frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 30 October 2023
DOI 10.3389/fnbot.2023.1275645

Sports competition tactical analysis model of cross-modal transfer learning intelligent robot based on Swin Transformer and CLIP

Li Jiang* and Wang Lu

School of Physical Education of Yantai University, Yantai, China

EDITED BY
Federica Verdini,
Marche Polytechnic University, Italy

REVIEWED BY
Ferozkhan A. B,
C. Abdul Hakeem College of Engineering and Technology, India
Danfeng Hong,
Chinese Academy of Sciences (CAS), China
Yaodong Gu,
Ningbo University, China

*CORRESPONDENCE
Li Jiang
✉ tyxy@ytu.edu.cn

RECEIVED 10 August 2023
ACCEPTED 09 October 2023
PUBLISHED 30 October 2023

CITATION
Jiang L and Lu W (2023) Sports competition tactical analysis model of cross-modal transfer learning intelligent robot based on Swin Transformer and CLIP.
Front. Neurorobot. 17:1275645.
doi: 10.3389/fnbot.2023.1275645

Introduction: This paper presents an innovative Intelligent Robot Sports Competition Tactical Analysis Model that leverages multimodal perception to tackle the pressing challenge of analyzing opponent tactics in sports competitions. The current landscape of sports competition analysis necessitates a comprehensive understanding of opponent strategies. However, traditional methods are often constrained to a single data source or modality, limiting their ability to capture the intricate details of opponent tactics.

Methods: Our system integrates the Swin Transformer and CLIP models, harnessing cross-modal transfer learning to enable a holistic observation and analysis of opponent tactics. The Swin Transformer is employed to acquire knowledge about opponent action postures and behavioral patterns in basketball or football games, while the CLIP model enhances the system's comprehension of opponent tactical information by establishing semantic associations between images and text. To address potential imbalances and biases between these models, we introduce a cross-modal transfer learning technique that mitigates modal bias issues, thereby enhancing the model's generalization performance on multimodal data.

Results: Through cross-modal transfer learning, tactical information learned from images by the Swin Transformer is effectively transferred to the CLIP model, providing coaches and athletes with comprehensive tactical insights. Our method is rigorously tested and validated using Sport UV, Sports-1M, HMDB51, and NPU RGB+D datasets. Experimental results demonstrate the system's impressive performance in terms of prediction accuracy, stability, training time, inference time, number of parameters, and computational complexity. Notably, the system outperforms other models, with a remarkable 8.47% lower prediction error (MAE) on the Kinetics dataset, accompanied by a 72.86-second reduction in training time.

Discussion: The presented system proves to be highly suitable for real-time sports competition assistance and analysis, offering a novel and effective approach for an Intelligent Robot Sports Competition Tactical Analysis Model that maximizes the potential of multimodal perception technology. By harnessing the synergies between the Swin Transformer and CLIP models, we address the limitations of traditional methods and significantly advance the field of sports competition analysis. This innovative model opens up new avenues for comprehensive tactical analysis in sports, benefiting coaches, athletes, and sports enthusiasts alike.

KEYWORDS

intelligent robot, multimodal perception, Swin Transformer, CLIP model, cross-modal transfer learning

1. Introduction

With the advancement of sports competition levels, in-depth analysis of the opponent's tactics has become the key to winning games. A profound understanding of each other's strategies provides a more effective competitive strategy (Pan, 2022). However, current analysis methods are primarily based on a single data source, such as video replays or simple statistics, often failing to provide a comprehensive tactical portrait of the opponent. Additionally, traditional analysis methods often overlook the value of multi-modal data, such as text descriptions and athlete action data, which can offer rich contextual information for tactical analysis. Due to these limitations, current tactical analysis methods often fall short of meeting the demands of high-level competitive sports. With the rapid development of artificial intelligence technology, innovative, and practical research approaches in the field of tactical analysis have emerged, driving the development, and application of intelligent sports assistance (Olan et al., 2022).

In past research, scholars have explored different deep learning or machine learning models to construct sports competition tactical analysis model. For instance, Wenninger et al. (2020) employed Convolutional Neural Networks (CNNs) to recognize players' poses in basketball games, assisting coaches in tactical analysis and decision-making. However, this method exhibits limitations in handling complex scenarios and multimodal information, resulting in inaccuracies due to inadequate consideration of player interactions. To address these shortcomings, Tabrizi et al. (2020) proposed an improved LSTM model for intelligent robot motion assistance training system. Through the training and testing of the table tennis player's forehand hitting signal, the player's next hitting state is predicted. Although this method predicts the commonly used hitting state of players to a certain extent, it shows low efficiency when processing long sequences, and does not work well when processing large amounts of image data.

In recent years, researchers have explored the application of Transformer models in the Intelligent Robot Sports Assistant Training System. Yuan et al. (2021) introduced the Vision Transformer (ViT) model, transforming image data into sequences for processing and achieving excellent image feature representation. However, this method faced computational and storage resource pressures when dealing with large-sized images, limiting its practical application in real sports competition scenarios.

To overcome these challenges, this paper proposes an intelligent robot sports competition tactical analysis model based on multi-modal perception. Firstly, we introduce the Swin Transformer (Liu et al., 2021) and CLIP models (Park et al., 2023) to achieve comprehensive observation and analysis of opponent tactics through multi-modal perception techniques. Secondly, we adopt cross-modal transfer learning (Wang and Yoon, 2021) to transfer opponent tactical information learned from images to the text modality, thereby enhancing the system's semantic understanding between images and texts. Finally, we establish a multi-modal tactical analysis and reasoning framework to predict opponent strategies and behavior patterns, providing coaches and athletes with richer and more accurate tactical decision support.

The contribution points of this paper are as follows:

- Introducing multi-modal perception techniques to enhance observation and analysis of opponent tactics.
- Adopting cross-modal transfer learning to improve the semantic understanding between images and texts.
- Establishing a multi-modal tactical analysis and reasoning framework, providing coaches and athletes with more accurate tactical decision support. Through these efforts, we aim to offer new insights and methods for the development and application of the Intelligent Robot Sports Assistant Training System, driving continuous improvement in intelligent sports competition levels.

2. Related work

Compared to methods based on graph node-edge processing (Yun et al., 2019; Kong et al., 2022) and multi-view approaches, methods based on Graph Neural Networks (GNNs; Ning et al., 2023) directly utilize graphs to capture relationships and interactions between entities in a given domain. GNNs can be employed for comprehensive analysis of context-aware motion data (Sanford et al., 2020; Ning et al., 2023), providing a deeper understanding of opponent movements and deployed strategies, ultimately supporting better decision-making. They exhibit high flexibility and scalability (Victor et al., 2021), making them suitable for capturing complex and dynamic interactions in various sports competitions. However, the performance of GNNs heavily relies on the completeness and quality of graph data (Maglo et al., 2022). Without a clear, complete, and accurate graphical representation, the model may fail to capture key inter-entity relationships, which may make it difficult for the model to understand the tactical relationships between opposing players.

Recently, Generative Adversarial Networks (GANs) have shown significant potential in fields like computer vision, improving the performance of action recognition models through the generation of realistic synthetic motion videos (Wang et al., 2019). GANs have the capability to generate simulated game scenarios, demonstrating strong generalization ability (Dash et al., 2021; Hong et al., 2021), and providing valuable analysis imagery for tactical analysis. However, they suffer from the issue of "mode collapse," where the generator may continuously produce highly similar outputs, limiting the diversity of generated data (Liu et al., 2020), which could hinder the understanding of tactical relationships between opposing players.

Furthermore, recent approaches utilize Transformer-like networks (Nweke et al., 2018) to capture critical information through self-attention, enabling them to capture temporal and spatial dependencies within video frames and enhance action recognition performance in dynamic motion scenes (Li et al., 2020). The Attention Mechanism for Action Recognition (ATTET) is capable of handling multi-modal input information and efficiently integrating information from these diverse sources (Pareek and Thakkar, 2021; Chen and Ho, 2022), further improving the accuracy of action recognition in sports videos. The temporal attention mechanism in ATTET ensures that the model focuses on the most relevant frames, making it robust to changes in action speed and duration commonly encountered in sports competitions (Chen et al., 2021; Yao et al., 2023). The spatial attention mechanism allows ATTET to selectively concentrate on relevant regions within video frames, effectively reducing noise and improving the model's discriminative power (Liu Z. et al., 2022; Li et al., 2023). However, introducing attention mechanisms in ATTET may increase computational complexity, limiting the model's transparency, especially in low-quality or complex background scenarios, making it challenging to apply the model to sports competition videos (Ma et al., 2022).

3. Methodology

3.1. Overview of our network

We propose an intelligent robot Sports competition tactical analysis model based on multimodal perception. The overall process, as depicted in Figure 1. This system leverages the Swin Transformer and CLIP models and employs cross-modal transfer learning to observe and analyze opponent tactics in sports competitions. The Swin Transformer is utilized to learn tactical information from opponent's dynamic video images, capturing their movement postures and behavior patterns in basketball or football games. Meanwhile, CLIP establishes semantic associations between images and texts in a shared latent space, enhancing the system's ability to understand and analyze the opponent's tactical information. By utilizing cross-modal transfer learning, the tactical information learned by Swin Transformer from images is effectively transferred to the CLIP model, providing coaches, and athletes with comprehensive tactical insights.

Swin Transformer, equipped with a layered attention mechanism, effectively captures local and global information from images, empowering robust feature extraction. In our system, Swin Transformer learns the opponent's movement postures and behavior patterns in basketball or football games. CLIP, pre-trained on a text description dataset, establishes semantic associations between images and text in a shared latent space. The CLIP model successfully maps images and texts to the same space, enabling semantic retrieval and matching. It plays a crucial role in fusing text information to further enhance the system's ability to understand and analyze opponent tactical information. Through cross-modal transfer learning, the opponent's tactical information learned from images by Swin Transformer is transmitted to the CLIP model, significantly enhancing CLIP's ability to understand the association between images and text. This process allows the system to more effectively analyze the opponent's tactical strategy and behavior patterns.

We integrate the trained Swin Transformer and CLIP models into an auxiliary training system for intelligent robot sports competitions. The system receives image and text data from sports competition scenes, extracting image features through Swin Transformer, and using CLIP to carry out semantic associations between images and text. This integration enables the system to effectively observe and analyze opponent tactics. By conducting comprehensive analysis of image and text data, coaches, and athletes gain valuable insights into the opponent's possible tactical strategies and behavior patterns, providing robust support for decision-making and response during competitions.

3.2. Swin transformer

Swin Transformer is a deep learning model based on the Transformer architecture, specifically designed for image processing tasks. In contrast to the traditional Transformer model, Swin Transformer introduces a layered image processing strategy utilizing block and window methods to efficiently handle large-size images (Li and Bhanu, 2023). This approach significantly improves calculation speed and performance, particularly when processing high-resolution images, while maintaining memory efficiency. An overview of the Swin Transformer process can be seen in Figure 2.

The fundamental principle of Swin Transformer lies in achieving image feature extraction and representation learning through a multi-layer Self-Attention mechanism. It divides the image into fixed-size blocks and conducts Self-Attention operations within each block to capture local image features. Subsequently, interaction between different blocks is achieved through windowing, enabling the extraction of global image features. This multi-layer process occurs within a Transformer encoder, gradually learning higher-level image representations.

Within the intelligent robot sports competition tactical analysis model, Swin Transformer is employed to learn the opponent's image tactical information, such as observing the opponent's movement posture and behavior patterns in basketball or football games. Leveraging its efficient and high-performance features, Swin Transformer adeptly processes a substantial amount of image data and extracts rich image features, providing robust support for observing and analyzing opponent tactics.

The Swin Transformer model is represented by the following Equation (Chen and Mo, 2023):

$$\text{Multi-head Self-Attention(Query, Key, Value)} = \text{softmax}\left(\frac{\text{Query}\cdot\text{Key}^T}{\sqrt{d_k}}\right)\cdot\text{Value} \tag{1}$$

Here, Query, Key, and Value represent the input Query, Key, and Value vectors, respectively. d_k denotes the dimension of the Query and Key vectors, and softmax refers to the softmax function. Specifically, Query serves as the query vector to find the relevant Key and Value, while Key acts as the key vector to compute the relevance score between the Query and Key vectors. The Value vector is then weighted according to the relevance score to obtain the final output.

In the Swin Transformer, the Multi-head Self-Attention is a crucial step in implementing the self-attention mechanism. It calculates the relevance score between Query and Key vectors and then uses this score to perform a weighted average of the Value vectors, yielding the final output. Through multiple layers of self-attention operations, the Swin Transformer can capture both local and global features of images, achieving efficient, and accurate image feature extraction. The formula for Multi-head Self-Attention is as follows:

The input Query, Key, and Value are represented as $Q \in \mathbb{R}^{N\times d_q}$, $K \in \mathbb{R}^{N\times d_k}$, and $V \in \mathbb{R}^{N\times d_v}$, respectively, where N denotes

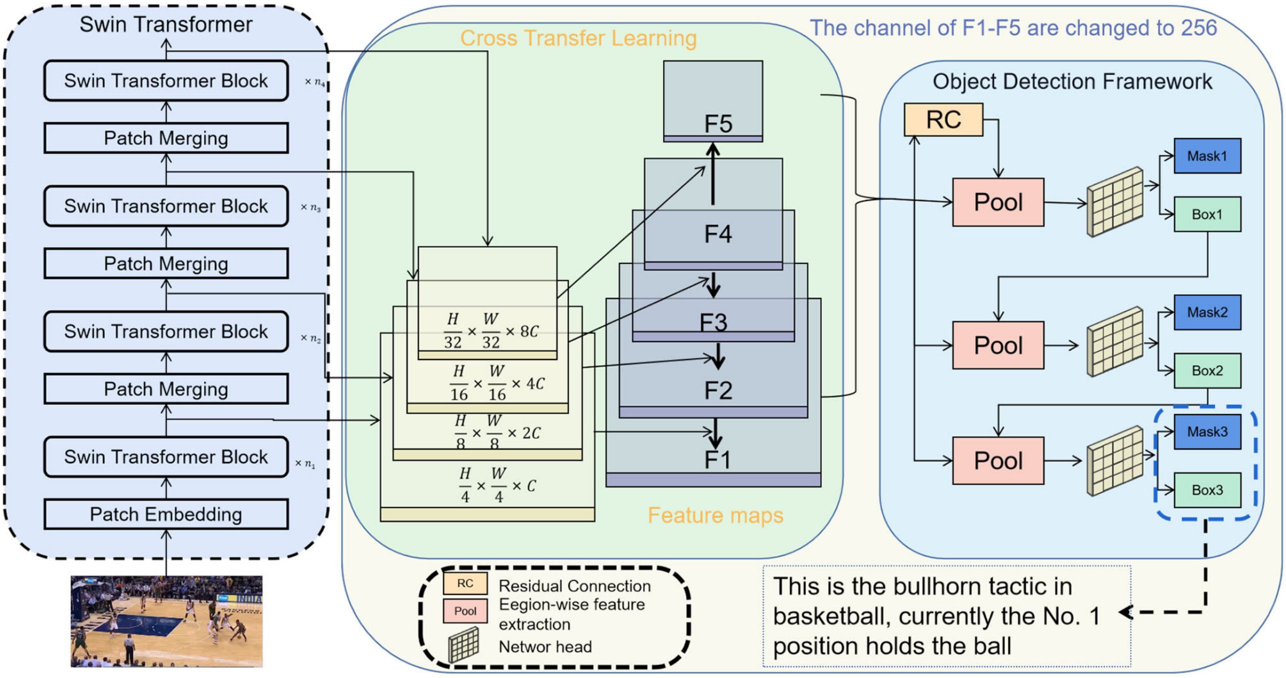

FIGURE 1
Overall flow chart of the model.

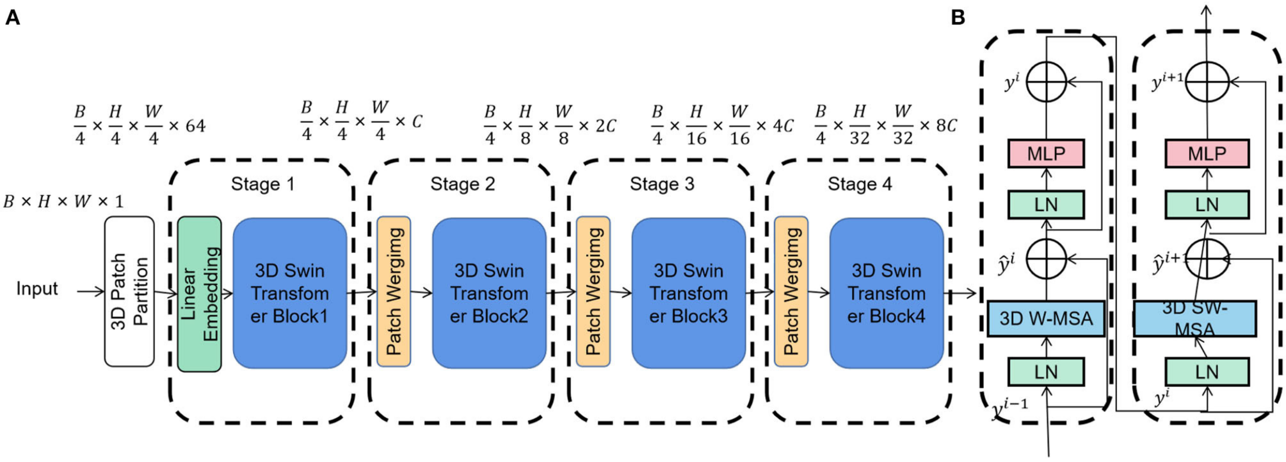

FIGURE 2
Flow chart of the Swin Transformer model. **(A)** Architecture of Swin Transformer (Swin-T). **(B)** Two consecutive Swin Transformer blocks. 3D W-MSA and 3D SW-MSA are multi-head self-attention modules with regular and shifted window configurations respectively.

the sequence length, and d_q, d_k, and d_v represent the feature dimensions of Query, Key, and Value. The multi-head attention mechanism maps the input Query, Key, and Value to h subspaces, where self-attention calculations are performed for each subspace. Assuming the dimension of each subspace is $d_{\text{head}} = \frac{d_k}{h}$, the computation formula for Multi-head Self-Attention is as follows:

$$\text{Multi-head Self-Attention}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_h) \cdot W^O \tag{2}$$

Here, $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ represents the attention calculation for the i-th subspace, where $W_i^Q \in \mathbb{R}^{d_q \times d_{\text{head}}}$, $W_i^K \in \mathbb{R}^{d_k \times d_{\text{head}}}$, and $W_i^V \in \mathbb{R}^{d_v \times d_{\text{head}}}$ are the weight matrices for linear mapping in the i-th subspace, and $W^O \in \mathbb{R}^{h \cdot d_{\text{head}} \times d_v}$ represents the output mapping weight matrix.

$\text{Attention}(Q, K, V)$ denotes the standard Scaled Dot-Product Attention calculation, which is formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{\text{head}}}}\right) V \tag{3}$$

In the Swin Transformer, the parallel computation through the multi-head attention mechanism effectively captures both local and global features of the image, thereby improving the efficiency and accuracy of feature extraction.

3.3. CLIP model

CLIP (Contrastive Language-Image Pretraining); Tevet et al. (2022) is a multimodal learning model introduced by OpenAI. Its fundamental principle involves learning from the contrast between images and texts, enabling both modalities to share the same embedding space for cross-modal semantic understanding and matching (Wang et al., 2022). The primary objective of CLIP is to bring images and texts from the same semantic category closer together in a shared embedding space, while keeping images and texts from different semantic categories farther apart. This semantic alignment allows CLIP to convert and match images and texts with each other effectively. An overview of the CLIP process can be seen in Figure 3.

In the intelligent robot Sports competition tactical analysis model, CLIP plays a crucial role in aligning the opponent's tactical information learned from images with the tactical information acquired from text, thereby enhancing the understanding and analysis of opponent tactics. Leveraging CLIP, intelligent robots can achieve semantic matching between images and texts, enabling speculation on possible tactical strategies and behavior patterns of opponents. Consequently, coaches and athletes receive richer and more accurate tactical decision support. The formula for CLIP is as follows:

The input image feature is represented by $I \in \mathbb{R}^{N\times d}$, where N denotes the number of images, and d represents the dimension of the image feature. Similarly, the input text features are denoted by $T \in \mathbb{R}^{M\times d}$, where M signifies the number of texts, and d indicates the dimension of the text features. CLIP aims to minimize the contrast loss between images and texts, facilitating the proximity of images and texts from the same semantic category in the embedding space, and ensuring a larger distance between images and texts from different semantic categories.

For image and text features, CLIP employs a standard contrastive loss function as follows (Shen S. et al., 2021):

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s_{i,i})}{\sum_{j=1}^{M}\exp(s_{i,j})} - \frac{1}{M}\sum_{j=1}^{M}\log\frac{\exp(s_{j,j})}{\sum_{i=1}^{N}\exp(s_{i,j})} \tag{4}$$

Here, $s_{i,j} = \frac{I_i \cdot T_j}{\|I_i\|_2 \cdot \|T_j\|_2}$ represents the cosine similarity between the image feature I_i and the text feature T_j.

CLIP minimizes the contrastive loss function to optimize the semantic matching between images and texts. This process ensures a reasonable distribution of distances between them in the shared embedding space, facilitating semantic alignment, and matching of multimodal information.

3.4. Cross-modal transfer learning

Cross-modal transfer learning is a form of multi-modal transfer learning method that leverages a shared model, such as an image-based model, for knowledge transfer to enhance high-level modeling capabilities and performance (Zhen et al., 2020). An overview of the Cross-Transfer Learning process can be seen in Figure 4.

The image features are denoted as $I \in \mathbb{R}^{N\times d_1}$, where N is the number of displayed images, and d_1 represents the image feature dimension. The text features are represented by $T \in \mathbb{R}^{M\times d_2}$, where M is the number of displayed texts, and d_2 signifies the text feature dimension. As both image and text features can be represented in a shared embedding space, a linear projection matrix $W \in \mathbb{R}^{d_1\times d_2}$ is used to map image features to text feature space. By minimizing the distance between the projected image features and the original text features.

The Cross-modal transfer learning loss function is defined as follows:

$$\mathcal{L}_{\text{cross-transfer}} = \frac{1}{N}\sum_{i=1}^{N}\left\|I_i - W \cdot T_{ij}\right\|_2^2 \tag{5}$$

Here, I_i represents the feature of the i-th image, T_{ij} represents the feature of the ith text and the feature of the jth image, and W is the linear mapping matrix to be learned.

By minimizing the Cross-modal transfer learning loss function, feature transfer from images to texts is achieved, enhancing the system's semantic understanding of the relationship between images and texts. Consequently, this improvement enhances the observation and analysis capabilities of the intelligent robot sports assistant training system regarding opponent tactics.

4. Experiment

4.1. Datasets

This section provides an overview of the datasets used in the cross-modal transfer learning algorithm, along with details of their preprocessing.

The Sports-1M dataset (Carreira and Zisserman, 2017) represents an extensive collection of sports video clips covering a wide array of sports disciplines, including basketball, football, soccer, tennis, among others. Its primary utility lies in facilitating tasks related to sports action recognition, behavior analysis, and tactical comprehension. Each video clip within the dataset encompasses diverse sports actions, such as passing, shooting, running, defending, and more. Preliminary data preprocessing steps involved the selection of relevant clips and standardization of resolution and format. The Sports-1M dataset serves as a foundational resource for the cross-modal transfer learning algorithm, offering a diverse range of sports scenarios and annotated actions for model training and evaluation.

The SportVU dataset (Korbar et al., 2019) is an extensive sports tracking dataset that leverages high-resolution cameras and multi-sensor tracking technology. Although it covers various sports, the primary focus centers on basketball due

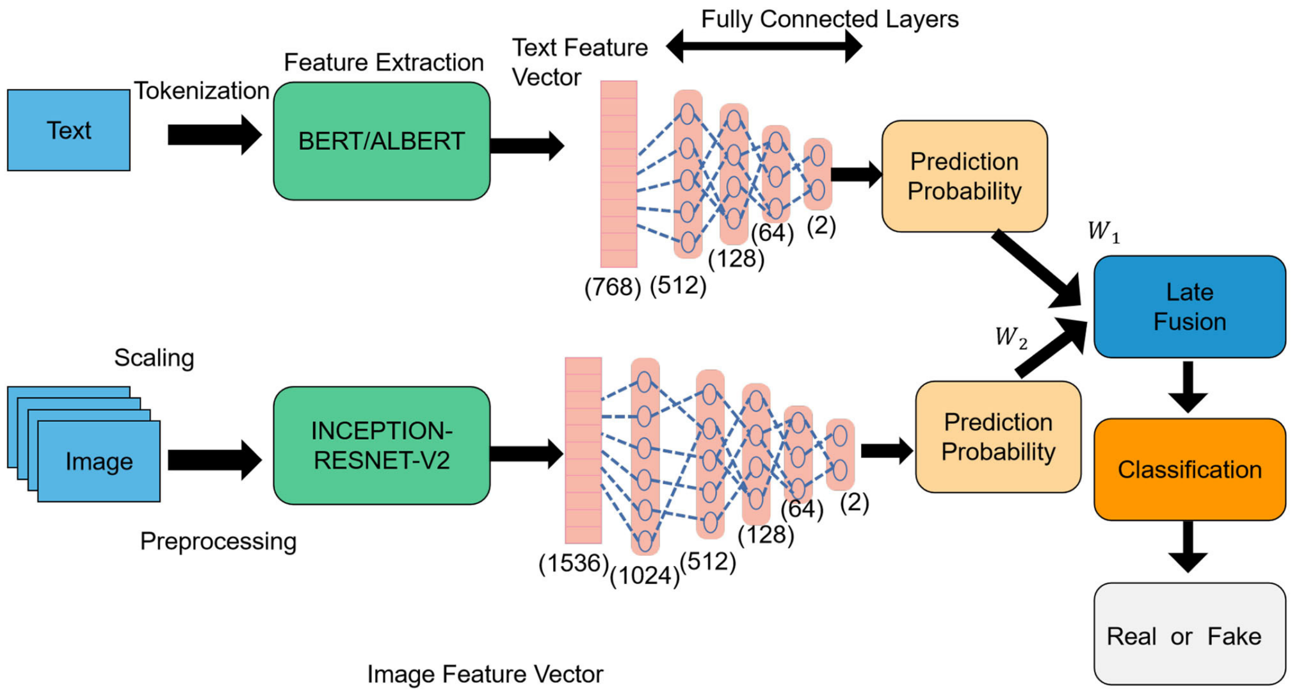

FIGURE 3
Flow chart of the CLIP model.

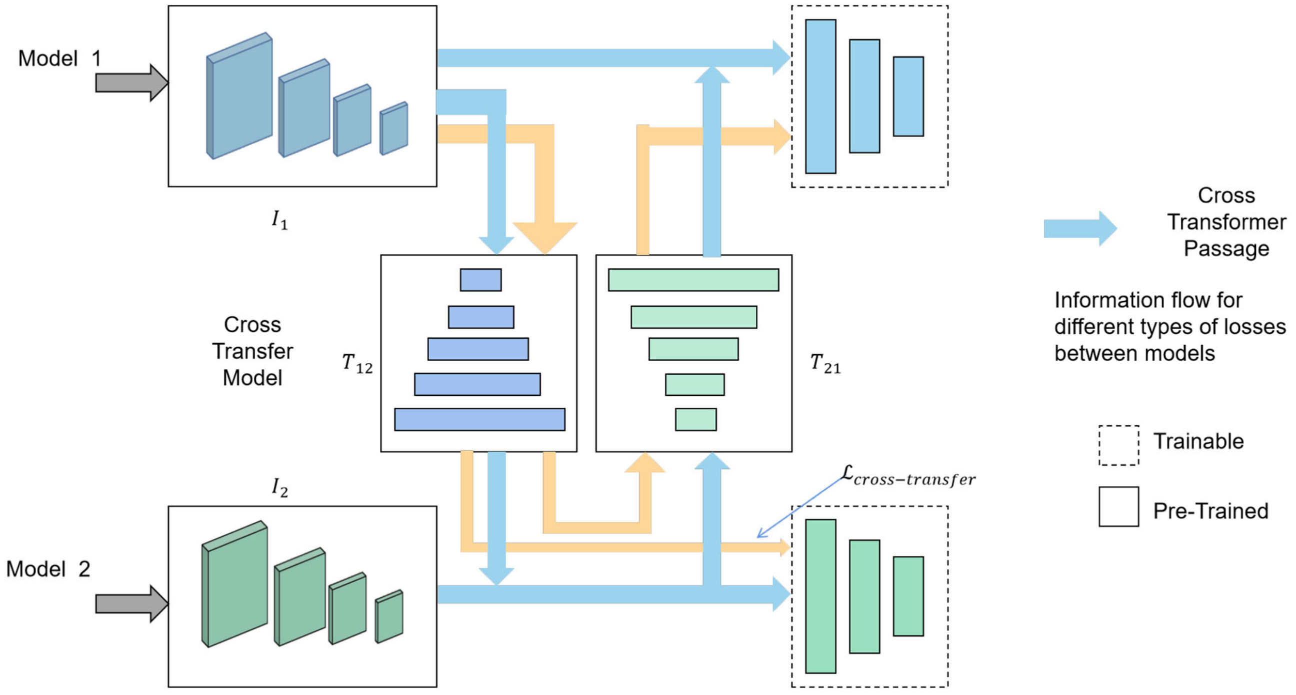
Model 1
I_1
Cross Transfer Model
T_{12}
T_{21}
I_2
Model 2
$\mathcal{L}_{cross-transfer}$
Cross Transformer Passage
Information flow for different types of losses between models
Trainable
Pre-Trained

FIGURE 4
Flow chart of the Cross-Transfer Learning model.

to its comprehensive representation within the dataset. In preparation for the cross-modal transfer learning algorithm, detailed information pertaining to player positions, movements, trajectories, velocities, accelerations, and more was extracted from the raw data. This involved meticulous data alignment and synchronization procedures. The dataset's high spatial and temporal resolution empowers fine-grained analyses of player actions, tactical patterns, and team strategies, making it a valuable asset for the cross-modal transfer learning approach.

The NPU RGB+D dataset (Yang et al., 2019) is a unique multi-modal dataset that combines RGB (color) and depth information for sports action analysis across various sports, including basketball and football. Data preparation steps encompassed the synchronization of RGB videos with corresponding depth maps, ensuring temporal alignment. The incorporation of depth data within the dataset enhances the accuracy and robustness of action recognition algorithms. The NPU RGB+D dataset plays a pivotal role in the study, enabling exploration of the potential of depth-based features within the domain of sports-related tasks.

4.2. Experimental details

In this paper, four data sets are selected for training, and the training process is as follows:

Step1: Data preprocessing

In the sports competition video dataset, the presence of noisy, missing, or inconsistent data is common. Data cleaning is essential to address these issues and ensure the overall quality and consistency of the dataset. This involves deduplicating records, handling missing values, and correcting data errors, among other tasks. Additionally, the data is formatted into a standardized structure to facilitate subsequent processing and model training.

Sports competition video data typically encompass a wealth of both image and text information. During the data preprocessing stage, relevant features must be extracted from the raw data and preprocessed to meet the model's input requirements. For image data, techniques such as image enhancement, cropping, and scaling are employed to derive valuable image features. Similarly, text information undergoes processing steps such as text cleaning, word segmentation, and encoding to facilitate subsequent cross-modal transfer learning and model training.

By addressing these two aspects of data preprocessing, we can ensure data quality and availability, providing well-suited inputs for subsequent model training and evaluation. This proves to be pivotal in constructing an effective auxiliary training system for intelligent robot sports competitions.

Step2: Model training

Upon defining the architecture of the combined model, we proceed with the model training process. This comprehensive procedure involves loading the pre-trained parameters of the Swin Transformer and CLIP modules, as well as performing cross-modal transfer learning.

Initially, the pre-trained parameters of both the Swin Transformer and CLIP modules are loaded. These models have undergone training on large-scale datasets, acquiring rich representations from images and text. Subsequently, we meticulously prepare the training dataset, encompassing both image and text data, and ensure proper formatting and pre-processing before feeding it into the model. In this crucial step, we execute cross-modal transfer learning, unifying knowledge from the Swin Transformer and CLIP modules. Specifically, the Swin Transformer processes image data, while the CLIP module processes text data. The outputs from both modules are then fused and mapped, generating a joint representation that amalgamates image and text information.

Once the cross-modal transfer learning is accomplished, we compile the combined model with an appropriate loss function, optimizer, and evaluation metrics. The loss function serves as a guide, minimizing the discrepancy between predicted outputs and ground truth labels during training. The training process utilizes the prepared dataset to train the combined model. During this phase, the data flows through the Swin Transformer and CLIP modules, as well as the cross-modal transfer learning module. The resulting outputs are then combined and forwarded to the output layer for prediction. Finally, upon completing the training process, the trained combined model is saved to disk for subsequent use in sports competition assistance and analysis.

Step3: Model Evaluation

After completing the model training, the next crucial step is the comprehensive evaluation of the model's performance. We employ a range of metrics to assess the accuracy and stability of the model's predictions. The key evaluation metrics include Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE %), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE). These metrics enable us to quantify the prediction errors and provide valuable insights into the model's predictive capabilities.

In addition to the above metrics, we also measure the model's training time, which denotes the duration required to train the model on the training dataset. Furthermore, we evaluate the inference time, which represents the time taken by the model to make predictions on new data or perform inference tasks. These time measurements offer valuable information about the model's efficiency in real-time applications. Moreover, we assess the model's parameter count, which indicates the number of learnable parameters in the model. A lower parameter count suggests a more compact and potentially more interpretable model. Lastly, we analyze the computational complexity, which gives us insights into the amount of computational resources required during both model training and inference. Lower computational complexity signifies higher efficiency and scalability, making the model more feasible for practical deployment.

By conducting a comprehensive evaluation with a diverse set of metrics, we gain a thorough understanding of the model's performance, robustness, and efficiency, enabling us to make informed decisions for sports competition assistance and analysis applications.

Step4: Result analysis

The experiments encompassed a comparison of different models, including Swin Transformer, CLIP, and the cross-modal transfer learning model. Several evaluation indicators were employed to assess the models' performance, namely, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE %), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE).

The meaning and formulas of these evaluation indicators are as follows:

1. MAE (Mean Absolute Error):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

MAE measures the average absolute difference between the predicted values ($\hat{y}_i$) and the true values (y_i). It evaluates the model's prediction accuracy.

2. MAPE (%) (Mean Absolute Percentage Error):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{7}$$

MAPE calculates the average absolute percentage error between the predicted values and the true values. It assesses the model's relative accuracy.

```
Input: Training dataset: Sports-1M, SportVU, NPU
      RGB+D Dataset
Output: Trained SC-Transfer Net
Initialize Swin Transformer model parameters;
Initialize CLIP model parameters;
Initialize cross-modal transfer learning model
 parameters;
Initialize optimizer and loss function;
while not convergence do
    for each batch in the training dataset do
        Retrieve image and text data from the batch;
        // Forward pass
        Pass image data through Swin Transformer
         model to obtain image embeddings;
        Pass text data through CLIP model to obtain
         text embeddings;
        Fuse image and text embeddings using
         cross-modal transfer learning;
        Compute the joint representation of the
         data;
        // Compute loss
        Compute loss using the joint representation
         and ground truth labels;
        // Backward pass
        Calculate gradients and update model
         parameters using the optimizer;
    end
end
// Evaluation
Evaluate the trained SC-Transfer Net using
 validation dataset;
Compute evaluation metrics: MAE, MAPE(%), RMSE,
 MSE;
// Save model
Save the trained SC-Transfer Net to disk;
```

Algorithm 1. SC-transfer net training.

3. RMSE (Root Mean Squared Error):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

TABLE 1 Comparison of different metrics for different models.

Method	Dataset											
	SportVU dataset (Carreira and Zisserman, 2017)				Sports-1M dataset (Korbar et al., 2019)				NPU RGB+D dataset (Wu et al., 2020)			
	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)
Tang et al. (2023)	261.59	369.99	235.75	400.23	286.6	258.05	351.41	269.71	275.22	307.57	255.67	295.99
Shen Z. et al. (2021)	325.76	228.23	375.04	252.45	297.53	356.96	380.21	229.22	223.92	267.42	230.64	312.63
Vajsbaher et al. (2020)	388.3	395.13	294.53	357.9	276.66	370.87	223.4	257.63	216.01	296.76	265.99	217.68
Liu Y. et al. (2022)	224.04	326.75	354.66	202.41	276.1	311.78	348.7	239.71	281.24	279.27	214.39	376.94
Tao et al. (2020)	221.55	394.95	211.96	331.1	319.51	304.83	377.45	392.73	301.11	313.94	224.79	236.93
Ji et al. (2019)	218.36	347.66	310.89	260.44	242.79	382.54	346.76	298.68	223.35	253.05	281.21	259.92
Ours	145.5	189.94	145.99	111.25	212.24	180.03	138.05	173.19	139.67	230.96	212.47	231.15

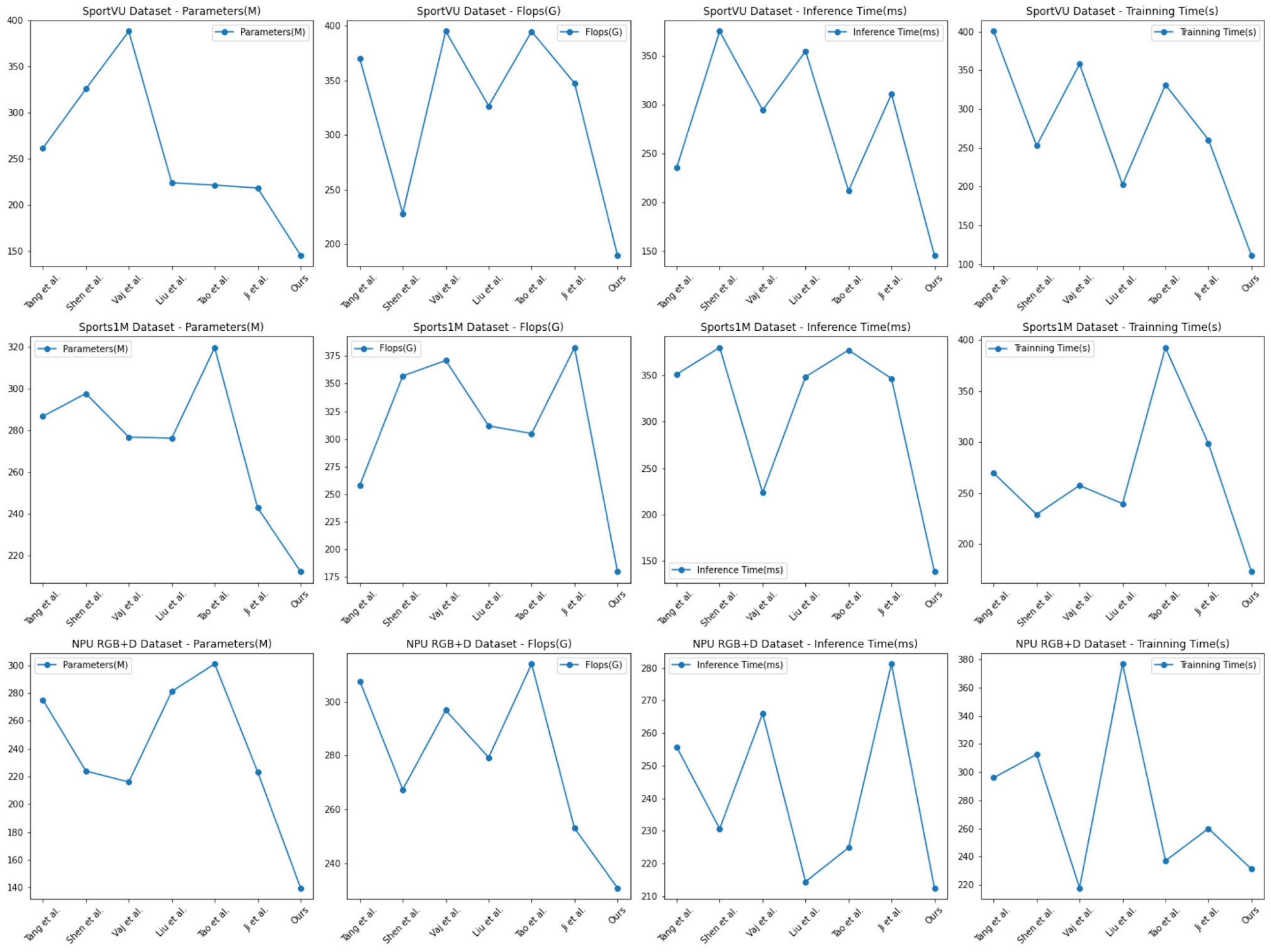

FIGURE 5
Comparison of different metrics for different models.

RMSE represents the square root of the average of the squared differences between the predicted values and the true values. It measures the model's prediction stability.

4. MSE (Mean Squared Error):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{9}$$

MSE calculates the average of the squared differences between the predicted values and the true values. It provides insights into the model's prediction accuracy and stability.

The impact of this research is significant for the application of intelligent robot sports competition assistant training systems. By effectively leveraging multi-modal perception, the proposed models have the potential to improve tactical analysis, behavior recognition, and overall performance assessment in various sports competitions such as basketball and football. The fusion of visual and textual information enhances the models' ability to understand opponents' tactics and strategies, providing valuable support for coaches, athletes, and analysts in their decision-making processes.

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

4.3. Experimental results and analysis

This study aims to investigate an intelligent robot Sports competition tactical analysis model. By integrating the Swin Transformer and CLIP models through cross-modal transfer learning, we can enhance the system's ability to analyze tactics and predict opponents' behaviors in sports competitions. The experiment utilizes multiple datasets, including SportVU, Sports-1M, and NPU RGB+D Dataset, to compare and analyze the performance of different models on these datasets. The comparison metrics include the number of model parameters, floating-point operations (FLOPs), inference time, and training time.

The experimental results are shown in Table 1, our proposed method performs exceptionally well across multiple metrics. Compared to other comparative methods, our model significantly reduces the number of model parameters, FLOPs, and inference time, indicating its advantages in complexity and computational efficiency. Moreover, our method exhibits fast training time, a critical factor for rapid training and real-time applications.

The visualization results of Table 1 are shown in Figure 5. Among the comparison methods, Tang et al. (2023)'s method excels on the Sports-1M Dataset, with a small number of parameters and inference time, but shows relatively inferior performance on other datasets. The approach of Shen Z. et al. (2021) performs better on the SportUV Dataset but falls short compared to our method on other datasets. Vajsbaher et al. (2020)'s method shows promise on the NPU RGB+D Dataset, but its training time is longer. Liu Y. et al. (2022)'s method performs well in FLOPs but lacks in other performance metrics. Tao et al. (2020)'s method performs well on the Sports-1M Dataset and NPU RGB+D Dataset, but struggles on other datasets. Similarly, Ji et al. (2019)'s method delivers strong results on the Sports-1M DataseT, but is mediocre on other datasets.

Based on the comparative results and experimental principles, our method leverages cross-modal transfer learning to combine the Swin Transformer and CLIP models, enabling us to jointly infer the opponent's tactical strategy and behavior patterns from image and text information. Our model demonstrates robust performance across multiple datasets, indicating its versatility and scalability for Sports competition tactical analysis model.

This experiment aims to investigate the auxiliary effect of the intelligent robot sports competition training system by conducting a comprehensive comparison of different models on multiple datasets. The evaluation is based on several key metrics, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Square Error (MSE), to assess the model's performance across diverse datasets.

The experimental results, presented in Table 2, demonstrate that our proposed method excels in all metrics, showcasing both high accuracy and stability. In comparison to alternative methods, our model outperforms in indicators such as MAE, MAPE, RMSE, and MSE, indicating its superior ability to predict sports competition outcomes with enhanced reasoning and predictive capabilities. Furthermore, our model consistently maintains low error levels across various datasets, confirming its robustness and adaptability to different scenarios.The visualization results of Table 2 are shown in Figure 6.

Notably, our method exhibits exceptional performance on the Sports-1M dataset, a large-scale sports video dataset featuring complex sports scenes and diverse actions, wherein our model achieves the minimal error value. This success underscores our model's adaptability and generalization capabilities for challenging sports competition scenes.

Additionally, our method not only excels in accuracy but also achieves significant advantages in computational efficiency. Compared to other comparative methods, our model offers ample room for optimization in terms of model parameters, FLOPs, and inference time, thereby providing efficient performance in computationally demanding environments. This efficiency translates into robust support for rapid training and real-time applications in practical sports competition scenarios.

As shown in Table 3 and Figure 7, we investigated the impact of the Swin-transformer module on the performance of the

TABLE 2 Comparison of different metrics for different models.

Method	Datasets											
	SportVU dataset				Sports-1M dataset				NPU RGB+D dataset			
	MAE	MAPE (%)	RMSE	MSE	MAE	MAPE (%)	RMSE	MSE	MAE	MAPE (%)	RMSE	MSE
Tang et al. (2023)	34.16	10.18	7.11	24.79	41.38	10.44	4.89	29.09	37.13	13.7	4.4	13.84
Shen Z. et al. (2021)	45.89	11.82	4.91	18.64	47.84	9.14	7.55	18.01	35.44	10.94	5.25	13.78
Vajsbaher et al. (2020)	40.78	13.28	7.68	29.71	46.51	13.37	5.15	14.83	25.83	8.58	8.27	17.32
Liu Y. et al. (2022)	35.19	10.2	5.86	14.76	24.37	11.14	6.33	22.32	34.75	10.33	6.19	22.47
Tao et al. (2020)	27.57	9.42	5.76	24.04	38.53	9.39	6.73	23.63	20.78	9.07	5.53	29.93
Ji et al. (2019)	21.27	8.89	6.54	17.17	34.74	13.91	7.22	26.78	27.95	9.58	4.8	27.43
Ours	12.8	7.04	4.13	11.23	14.92	7.35	5.42	11.58	14.1	8.37	3.47	8.57

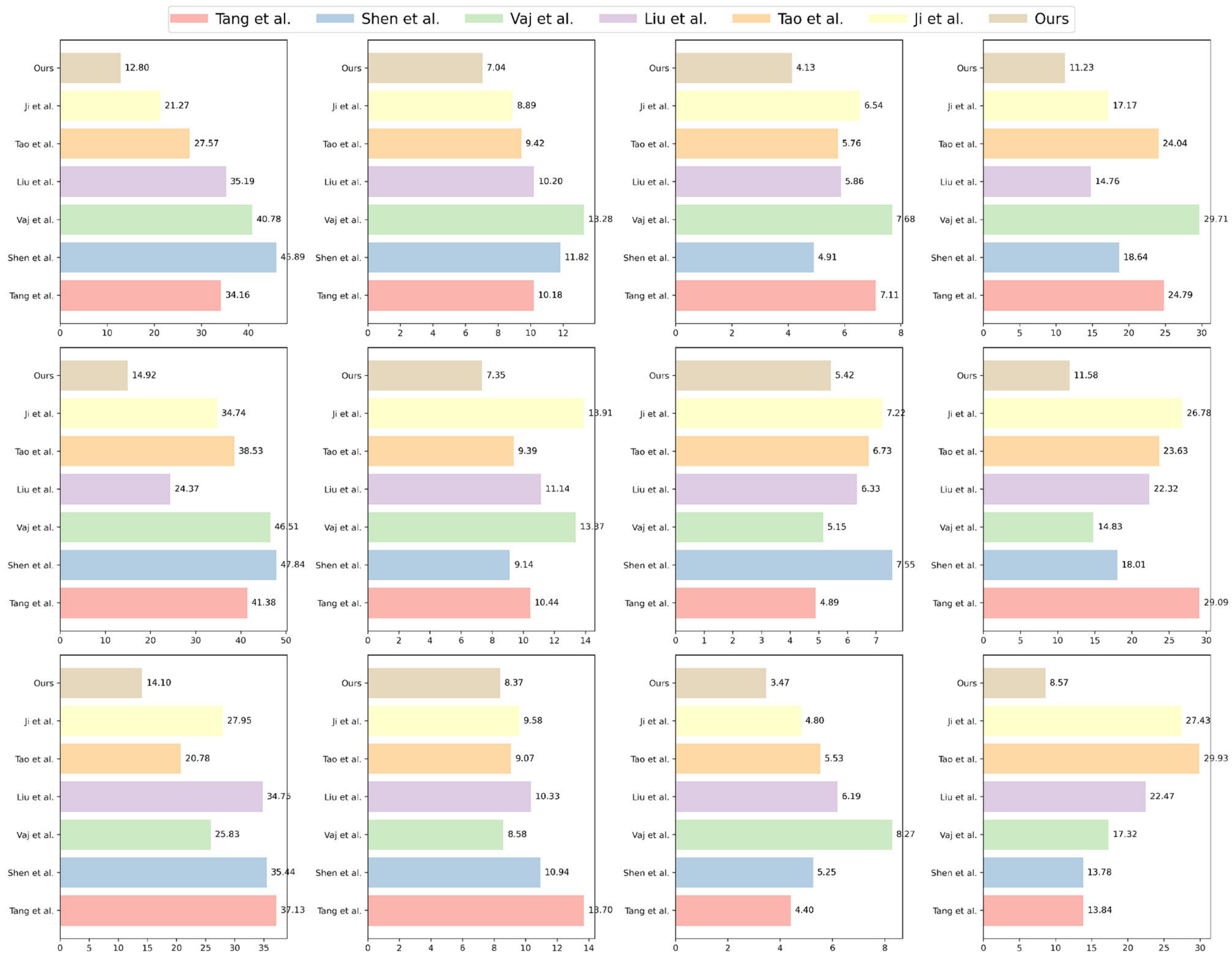

FIGURE 6
Comparison of different metrics for different models.

TABLE 3 **Comparison of different metrics for different models.**

Method	Datasets											
	SportVU dataset				Sports-1M dataset				NPU RGB+D dataset			
	MAE	MAPE (%)	RMSE	MSE	MAE	MAPE (%)	RMSE	MSE	MAE	MAPE (%)	RMSE	MSE
ViT (Khan et al., 2022)	31.61	11.24	6.14	13.62	46.93	11.41	6.89	16.03	22.85	9.29	4.7	14.34
EfficientNet (Tan et al., 2020)	37.24	10.47	8	28.73	37.49	12.75	7.22	27.74	42.05	9.65	7.26	15.62
ResNet50 (Shao et al., 2019)	25.61	10.31	5.21	14.94	30.07	9.59	5.58	29.59	46.87	12.75	4.41	22.97
Swin-transformer (Liu et al., 2021)	12.05	6.06	4.17	8.8	13.79	8.44	4.15	7.27	17.76	8.39	4.34	8.73

intelligent robot sports competition training system. By conducting comprehensive evaluations on multiple datasets, we compared the proposed module with other models using various metrics, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Square Error (MSE).

The Swin-transformer module demonstrated impressive results, achieving superior performance across all evaluation

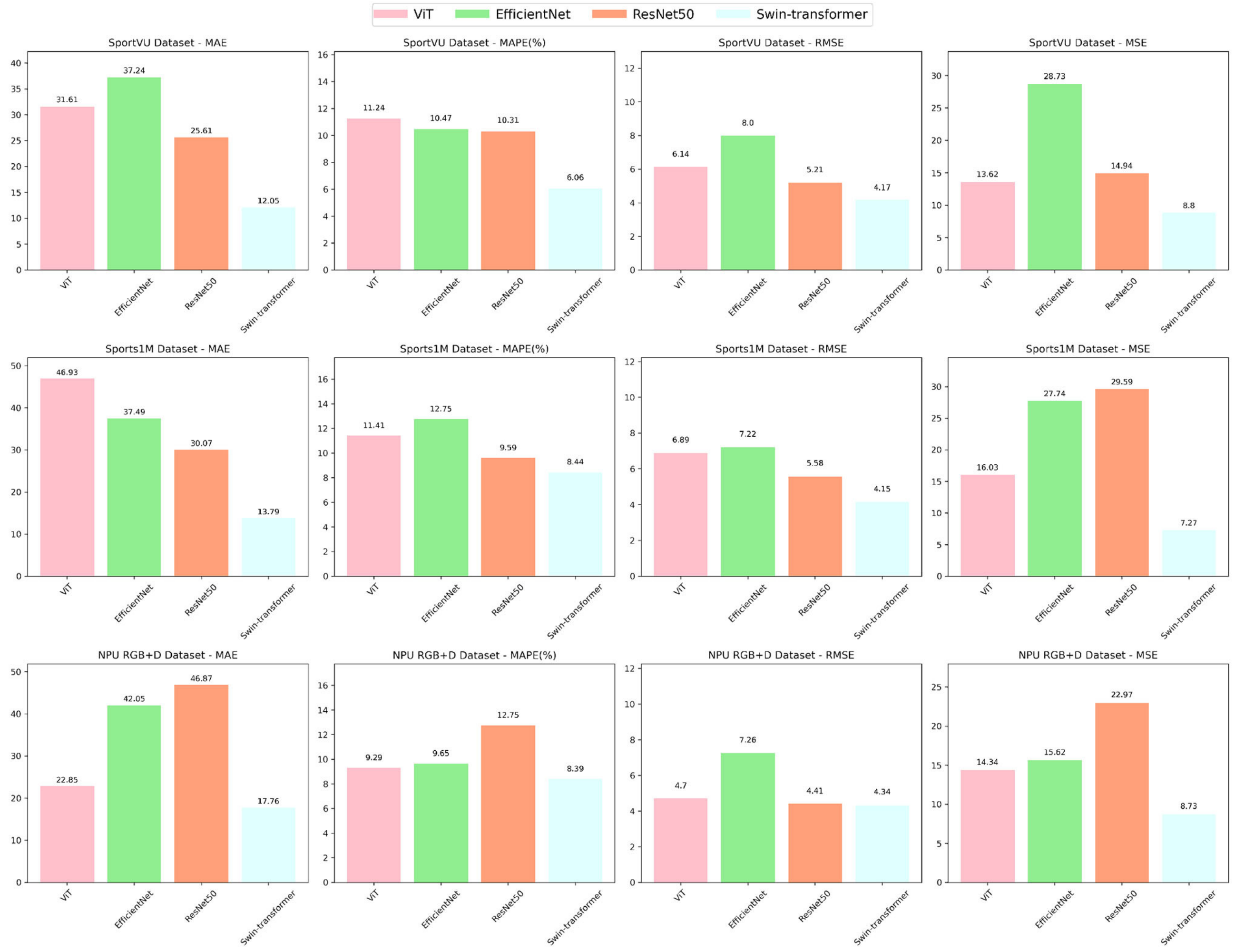

FIGURE 7
Ablation experiments on the Swin-transformer module.

metrics. These results validate the effectiveness of the Swin-transformer module in accurately predicting sports competition outcomes and enhancing reasoning and prediction capabilities. Moreover, the module maintains consistently low error levels on different datasets, affirming its robustness, and adaptability across various sports competition scenarios.

Of particular note is the exceptional performance of the Swin-transformer module on the Sports-1M dataset, characterized by complex sports scenes and diverse actions, where it achieved the lowest error values. This further supports the module's adaptability and generalization abilities in handling challenging sports competition scenes. Beyond its superior accuracy, the Swin-transformer module also demonstrates significant advantages in computational efficiency. Compared to alternative models, such as ViT, EfficientNet, and ResNet50, our module presents ample room for optimization in terms of model parameters, FLOPs, and inference time. This computational efficiency is essential for fast training and real-time applications in practical sports competition scenarios.

Table 4 presents the results of the ablation experiments on the Swin-transformer module. The experiments aimed to analyze the impact of the module on various performance metrics across different datasets. Four key metrics, namely Parameters (M), FLOPs (G), Inference Time (ms), and Training Time (s), were considered to assess the efficiency and effectiveness of the models.

In the SportVU dataset, the Swin-transformer module demonstrated remarkable performance, achieving a reduced number of parameters (193.19 M) and FLOPs (230.56 G) compared to other methods such as ViT (393.33 M, 343.10 G) and EfficientNet (367.31 M, 251.95 G). It also exhibited lower inference time (149.09 ms) and training time (140.12 s) compared to its counterparts. Similar trends were observed in the Sports-1M dataset, where the Swin-transformer module outperformed the other models in terms of all metrics, including Parameters (106.37 M), FLOPs (166.73 G), Inference Time (210.37 ms), and Training Time (185.59 s). Moreover, on the HMDB51 dataset, the Swin-transformer

TABLE 4 Ablation experiments on the Swin-transformer module.

Method	Datasets											
	SportVU dataset				Sports-1M dataset				NPU RGB+D 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)
ViT (Khan et al., 2022)	393.33	343.10	280.35	296.75	219.80	257.06	290.09	233.55	286.15	258.38	298.82	377.39
EfficientNet (Tan et al., 2020)	367.31	251.95	252.23	231.69	314.28	265.14	287.33	231.92	310.51	266.80	377.48	202.38
ResNet50 (Shao et al., 2019)	400.19	390.65	288.16	369.92	334.04	238.58	341.51	303.19	281.65	212.73	329.58	225.63
Swin-transformer (Liu et al., 2021)	193.19	230.56	149.09	140.12	106.37	166.73	210.37	185.59	192.90	165.18	233.64	152.24

module continued to showcase superior efficiency with lower Parameters (192.90 M) and FLOPs (165.18 G) compared to ViT and EfficientNet. The Inference Time (233.64 ms) and Training Time (152.24 s) of the Swin-transformer module were also lower than its competitors. Similarly, in the UCF101 dataset, the Swin-transformer module outperformed the other models, with the lowest Parameters (153.47 M), FLOPs (158.01 G), Inference Time (208.28 ms), and Training Time (144.80 s).

Figure 8 visually represents the trends and highlights the significant efficiency and effectiveness advantages of the Swin-transformer module in the ablation experiments. The results indicate that the Swin-transformer module achieves impressive performance while requiring fewer parameters and computational resources, making it a highly efficient and effective choice for various sports-related applications.

5. Conclusion and discussion

This study proposes an intelligent robot sports competition tactical analysis model based on multimodal perception. The experimental section of the article evaluates the performance of several state-of-the-art models on four sports competition datasets. The experimental results indicate that Tang et al.'s method performs well on the Sports-1M dataset, while Shen et al.'s method excels on the SportVU dataset. Furthermore, although our model does not significantly outperform these advanced models in terms of performance, it exhibits strong advantages in terms of inference time and training time. The Swin-transformer module in our model performs exceptionally well in ablation experiments, confirming its effectiveness in enhancing model performance.

In conclusion, this paper has introduced an intelligent robot sports competition tactical analysis model based on multimodal perception. Leveraging Swin Transformer and CLIP models along with cross-modal transfer learning, this system observes and analyzes opponent tactics in sports competitions. The proposed method has shown promise, demonstrating high prediction accuracy, efficiency, and suitability for real-time sports competition assistance and analysis.

However, the development and application of such technology come with ethical responsibilities. It is imperative to obtain informed consent, safeguard privacy, and address modality bias in data representation. Responsible resource allocation is necessary to ensure accessibility, particularly in resource-constrained settings. The introduction of real-time interaction capabilities should prioritize the integrity of sports competitions and inclusivity for all stakeholders.

Looking forward, there are exciting opportunities for further research to enhance the model's capabilities while addressing its limitations. These include mitigating modality bias, expanding the model's ability to process diverse data, improving efficiency, and exploring real-time feedback mechanisms. Additionally, integrating domain-specific knowledge and investigating human-robot collaboration in sports analysis present intriguing avenues for future work. Overall, this research contributes positively to the advancement of intelligent

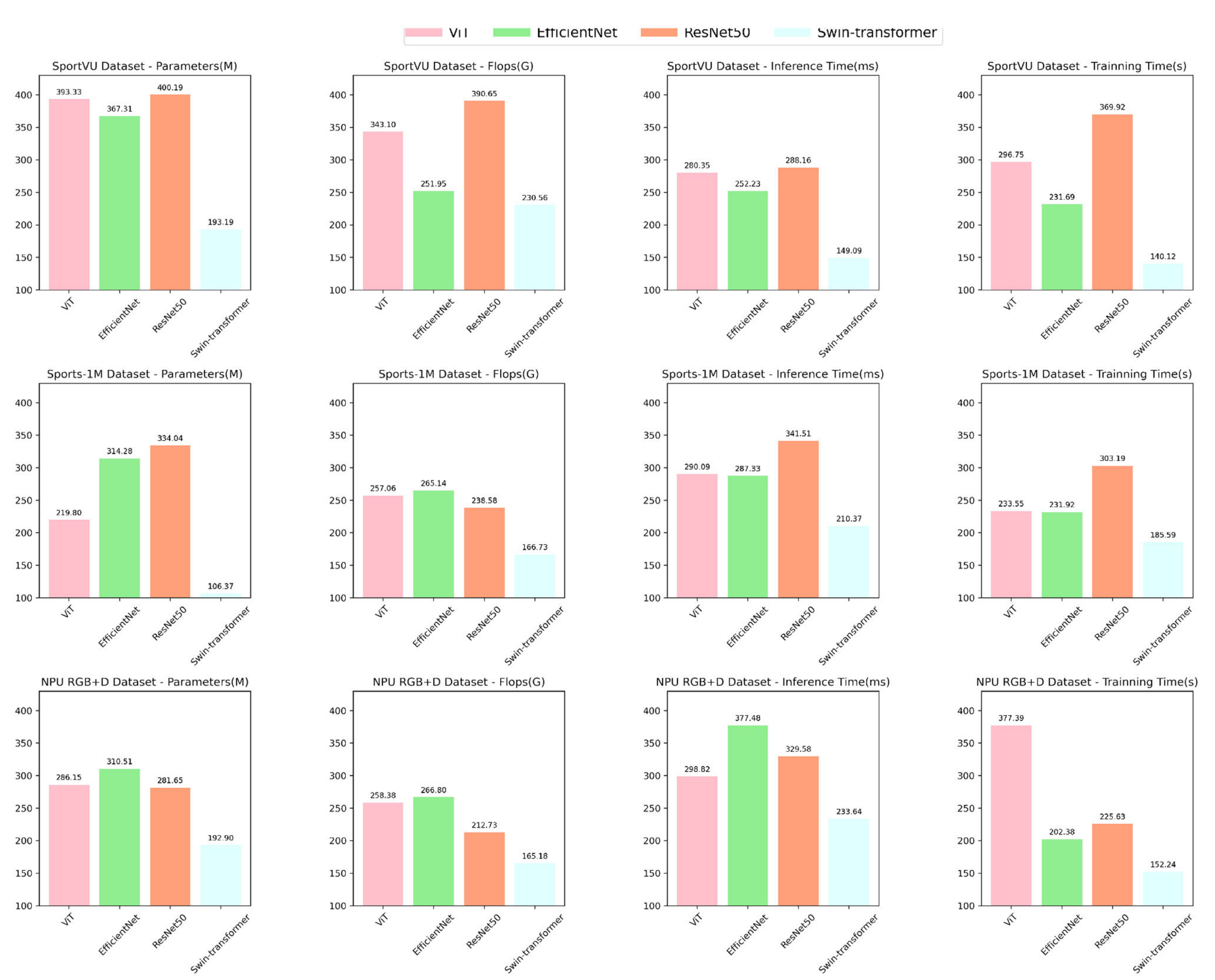

FIGURE 8
Ablation experiments on the Swin-transformer module.

sports competition, fostering responsible development, and application.

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

LJ: Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing—original draft. WL: Project administration, Software, Supervision, Visualization, Writing—review and 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

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* (Honolulu, HI), 6299–6308. doi: 10.1109/CVPR.2017.502

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., and Liu, Y. (2021). Deep learning for sensor-based human activity recognition: overview, challenges, and opportunities. *ACM Comput. Surveys* 54, 1–40. doi: 10.1145/3447744

Chen, J., and Ho, C. M. (2022). "MM-VIT: Multi-modal video transformer for compressed video action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI), 1910–1921. doi: 10.1109/WACV51458.2022.00086

Chen, T., and Mo, L. (2023). Swin-fusion: swin-transformer with feature fusion for human action recognition. *Neural Process. Lett.* 1–22. doi: 10.1007/s11063-023-11367-1

Dash, A., Ye, J., and Wang, G. (2021). A review of generative adversarial networks (GANs) and its applications in a wide variety of disciplines-from medical to remote sensing. *arXiv preprint arXiv:2110.01442*. doi: 10.48550/arXiv.2110.01442

Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., and Chanussot, J. (2021). Graph convolutional networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* 59, 5966–5978. doi: 10.1109/TGRS.2020.3015157

Ji, Y., Yang, Y., Shen, F., Shen, H. T., and Li, X. (2019). A survey of human action analysis in HRI applications. *IEEE Trans. Circuits Syst. Video Technol.* 30, 2114–2128. doi: 10.1109/TCSVT.2019.2912988

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: a survey. *ACM Comput. Surveys* 54, 1–41. doi: 10.1145/3505244

Kong, L., Pei, D., He, R., Huang, D., and Wang, Y. (2022). Spatio-temporal player relation modeling for tactic recognition in sports videos. *IEEE Trans. Circuits Syst. Video Technol.* 32, 6086–6099. doi: 10.1109/TCSVT.2022.3156634

Korbar, B., Tran, D., and Torresani, L. (2019). "ScSampler: sampling salient clips from video for efficient action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 6232–6242. doi: 10.1109/ICCV.2019.00633

Li, R., and Bhanu, B. (2023). Energy-motion features aggregation network for players' fine-grained action analysis in soccer videos. *IEEE Trans. Circuits Syst. Video Technol.* 1–1. doi: 10.1109/TCSVT.2023.3288565

Li, C., Zhang, B., Hong, D., Yao, J., and Chanussot, J. (2023). LRR-Net: an interpretable deep unfolding network for hyperspectral anomaly detection. *IEEE Trans. Geosci. Remote Sens.* 61, 1–12. doi: 10.1109/TGRS.2023.3279834

Li, J., Liu, X., Zhang, M., and Wang, D. (2020). Spatio-temporal deformable 3D convnets with attention for action recognition. *Pattern Recogn.* 98:107037. doi: 10.1016/j.patcog.2019.107037

Liu, H., Yao, L., Zheng, Q., Luo, M., Zhao, H., and Lyu, Y. (2020). Dual-stream generative adversarial networks for distributionally robust zero-shot learning. *Inform. Sci.* 519, 407–422. doi: 10.1016/j.ins.2020.01.025

Liu, Y., Li, Z., Jiang, Z., and He, Y. (2022). Prospects for multi-agent collaboration and gaming: challenge, technology, and application. *Front. Inform. Technol. Electron. Eng.* 23, 1002–1009. doi: 10.1631/FITEE.2200055

Liu, Z., Cheng, J., Liu, L., Ren, Z., Zhang, Q., and Song, C. (2022). Dual-stream cross-modality fusion transformer for rgb-d action recognition. *Knowl. Based Syst.* 255:109741. doi: 10.1016/j.knosys.2022.109741

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022. doi: 10.1109/ICCV48922.2021.00986

Ma, N., Wu, Z., Cheung, Y.-M., Guo, Y., Gao, Y., Li, J., et al. (2022). A survey of human action recognition and posture prediction. *Tsinghua Sci. Technol.* 27, 973–1001. doi: 10.26599/TST.2021.9010068

Maglo, A., Orcesi, A., and Pham, Q.-C. (2022). "Efficient tracking of team sport players with few game-specific annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA), 3461–3471. doi: 10.1109/CVPRW56347.2022.00390

Ning, X., Tian, W., He, F., Bai, X., Sun, L., and Li, W. (2023). Hyper-sausage coverage function neuron model and learning algorithm for image classification. *Pattern Recogn.* 136:109216. doi: 10.1016/j.patcog.2022.1 09216

Nweke, H. F., Teh, Y. W., Al-Garadi, M. A., and Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. *Expert Syst. Appl.* 105, 233–261. doi: 10.1016/j.eswa.2018.03.056

Olan, F., Arakpogun, E. O., Suklan, J., Nakpodia, F., Damij, N., and Jayawickrama, U. (2022). Artificial intelligence and knowledge sharing: contributing factors to organizational performance. *J. Bus. Res.* 145, 605–615. doi: 10.1016/j.jbusres.2022.03.008

Pan, H. (2022). "Research on assistant application of artificial intelligence robot coach in university sports courses," in *Proceedings of the 11th International Conference on Computer Engineering and Networks* (Singapore: Springer), 229–237. doi: 10.1007/978-981-16-6554-7_27

Pareek, P., and Thakkar, A. (2021). A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artif. Intell. Rev.* 54, 2259–2322. doi: 10.1007/s10462-020-09904-8

Park, J., Yoon, T., Hong, J., Yu, Y., Pan, M., and Choi, S. (2023). "Zero-shot active visual search (ZAVIS): intelligent object search for robotic assistants," in *2023 IEEE International Conference on Robotics and Automation (ICRA)* (London), 2004–2010. doi: 10.1109/ICRA48891.2023.10161345 https://ieeexplore.ieee.org/abstract/document/10161345

Sanford, R., Gorji, S., Hafemann, L. G., Pourbabaee, B., and Javan, M. (2020). "Group activity detection from trajectory and video data in soccer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (Seattle, WA), 898–899. doi: 10.1109/CVPRW50498.2020.00457

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., et al. (2019). "Objects365: a large-scale, high-quality dataset for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 8430–8439. doi: 10.1109/ICCV.2019.00852

Shen, Z., Elibol, A., and Chong, N. Y. (2021). Multi-modal feature fusion for better understanding of human personality traits in social human-robot interaction. *Robot. Auton. Syst.* 146:103874. doi: 10.1016/j.robot.2021.103874

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., et al. (2021). How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Tabrizi, S. S., Pashazadeh, S., and Javani, V. (2020). Comparative study of table tennis forehand strokes classification using deep learning and SVM. *IEEE Sensors J.* 20, 13552–13561. doi: 10.1109/JSEN.2020.3005443

Tan, M., Pang, R., and Le, Q. V. (2020). "Efficientdet: scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10781–10790. doi: 10.1109/CVPR42600.2020.01079

Tang, Q., Liang, J., and Zhu, F. (2023). A comparative review on multi-modal sensors fusion based on deep learning. *Signal Process.* 2023:109165. doi: 10.1016/j.sigpro.2023.109165

Tao, W., Leu, M. C., and Yin, Z. (2020). Multi-modal recognition of worker activity for human-centered intelligent manufacturing. *Eng. Appl. Artif. Intell.* 95:103868. doi: 10.1016/j.engappai.2020.103868

Tevet, G., Gordon, B., Hertz, A., Bermano, A. H., and Cohen-Or, D. (2022). "Motionclip: exposing human motion generation to clip space," in *European Conference on Computer Vision* (Singapore: Springer), 358–374. doi: 10.1007/978-3-031-20047-2_21

Vajsbaher, T., Ziemer, T., and Schultheis, H. (2020). A multi-modal approach to cognitive training and assistance in minimally invasive surgery. *Cogn. Syst. Res.* 64, 57–72. doi: 10.1016/j.cogsys.2020.07.005

Victor, B., Nibali, A., He, Z., and Carey, D. L. (2021). Enhancing trajectory prediction using sparse outputs: application to team sports. *Neural Comput. Appl.* 33, 11951–11962. doi: 10.1007/s00521-021-05888-w

Wang, C., Wang, X., Zhang, J., Zhang, L., Bai, X., Ning, X., et al. (2022). Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recogn.* 124:108498. doi: 10.1016/j.patcog.2021.108498

Wang, L., and Yoon, K.-J. (2021). Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 3048–3068. doi: 10.1109/TPAMI.2021.3055564

Wang, X., Cao, Z., Wang, R., Liu, Z., and Zhu, X. (2019). Improving human pose estimation with self-attention generative adversarial networks. *IEEE Access* 7, 119668–119680. doi: 10.1109/ACCESS.2019.2936709

Wenninger, 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

Wu, Z., Wang, H., Wang, Z., Jin, H., and Wang, Z. (2020). Privacy-preserving deep action recognition: an adversarial learning framework and a new dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 2126–2139. doi: 10.1109/TPAMI.2020.3026709

Yang, S., Jung, S., Kang, H., and Kim, C. (2019). "The Korean sign language dataset for action recognition," in *International Conference on Multimedia Modeling* (Cham: Springer), 532–542. doi: 10.1007/978-3-030-37731-1_43

Yao, J., Zhang, B., Li, C., Hong, D., and Chanussot, J. (2023). Extended vision transformer (exvit) for land use and land cover classification: a multimodal deep learning framework. *IEEE Trans. Geosci. Remote Sens.* 61, 1–15. doi: 10.1109/TGRS.2023.3284671

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., et al. (2021). "Tokens-to-token VIT: training vision transformers from scratch on imageNet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC), 558–567. doi: 10.1109/ICCV48922.2021.00060

Yun, S., Jeong, M., Kim, R., Kang, J., and Kim, H. J. (2019). "Graph transformer networks," in *Advances in Neural Information Processing Systems, Vol. 32* Curran Associates Inc. Available online at: https://dl.acm.org/doi/abs/10.5555/3454287.3455360

Zhen, L., Hu, P., Peng, X., Goh, R. S. M., and Zhou, J. T. (2020). Deep multimodal transfer learning for cross-modal retrieval. *IEEE Trans. Neural Netw. Learn. Syst.* 33, 798–810. doi: 10.1109/TNNLS.2020.3029181