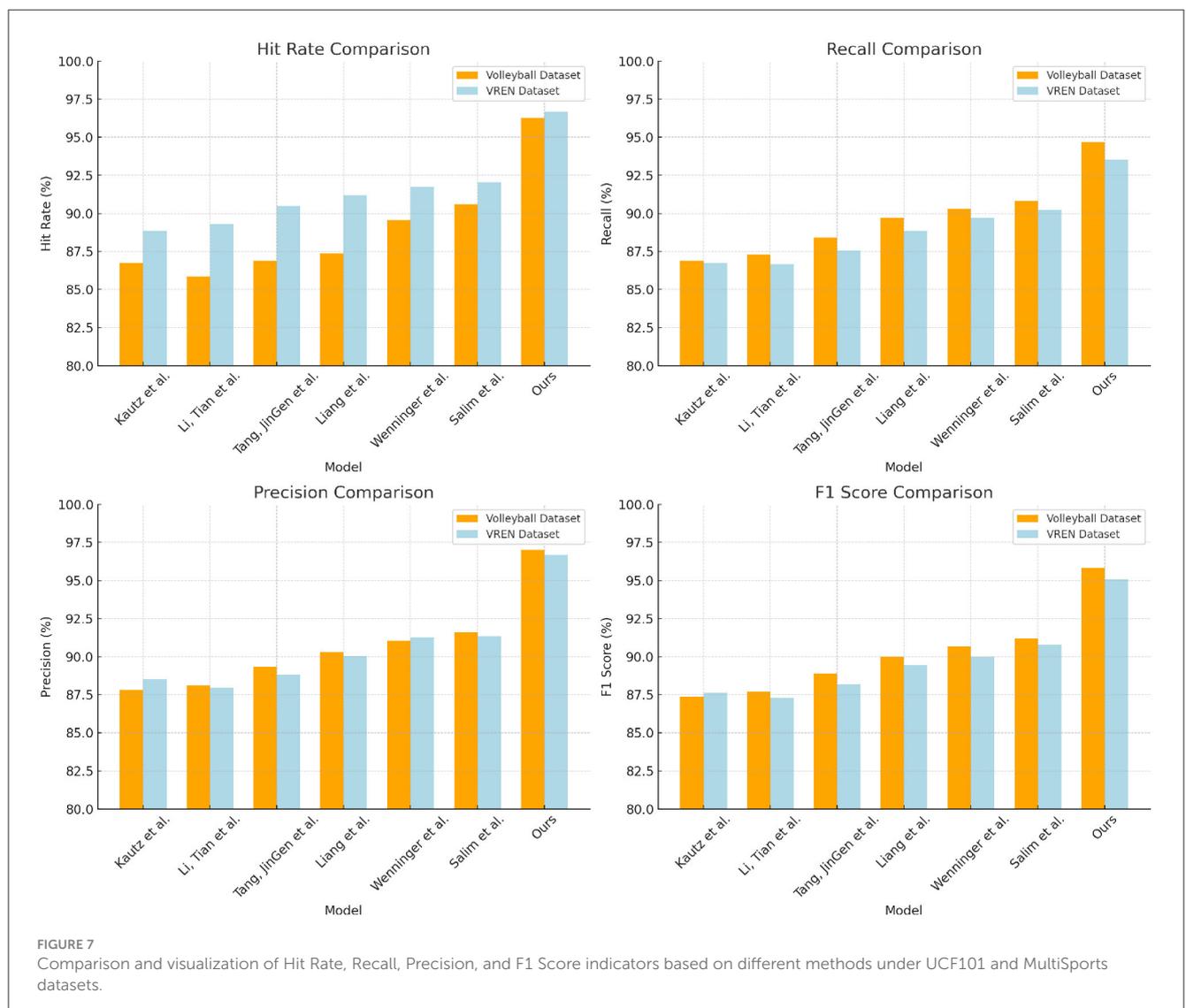


According to the data in Table 3 above, with the improvement of the model structure, the performance of our proposed method on the two classic volleyball data sets has been significantly

improved. Specifically, compared with the baseline model, after adding the self-attention mechanism, the hit rate on the Volleyball dataset increased by 6.78%, the recall rate increased by 11.76%,

TABLE 2 Comparison of Hit Rate, Recall, Precision, and F1 Score indicators based on different methods under UCF101 and MultiSports datasets.

Model	Datasets							
	UCF101 dataset (Soomro et al., 2012)				MultiSports dataset (Li et al., 2021)			
	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)
Kautz et al. (2017)	86.71	86.88	87.84	87.36	88.83	86.75	88.51	87.62
Li and Tian (2023)	85.87	87.32	88.11	87.71	89.3	86.67	87.98	87.32
Tang (2021)	86.89	88.41	89.35	88.88	90.49	87.56	88.84	88.2
Liang and Liang (2022)	87.36	89.69	90.3	89.99	91.19	88.83	90.05	89.44
Wenninger et al. (2020)	89.54	90.29	91.04	90.66	91.74	89.7	91.27	90.02
Salim et al. (2019)	90.6	90.81	91.58	91.19	92.06	90.22	91.34	90.78
Ours	96.28	94.67	97.03	95.84	96.68	93.53	96.68	95.08



and the F1 score increased by 7.15%; the corresponding increase in the VREN dataset They are 7.49, 5.27, and 6.41%, respectively. After adding the generative adversarial network to the attention

model, the indicators of the two data sets have been further improved. Among them, the hit rate and F1 score of the Volleyball data set have increased by about 9.73 and 7.5%,

TABLE 3 Comparison and visualization of Hit Rate, Recall, Precision, and F1 Score indicators of different modules based on Volleyball and VREN datasets.

Module	Dataset							
	Volleyball dataset (Ibrahim et al., 2016)				VREN dataset (Nia et al., 2022)			
	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)
baseline	65.49	64.34	67.57	65.92	66.52	67.09	67.39	67.24
+satt	72.27	76.10	70.28	73.07	74.01	72.36	74.99	73.65
+gan	82.00	78.58	82.67	80.57	80.12	84.36	77.02	80.52
+satt gan(our)	95.81	93.72	95.71	94.70	96.15	94.85	96.15	95.49

“satt” is the self-attention mechanism, and “gan” is the generative adversarial network.

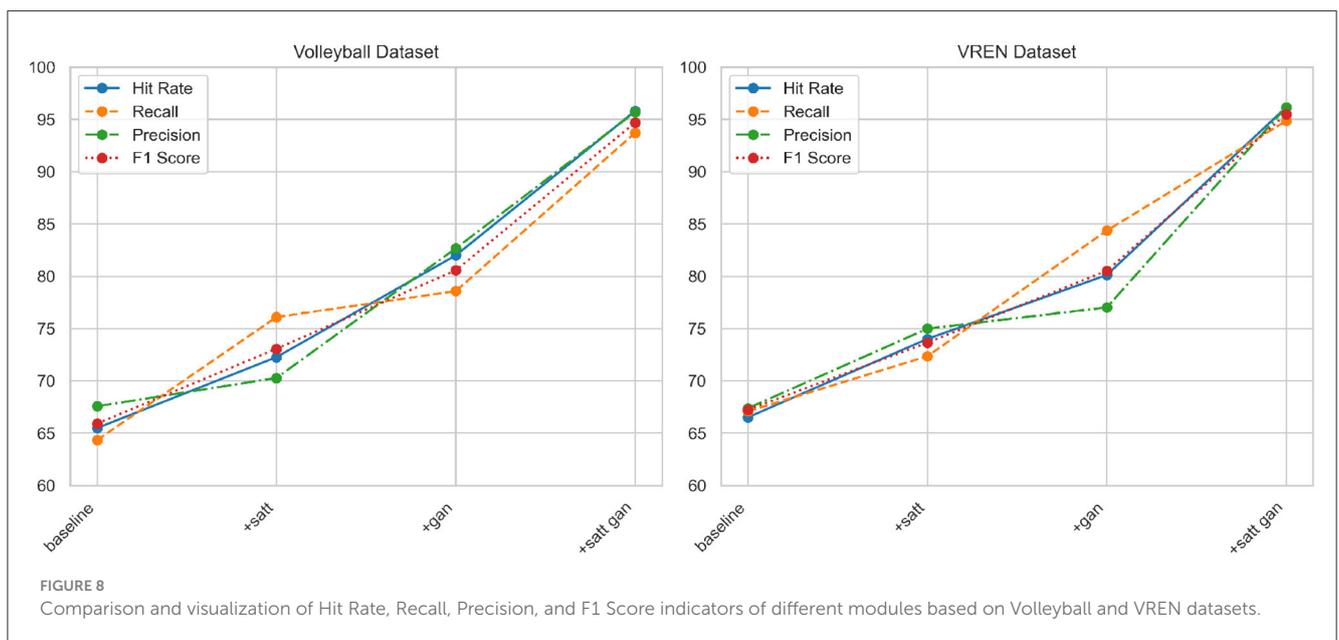


FIGURE 8 Comparison and visualization of Hit Rate, Recall, Precision, and F1 Score indicators of different modules based on Volleyball and VREN datasets.

TABLE 4 Comparison of Hit Rate, Recall, Precision, and F1 Score indicators of different modules based on UCF101 and MultiSports datasets.

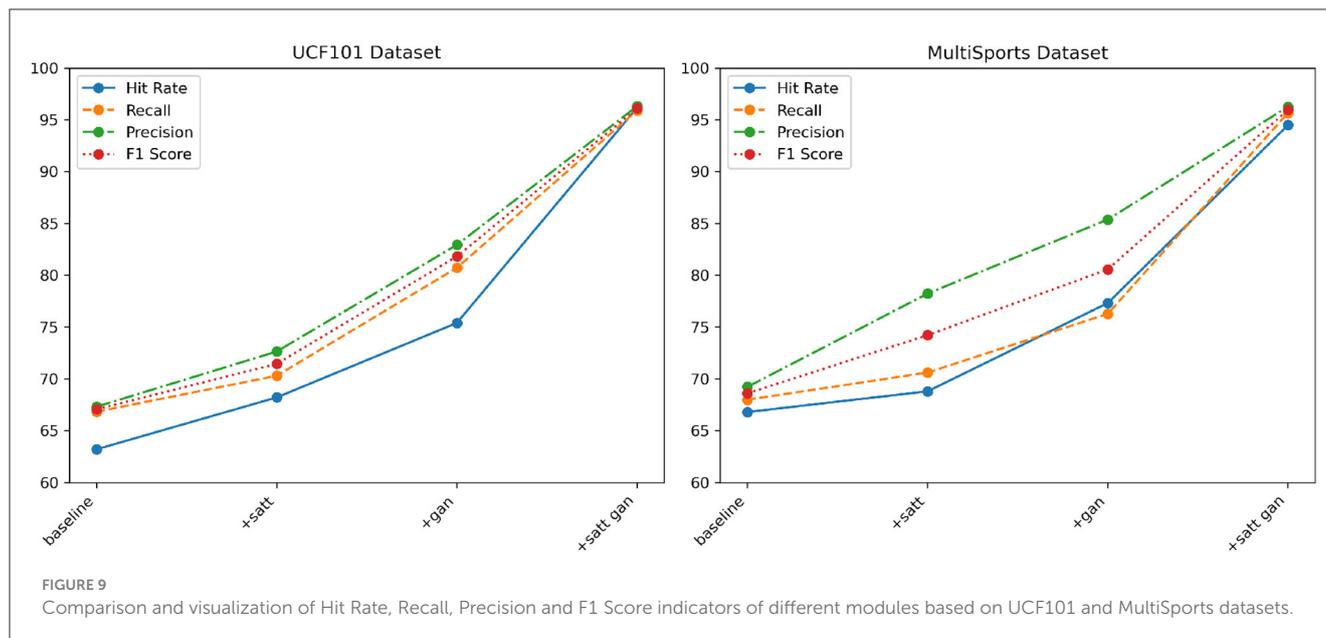
Module	Dataset							
	UCF101 dataset (Soomro et al., 2012)				MultiSports dataset (Li et al., 2021)			
	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)	Hit rate (%)	Recall (%)	Precision (%)	F1 Score (%)
baseline	63.21	66.84	67.32	67.07	66.81	68.0	69.24	68.61
+satt	68.22	70.31	72.65	71.46	68.81	70.62	78.24	74.23
+gan	75.41	80.73	82.94	81.82	77.33	76.29	85.39	80.58
+satt gan(our)	96.18	95.91	96.32	96.11	94.5	95.68	96.28	95.98

“satt” is the self-attention mechanism, and “gan” is the generative adversarial network.

respectively; Indicators increased by 6 to 7%. In the end, these two key modules were applied in series, not only achieved the highest hit rate of more than 95% on the two data sets, the precision index also exceeded 95 and 96%, and the recall rate was increased to 93.72 and 94.85% of the top level. This fully confirms the important role of attention mechanism and adversarial learning in improving the ability of deep network action recognition, and also highlights the advantages of our improved method in mining multi-modal features. At the same

time, we compared and visualized the results in Table 3, as shown in Figure 8.

From the data in Table 4 above, it can be seen that with the continuous optimization of the model structure in our proposed method, the action recognition ability on these two large-scale general-purpose action datasets UCF101 and MultiSports has been greatly improved. Specifically, in comparison with the baseline module, after only adding the self-attention module, the three core evaluation indicators on the MultiSports dataset, namely hit



rate, recall rate and F1 score, have been improved by more than 2%, respectively; UCF101 dataset The corresponding improvements on the above are even greater, reaching 6.78, 11.76, and 7.15%, respectively, which has verified the role of the attention mechanism in extracting cross-modal correlation features. After adding deep adversarial training on this basis, the improvement of evaluation indicators on the two data sets continues to expand. Among them, the three indicators of the UCF101 dataset all improved within the range of 2 to 9%; the corresponding indicators of the MultiSports dataset increased the most, reaching 10.52, 8.29, and 11.97%, respectively, which further verified how adversarial learning can effectively improve model generalization ability. Finally, the optimization model that integrates attention and confrontation mechanism is adopted, not only makes multiple indicators on UCF101 and MultiSports data sets break through the high level of about 94% for the first time, but also has a recall rate of more than 95.68% on the MultiSports data set; this shows the effectiveness of our method. The optimization effect has achieved generalizability on different types of large-scale action recognition tasks. We compared and visualized the results in Table 4, as shown in Figure 9.

In conclusion, the multimodal deep learning-based robot action recognition method proposed in this study demonstrates significant advantages in experiments conducted on multiple classic volleyball datasets and a large-scale diverse action dataset. By leveraging attention mechanisms to integrate visual and motion features, along with the incorporation of deep adversarial mechanisms to enhance model generalization, the accuracy and recall rate of action recognition have both been notably improved. Particularly, with the integration of the optimized model structure, our method achieves impressive recognition performance across all tested datasets, thus fully validating the reliability and potential of this approach in action recognition tasks.

Through detailed data comparison and analysis, we can clearly witness how the seamless integration of various modules within the model's structure drives the continuous enhancement of

recognition capabilities. This not only underscores the correctness of the deep learning architectural approach but also confirms the vital roles of attention mechanisms and adversarial learning in multimodal feature learning. While rooted in the context of volleyball robot requirements, experimental results indicate its promising applicability to other action recognition tasks, further showcasing the method's versatility.

In summary, this work successfully designs and implements a deep multimodal learning algorithm to optimize action recognition capabilities, laying down a methodological foundation for the advancement of robotic sports skills.

4. Conclusion

In preceding chapters, we provided an extensive account of the application of multimodal deep learning methods to enhance robotic cyclic motion skills. In this chapter, we delve into a comprehensive discussion of research outcomes, summarizing key findings from experiments, exploring the significance and contributions of this study, analyzing the strengths and limitations of our approach, and outlining potential avenues for future research.

Through meticulous experimentation and analysis, we observed substantial accomplishments in enhancing robotic skills via multimodal deep learning. The introduction of the cross-modal self-attention mechanism proficiently fuses information from distinct sensors, culminating in comprehensive scene perception. Leveraging Generative Adversarial Networks (GANs) imbues the model with superior data generation and training capabilities, enriching the diversity and practicality of skill training. The implementation of transfer learning further expedites skill augmentation, minimizing the temporal cost of relearning in new environments. The confluence of these modules facilitates remarkable skill enhancement across several pivotal metrics, presenting a positive contribution to the realm of sports robotics.

The significance of this study resides in its insightful and empirical contribution to the progression of cyclic motion robotics. The seamless integration of multimodal perception and deep learning not only elevates robotic prowess in volleyball matches but also ushers in novel prospects for intelligent sports competition and human-robot collaboration. Our research not only theoretically validates this approach but also substantiates its practical efficacy, offering a valuable reference for researchers in related domains.

Throughout this study, we harnessed the inherent advantages of multimodal perception, synergizing information from diverse sensors. This multimodal data processing strategy not only heightens model performance but also enhances robot scene awareness. Simultaneously, our research introduces pivotal technologies such as self-attention mechanisms, GANs, and transfer learning, fully harnessing the potential of deep learning and providing diverse tools and avenues for skill augmentation. However, we acknowledge certain limitations, such as potential model generalization issues stemming from experimental data distribution and the possible challenges and constraints in real-world applications.

Future research directions could encompass the expansion of our approach to diverse sports domains, unraveling the broader potential of multimodal perception and deep learning. Concurrently, optimizing model architectures and algorithms could enhance the efficacy and swiftness of skill augmentation. Furthermore, applying our approach to real volleyball match scenarios could authenticate its viability and efficacy in actual competition. Ultimately, we anticipate our continued research and practical efforts will contribute significantly to the advancement of sports robotics and intelligent sports competition.

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

MW: Conceptualization, Data curation, Project administration, Resources, Writing—original draft. ZL: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, 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 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.

References

- Ajakan, H., Germain, P., Laroche, H., Laviolette, F., and Marchand, M. (2014). Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*. doi: 10.48550/arXiv.1412.4446
- Chen, S., Cao, Y., Sarparast, M., Yuan, H., Dong, L., Tan, X., et al. (2020). Soft crawling robots: design, actuation, and locomotion. *Adv. Mater. Technol.* 5, 1900837. doi: 10.1002/admt.201900837
- Hong, C., Jeong, I., Vecchietti, L. F., Har, D., and Kim, J.-H. (2021). AI world cup: robot-soccer-based competitions. *IEEE Trans. Games* 13, 330–341. doi: 10.1109/TG.2021.3065410
- Hu, Y., Li, Z., and Yen, G. G. (2023). A knee-guided evolutionary computation design for motor performance limitations of a class of robot with strong nonlinear dynamic coupling. *IEEE Trans. Syst. Man Cybernet. Syst.* 53, 4429–4441. doi: 10.1109/TSMC.2023.3249123
- Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., and Mori, G. (2016). "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1971–1980. Available online at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Ibrahim_A_Hierarchical_Deep_CVPR_2016_paper.html
- Ji, Y., Li, Z., Sun, Y., Peng, X. B., Levine, S., Berseth, G., et al. (2022). "Hierarchical reinforcement learning for precise soccer shooting skills using a quadrupedal robot," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Kyoto: IEEE), 1479–1486. Available online at: <https://ieeexplore.ieee.org/abstract/document/9981984>
- Kautz, T., Groh, B. H., Hannink, J., Jensen, U., Strubberg, H., and Eskofier, B. M. (2017). Activity recognition in beach volleyball using a deep convolutional neural network: leveraging the potential of deep learning in sports. *Data Mining Knowledge Discov.* 31, 1678–1705. doi: 10.1007/s10618-017-0495-0
- Li, B., and Tian, M. (2023). Volleyball movement standardization recognition model based on convolutional neural network. *Comput. Intell. Neurosci.* 2023, 6116144. doi: 10.1155/2023/6116144
- Li, R., and Peng, B. (2022). Implementing monocular visual-tactile sensors for robot manipulation. *Cyborg Bionic Syst.* 2022, 9797562. doi: 10.34133/2022/9797562
- 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* (Montreal, QC), 13536–13545. Available online at: <https://ieeexplore.ieee.org/abstract/document/9711267>
- Liang, C., and Liang, Z. (2022). The application of deep convolution neural network in volleyball video behavior recognition. *IEEE Access* 10, 125908–125919. doi: 10.1109/ACCESS.2022.3221530
- Mi, Z., Jiang, X., Sun, T., and Xu, K. (2020). Gan-generated image detection with self-attention mechanism against gan generator defect. *IEEE J. Select. Top. Signal Process.* 14, 969–981. doi: 10.1109/JSTSP.2020.2994523

- Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing* 452, 48–62. doi: 10.1016/j.neucom.2021.03.091
- Olaniyan, O. T., Adetunji, C. O., Adeniyi, M. J., and Hefft, D. I. (2022). “Computational intelligence in iot healthcare,” in *Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics* (CRC Press), 297–310. Available online at: <https://www.taylorfrancis.com/chapters/edit/10.1201/9780367548445-19/computational-intelligence-iot-healthcare-olugbemi-olaniyan-charles-adetunji-mayowa-adeniyi-daniel-ingo-hefft>
- Oliff, H., Liu, Y., Kumar, M., Williams, M., and Ryan, M. (2020). Reinforcement learning for facilitating human-robot-interaction in manufacturing. *J. Manufact. Syst.* 56, 326–340. doi: 10.1016/j.jmsy.2020.06.018
- Oliveira, L. S., Moura, T. B. M. A., Rodacki, A. L. F., Tilp, M., and Okazaki, V. H. A. (2020). A systematic review of volleyball spike kinematics: implications for practice and research. *Int. J. Sports Sci. Coach.* 15, 239–255. doi: 10.1177/1747954119899881
- Salim, F. A., Haider, F., Tasdemir, S. B. Y., Naghashi, V., Tengiz, I., Cengiz, K., et al. (2019). “Volleyball action modelling for behavior analysis and interactive multi-modal feedback,” in *Proceedings of the 15th International Summer Workshop on Multimodal Interfaces* (Ankara), 50.
- Siedentop, D., and Van der Mars, H. (2022). *Introduction to Physical Education, Fitness, and Sport*. Human Kinetics.
- Siegel, J., and Morris, D. (2020). “Robotics, automation, and the future of sports,” in *21st Century Sports: How Technologies Will Change Sports in the Digital Age*, ed S. L. Schmidt (Springer), 53–72.
- So, J., Kim, U., Kim, Y. B., Seok, D.-Y., Yang, S. Y., Kim, K., et al. (2021). Shape estimation of soft manipulator using stretchable sensor. *Cyborg Bionic Syst.* 2021, 9843894. doi: 10.34133/2021/9843894
- Soomro, K., Zamir, A. R., and Shah, M. (2012). UCF101: a dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.
- Tang, J. (2021). An action recognition method for volleyball players using deep learning. *Sci. Prog.* 2021, 1–9. doi: 10.1155/2021/3934443
- Thuruthel, T. G., Shih, B., Laschi, C., and Tolley, M. T. (2019). Soft robot perception using embedded soft sensors and recurrent neural networks. *Sci. Robot.* 4, eaav1488. doi: 10.1126/scirobotics.aav1488
- Wang, H., Sahoo, D., Liu, C., Shu, K., Achananuparp, P., Lim, E.-P., et al. (2021). Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Trans. Multimedia* 24, 2515–2525. doi: 10.1109/TMM.2021.3083109
- Weiss, A., Wortmeier, A.-K., and Kubicek, B. (2021). Cobots in industry 4.0: a roadmap for future practice studies on human–robot collaboration. *IEEE Trans. Hum. Mach. Syst.* 51, 335–345. doi: 10.1109/THMS.2021.3092684
- Weninger, S., Link, D., and Lames, M. (2020). Performance of machine learning models in application to beach volleyball data. *Int. J. Comput. Sci. Sport* 19, 24–36. doi: 10.2478/ijcss-2020-0002
- Xia, H., Tracy, R., Zhao, Y., Fraise, E., Wang, Y.-F., and Petzold, L. (2022). “VREN: volleyball rally dataset with expression notation language,” in *2022 IEEE International Conference on Knowledge Graph (ICKG)* (Orlando, FL: IEEE), 337–346.
- Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). “Modeling tabular data using conditional gan,” in *Advances in Neural Information Processing Systems* 32. Available online at: <https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html>
- Zhang, J. (2019). Gradient descent based optimization algorithms for deep learning models training. *arXiv preprint arXiv:1903.03614*. doi: 10.48550/arXiv.1903.03614
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2020). A comprehensive survey on transfer learning. *Proc. IEEE* 109, 43–76. doi: 10.1109/JPROC.2020.3004555