Spatial-Temporal Pyramid Reasoning for Fine-grained Classification and Recognition in Sports Videos

Zichen Wang[1], Qianze Liu[1], and Yufei Zhang[1,2]

[1]Beijing University of Technology
[2]Corresponding Author : denny_0601@126.com

***Abstract*—The diverse and intricate world of sports, an integral part of human life, has ignited the emergence of a variety of technologies. These technologies decode the content of sports event videos, encompassing athletes' movements, postures, gaits, and performances. Within the realm of sports, these technologies prove invaluable for coaches, trainers, athletes, and others in analyzing the intricacies of top-tier sporting events. Simultaneously, they fulfill the desires of sports enthusiasts aiming to replicate professional skills like passing and shooting in soccer games. In the realm of technology, these advancements find applications due to the complexity and real-time nature of sports videos, ensuring their robustness across various contexts. However, it is within this context that the automated sports analysis community faces three primary challenges. First, the complex and diverse nature of movement patterns poses difficulties in effective action recognition and fine-grained classification, as local receptive fields struggle to encompass all scenarios. Second, the uneven distribution of data and the limited discriminative nature of temporal features complicate the differentiation of distinct actions. Third, current interpretations remain coarse-grained, falling short of meeting the demands for in-depth analysis. To tackle these issues, we propose a lightweight spatial-temporal pyramid reasoning approach. It aims to capture short-term, mid-term, and long-term temporal contexts, enhancing the learning and analysis of intricate sports video information. In this approach, the spatial-temporal pyramid expands the temporal receptive field of the network through multi-temporal kernel decompositions, yielding more discriminative features without significantly inflating computational costs. Additionally, we employ dense sampling to train video networks of varying lengths, catering to the requirements of fine-grained queries. Experimental results validate that our proposed method achieves significant enhancements across various video benchmarks, including soccer highlight video datasets.**

***Index Terms*—Fine-grained classification, Player action analysis, Video analysis, Sports analytics, Spatial-temporal network.**

I. Introduction

SPORTS activities, as an essential component of human societal life [1], fulfill not only the need for physical exercise and entertainment but also encompass the essence of teamwork, competitive spirit, and cultural exchange. With the continuous progress of society and rapid technological development, the significance of sports activities has become increasingly prominent. To better cater to the demands of coaches, athletes, and enthusiasts for event analysis, sports event video analysis has become a crucial part of sports [2]–[4], encompassing the analysis of athlete movements, postures, gaits, performances, team positioning, collective attacking trends, and more [5]. This analysis aids in dissecting team strategies for offense, defense, and technical enhancement.

The aforementioned sports activities within video analysis can be defined as a series of activities conducted by individuals or teams [6], [7], involving physical exertion and skillful competition against opponents. Broadly speaking, sports can be categorized into indoor leisure sports (such as badminton, snooker, chess, table tennis) and outdoor leisure sports (such as cricket, soccer, basketball). While many sports share similarities in terms of body posture, objective (defeating opponents), and gameplay, distinctions do exist among them. For instance, badminton emphasizes forearm strength, while tennis places more emphasis on arm strength. Furthermore, each sport has unique attributes such as game pace, requirements, dynamics, outcomes, and physical demands on athletes. It's these distinct qualities that make it challenging for researchers to capture and analyze the intricacies of movements [8]–[10], particularly in team sports like soccer and ice hockey, which involve numerous players, real-time actions, and intricate maneuvers.

Traditional sports event video analysis primarily relies on manual methods [11], [12]. In the case of soccer matches [13], professional analysis teams observe and analyze the movements, actions, overall progress trends, positioning, and final scores of players at various positions on the field. This macro-level analysis helps with controlling the flow of a regular match and making timely adjustments for subsequent matches. However, the complexity of movements, challenges posed by group strategies, and intricacies of actions make this approach inefficient. Strategies and action analysis within sports events present solid technological barriers, demanding significant human, financial, and material resources.

In response to the substantial demand in this field, automated sports video analysis has made rapid advancements. Traditional approaches often leverage sensors; researchers combine real-time inference with Internet of Things (IoT) wearable sensors to capture athlete movement, positioning, and other data [14]–[16]. For example, Connaghan et al. [17] utilized sensors to capture hand movements of tennis players, categorizing their actions as backhand, forehand, or serve, and classifying players as advanced, intermediate, or

beginners. Kiang et al. [18] used acoustic and acceleration sensors to analyze volleyball spikes and determine player speed and competitiveness. Steels et al. [19] employed three sensor positions to monitor changes in motion at the handle, wrist, and upper arm of badminton players, capturing seven shot types and two actions (running and standing). However, these sensor-based methods, while effective, rely on subsequent manual analysis and are less efficient. Additionally, the complexity of actions, diverse postures, rapidly changing gaits, and impressive performances all contribute to high precision requirements not always met by sensor-based methods [20].

In recent years, deep learning algorithms have shown promise in overcoming these limitations with their strong pattern recognition and data processing capabilities [21]–[23]. Rahmad et al. [24] used computer vision to classify badminton shots. Cai et al. [25] introduced HARPET, a convolutional neural network-based approach for recognizing four ice hockey actions: forward skating, backward skating, passing, and shooting. McNally et al. [26] introduced SwingNet, a lightweight deep neural network combining convolutional and recurrent architectures to analyze golf swing videos from championship tours, identifying eight primary golf swing actions. While these methods use repeated local operations like convolution and recurrence to capture long-range dependencies in space and time, they are often limited in capturing comprehensive action representations due to their local scope and lack of multi-scale consideration.

However, challenges persist in sports video analysis. First, complex and varied movement patterns, along with significant occlusion, make recognizing actions in sports events, especially team sports like soccer, difficult. Second, due to uneven data distributions, the discriminative ability of temporal features can be restricted. Lastly, current interpretation methods suffer from coarse granularity, unable to capture intricate motion details.

To address these challenges, this paper proposes an innovative solution—a lightweight Adaptive Spatial-Temporal Pyramid Reasoning (STPR) method. This approach aims to dynamically capture short-term, mid-term, and long-term temporal contexts and integrate them with spatial features to better learn and analyze information in complex sports videos. By building a pyramid structure in both space and time, different scales of action features can be effectively captured, enhancing the accuracy of action recognition and classification. The lightweight design of this approach ensures practical efficiency, holding the potential to bring breakthroughs to the field of sports video analysis.

Our research focuses on macro-level analysis of complex actions and group strategies in sports event videos, particularly using soccer matches as examples due to their intricate positioning involving numerous players. We have successfully employed deep learning techniques to address three major challenges in sports event analysis, also :

- Addressing complex and varied action patterns and occlusion: We introduce a novel spatial-temporal pyramid framework, which attempts to infer spatial-temporal relationships using raw videos as input. By aggregating action-sensitive features from multi-scale images, a pyramid graph structure is constructed, resulting in a more robust and comprehensive action representation. This approach performs well in handling complex actions and occlusion.
- Overcoming data distribution imbalance and low temporal feature discriminability: We integrate an attention mechanism within the spatial-temporal pyramid framework to determine the importance of each frame and spatial-temporal region at the feature level. This mechanism allows us to focus more precisely on critical temporal and spatial information, beyond just channel features. This aids in improving the accuracy of action recognition and analysis while overcoming challenges related to data distribution imbalance and low temporal feature discriminability.
- Breaking the limitations of coarse-grained group strategy analysis: We introduce an Energy-Motion Feature Aggregation module, utilizing both the energy representation of soccer players in video sequences and explicit motion dynamics. This module enhances fine-grained analysis of soccer player movements, resulting in a better understanding of player behavior and strategies. Through this innovative approach, we successfully surpass the limitations of coarse-grained group strategy analysis.

In terms of experimentation, we conducted extensive tests that demonstrated the effectiveness and performance of our proposed method. This included experiments on various sports video datasets, comparisons with existing methods, and performance evaluations in fine-grained classification and strategy analysis tasks. These experimental results further confirm the practical application potential and superiority of our method.

The remaining sections of this paper are organized as follows. Section II provides a concise overview of the relevant literature in the field of study, including video action recognition, fine-grained classification, and spatial-temporal analysis. Section III elaborates on the Spatial-Temporal Pyramid Reasoning Framework (STPR-Net) proposed in this paper. Section IV presents comprehensive ablation studies, experimental results, and visualizations. Finally, a summary of this paper is provided in Section V.

II. Related work

In this section, we introduce related works in terms of four aspects: 1) video action recognition; 2) fine-grained classification; and 3) spatial-temporal analysis.

A. Video Action Recognition

Video action recognition is an essential aspect of computer vision, and it has garnered widespread attention due to its applications in various fields such as sports analysis, surveillance, and human-computer interaction [27], [28]. Early research in this field primarily revolved around handcrafted features and simple classifiers [29], [30]. Methods like Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) were employed to extract spatiotemporal information from videos [31], [32]. While these methods laid the foundation,

they still struggled to handle complex actions as well as variations in appearance and lighting.

With the emergence of deep learning, there has been a paradigm shift in video action recognition. Convolutional Neural Networks (CNNs) have been adapted to the temporal dimension, leading to the development of two-stream CNNs [33], [34]. These networks combine spatial and temporal streams to capture motion patterns and appearance information. The fusion of these streams enhances the accuracy of recognizing various actions. By extending traditional 2D convolutions to the temporal dimension, 3D CNNs were introduced to directly capture spatiotemporal features from raw video frames. However, a major drawback of 3D CNNs is their involvement of a large number of parameters [35], [36], resulting in increased computational demands and the need for extensive pretraining datasets.

To address this limitation, the use of dilated 2D convolution kernels as 3D convolution kernels was proposed [37], [38]. Nevertheless, these algorithms can only handle single frames or short-term video data, making it challenging to achieve long-term temporal modeling. Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), emerged to tackle this predicament [22], [39], [40]. These architectures can identify actions that unfold over longer time intervals. Furthermore, recent RNN research has explored the integration of graph-based representations to capture interactions among multiple actors within a scene. Despite these methods achieving recognition of actions in video data, challenges persist in recognizing complex group actions and handling scenarios with severe noise like occlusions.

B. Fine-grained Classification

Fine-grained classification refers to the more detailed and accurate categorization of objects or entities with similar appearances in the fields of machine learning and computer vision. Unlike traditional coarse-grained classification (such as categorizing cats, dogs, birds, etc., into distinct classes), fine-grained classification aims to identify different subcategories within the same class, where the differences between these subcategories are often subtle, requiring higher sensitivity and feature distinctiveness. Yang et al. [41] introduced a novel self-supervised mechanism known as NTS-Net (Navigator-Teacher-Scrutinizer Network), composed of Navigator, Teacher, and Scrutinizer agents. This mechanism effectively localizes informative regions without the need for fine-grained bounding box/partial annotations. Zhuang et al. [42] proposed a simple yet effective Attention Pairing Interaction Network (API-Net) that progressively identifies pairs of fine-grained images through interactions.

However, achieving fine-grained classification in the sports domain is challenging. People typically rely more on coarse-grained classification [43] to categorize different sports actions, sports equipment, game types, etc. For instance, accurate recognition of different sports actions has been achieved by combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [41], [43]. Alternatively, transfer learning has been employed for equipment classification in basketball games. Nevertheless, these methods achieve category distinctions, but they struggle to achieve similar effects for actions with subtle differences or combinations that rely on coherence.

C. Spatial-temporal network

A spatial-temporal network is an advanced data analysis method that involves spatial and temporal information. Its primary objective is to utilize the changing trends, patterns, and correlations of data in both spatial and temporal dimensions to achieve goals such as detection and recognition [44]. This analytical approach finds widespread applications in various fields [45]–[47], including environmental science, sociology, epidemiology, urban planning, and more. In the realm of sports, particularly within the context of the domain of interest in this study, spatial-temporal analysis assists researchers and decision-makers in comprehensively understanding the changing trends of athletes' performances, competitions, and training. This understanding further provides robust support for devising more efficient training plans, competition strategies, and decision-making.

In practical applications, Aljoufie et al. [48] have proposed a methodology based on eight city growth and traffic indices to analyze the relationship between urban spatial growth and traffic changes. They employed techniques such as remote sensing (RS) and geographic information systems (GIS) to quantify and analyze the spatiotemporal relationship between the development of Jeddah city and its traffic dynamics. Furthermore, Yao et al. [49] have introduced a framework called Deep Multi-View Spatio-Temporal Network (DMVST-Net) to address spatial and temporal relationship modeling in traffic prediction problems. As an integrated method considering both spatial and temporal information, the spatial-temporal network demonstrates substantial potential across various domains. In the realm of sports, it provides valuable insights and support for optimizing athlete performance, competition trends, and training plans.

III. The Proposed Approach

In this section, we propose an innovative solution, a lightweight Adaptive Spatial-Temporal Pyramid Reasoning (STPR) method, to address the various challenges mentioned earlier in complex sports activities. For the first time, this framework attempts to utilize raw, unprocessed videos as input for spatial-temporal relationship inference regarding strategic analysis and action classification. Firstly, by integrating action-sensitive features from multi-scale images, a pyramid graph structure is constructed to attain a more robust and comprehensive representation of actions. Next, within the spatial-temporal pyramid framework, an attention mechanism is introduced to determine the importance of each frame and each spatiotemporal region at the feature level. This attention mechanism enables us to focus more on temporal and spatiotemporal cues, beyond just channel features. This way, we can capture crucial information in videos, enhancing the

Fig. 1. An ideal video sample of a sports match in original Bundesliga matches. In spite of employing high-speed cameras for recording the diverse motions of the participants, the precise extraction of individual players' actions and the subsequent analysis of strategic alterations during a standard match pose significant difficulties owing to the numerous player positions and the extensive field coverage.

precision of action recognition and analysis. During the feature extraction process, we introduce a multi-temporal kernel decomposition approach to effectively expand the network's temporal receptive field. Through this method, we are able to broaden the network's perception of temporal sequence data without significantly increasing computational costs. This further reinforces the model's ability to comprehend temporal information. Lastly, we present an energy-motion feature aggregation module to fully utilize energy representations and distinct motion dynamics of soccer players in video sequences. This module aids in fine-grained analysis of soccer player actions, leading to a better understanding of player behavior and strategies. This novel approach introduces fresh perspectives to the field of sports video analysis. It represents the first attempt at analyzing and finely classifying complex actions and group strategies in sports event videos. Figure 2 shows the structure of our proposed framework.

In this section, we first introduce the problem addressed in this study, as explained in Section III-A. Following that, we present the proposed STPR framework and its utilization of a pyramid structure in Section III-B. In particular, the adaptive mechanism employed to capture hort-term, mid-term, and long-term spatio-temporal relationships in multi-scale video sequences will be detailed in Section III-C. Section III-D will elaborate on the multi-temporal kernel decomposition method employed in the feature extraction process of the STPR framework. Lastly, we will provide an introduction and explanation of the energy-motion feature aggregation module in Section III-E.

A. Problem Definition

Fine-grained action detection is a task that aims to locate and recognize the specific actions of individual players or groups of players in a video sequence, and to output the temporal boundaries and the action labels of each detected action. For example, in a football match, the actions of players may include passing, shooting, dribbling, tackling, etc., and the action detection task is to output the start and end time of each action, as well as the corresponding action label. Formally, given a video sequence $V = v_1, v_2, ..., v_T$, where v_t is the t-th frame image, and a set of action categories $C = c_1, c_2, ..., c_K$, where c_k is the k-th action category, the fine-grained action detection task is to output a set of action instances $A = a_1, a_2, ..., a_N$, where $a_n = (s_n, e_n, l_n)$ is the n-th action instance, consisting of the start time s_n, the end time e_n, and the action label $l_n \in C$. The output set A should satisfy the following constraints:

- For each action instance $a_n = (s_n, e_n, l_n)$, we have $1 \leq s_n < e_n \leq T$ and $l_n \in C$.
- For any two action instances $a_i = (s_i, e_i, l_i)$ and $a_j = (s_j, e_j, l_j)$, if $i \neq j$, then we have either $e_i < s_j$ or $e_j < s_i$, that is, no overlap between different action instances.

However, this task faces many difficulties. One of them is that recognizing actions in sports events, especially team sports like soccer, is hard because of the complex and varied movement patterns and the significant occlusion. Another one is that the temporal features may not be discriminative enough because of the uneven data distributions. The last one is that the current interpretation methods are too coarse-grained and cannot capture the intricate motion details. Figure 1 illustrates an ideal video sample of a sports match. Despite the utilization of high-speed cameras to capture the various movements of the participants, accurately extracting the actions of individual players and subsequently analyzing the strategic changes during a regular match proves to be quite challenging due to the multitude of player positions and the expansive field area. Furthermore, although the aforementioned action examples belong to the same category of movements, they significantly differ in terms of speed, duration, as well as starting and ending positions.

B. The Overview of STPR

We propose a lightweight Adaptive Spatial-Temporal Pyramid Reasoning (STPR) method to address the various chal-

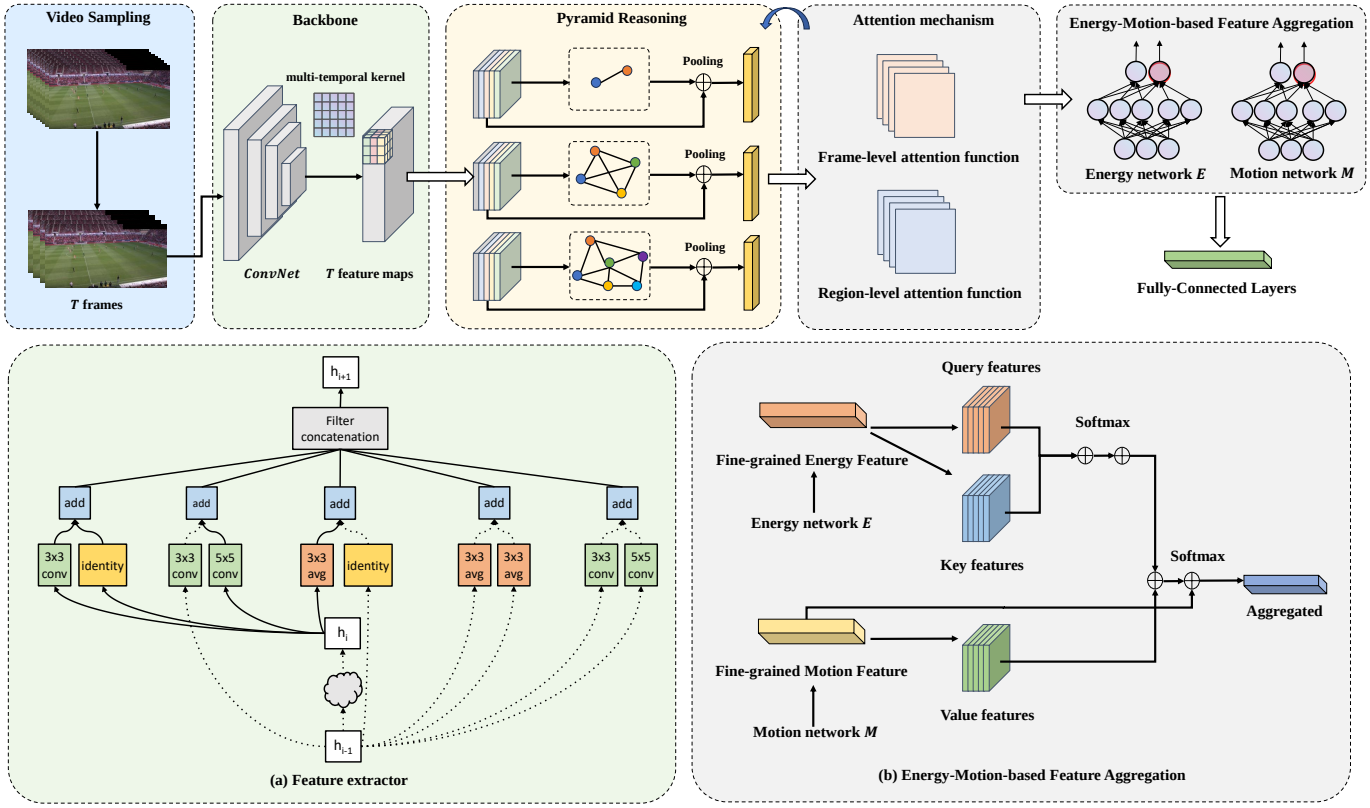

Fig. 2. The structure of our proposed framework STPR.

lenges in complex sports video analysis. This method attempts to use raw videos as input for the first time, and perform spatial-temporal relationship inference for action recognition and strategy analysis. Figure 2 shows the structure of our proposed framework.

Specifically, we first input the raw video sequence $V = \{v_1, v_2, ..., v_T\}$, where v_t is the t-th frame image. Then, we process each frame image with multi-scale image processing, and extract features at different levels. Specifically, for each scale s, we use a convolutional layer C_s to process each frame image, and obtain the features at that scale $F_s = \{f_{s,1}, f_{s,2}, ..., f_{s,T}\}$, where $f_{s,t} = C_s(v_t)$ is the feature of the t-th frame at the s-th scale. We build a pyramid graph structure with multi-scale features, where each node represents a spatio-temporal region, and each edge represents the relationship between two regions. Specifically, for each scale s's feature F_s, we divide it into N_s spatio-temporal regions $R_s = \{r_{s,1}, r_{s,2}, ..., r_{s,N_s}\}$, where $r_{s,i}$ is the feature vector of the i-th region. We use an undirected graph $G_s = (V_s, E_s)$ to represent these regions and their relationships, where the node set $V_s = R_s$, and the edge set $E_s = \{(r_{s,i}, r_{s,j}) | i, j \in [1, N_s]\}$. We stack the graph structures at different scales to form a pyramid graph structure $P = (G_1, G_2, ..., G_S)$.

Next, we introduce an attention mechanism in the pyramid graph structure to calculate the importance weights of each frame and each spatio-temporal region. Specifically, for each scale s's graph structure $G_s = (V_s, E_s)$, we define a frame-level attention function $A_f : V_s \rightarrow [0, 1]$ and a region-level attention function $A_r : E_s \rightarrow [0, 1]$. The frame-level attention function calculates the importance weight of each frame, and the region-level attention function calculates the importance weight of each spatio-temporal region. We use the attention weights to perform weighted averaging on the features, and obtain the spatial-temporal pyramid features. For each scale s's graph structure $G_s = (V_s, E_s)$, we use the attention weights to perform weighted averaging on its nodes and edges, and obtain the spatial-temporal feature at that scale h_s. The formula is as follows:

$$h_s = \frac{1}{N_s} \sum_{i=1}^{N_s} A_f(r_{s,i}) \sum_{j=1}^{N_s} A_r(r_{s,i}, r_{s,j}) r_{s,j} \quad (1)$$

We concatenate the spatial-temporal features at different scales to obtain the spatial-temporal pyramid feature $h = [h_1; h_2; ...; h_S]$. After that, in order to expand the network's temporal receptive field, we use a multi-temporal kernel decomposition method in the feature extraction process to decompose the convolutional kernel into sub-kernels with different lengths. We use these sub-kernels to convolve the input feature, and concatenate the output features to obtain a larger temporal receptive field.

After that, we use the energy-motion feature aggregation module to fuse the energy representation and motion dynamics in the video sequence, and obtain the fine-grained feature. In particular, with regards to the given input video sequence denoted as $V = v_1, v_2, ..., v_T$, an energy network denoted as E and a motion network denoted as M are employed

for extracting the energy-based representation and motion dynamics correspondingly. The energy network utilizes a two-dimensional convolutional layer for the processing of each individual frame image, yielding the energy feature $E(V)$. The motion network employs a three-dimensional convolutional layer to process the entire video sequence, resulting in the motion feature $M(V)$. The fusion of the energy feature and motion feature is performed to derive the fine-grained feature denoted as g.

Finally, we use the spatial-temporal pyramid feature h and the fine-grained feature g to perform action recognition and strategy analysis, and output the classification results and evaluation metrics. Specifically, for each video sequence V_i, we use a classification network P to perform action recognition on the spatial-temporal pyramid feature, and obtain the classification probability distribution $p_i = P(h_i)$. We use an evaluation network Q to perform strategy analysis on the fine-grained feature, and obtain the evaluation metric $q_i = Q(g_i)$. The loss function is defined as the weighted sum of the action recognition loss and the strategy analysis loss. Specifically, for each video sequence V_i's true label y_i and true evaluation value z_i, we use cross-entropy loss function to measure the difference between the classification result and the true label, and use mean squared error loss function to measure the difference between the evaluation metric and the true value. The formulas are as follows:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i \tag{2}$$

$$L_a = \frac{1}{N}\sum_{i=1}^{N} (q_i - z_i)^2 \tag{3}$$

where N is the number of video sequences.

C. Attention mechanism

Attention mechanism is a method to improve the model's attention and selectivity to the input data. In our method, we use two types of attention mechanisms: frame-level attention and region-level attention. Frame-level attention is used to calculate the importance of each frame image in the temporal domain, and region-level attention is used to calculate the importance of each spatio-temporal region in the spatial domain. Both types of attention mechanisms are based on feature vector self-attention, that is, using the similarity between feature vectors to calculate weights.

Specifically, for each scale s's graph structure $G_s = (V_s, E_s)$, we first calculate the dot product of each node (i.e., each spatio-temporal region)'s feature vector $r_{s,i}$ and a learnable weight vector w_f, and then use an exponential function to perform a nonlinear transformation, obtaining a non-negative scalar value. This value represents the frame-level importance of that node at that scale. We normalize the scalar values of all nodes so that their sum is 1, and obtain the frame-level attention weight $A_f(r_{s,i})$. The formula is as follows:

$$A_f(r_{s,i}) = \frac{\exp(w_f^\top r_{s,i})}{\sum_{j=1}^{N_s} \exp(w_f^\top r_{s,j})} \tag{4}$$

where w_f is a learnable weight vector.

Next, for each edge (i.e., the relationship between every two spatio-temporal regions), we calculate the Hadamard product of the feature vectors of the two nodes $r_{s,i}$ and $r_{s,j}$, i.e., element-wise multiplication, obtaining a new feature vector. This feature vector represents the interaction between the two nodes. We then use another learnable weight vector w_r to perform a nonlinear transformation on the dot product of this feature vector, obtaining a non-negative scalar value. This value represents the region-level importance of that edge at that scale. We normalize the scalar values of all edges so that their sum is 1, and obtain the region-level attention weight $A_r(r_{s,i}, r_{s,j})$:

$$A_r(r_{s,i}, r_{s,j}) = \frac{\exp(w_r^\top (r_{s,i} \odot r_{s,j}))}{\sum_{k=1}^{N_s} \exp(w_r^\top (r_{s,i} \odot r_{s,k}))} \tag{5}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication), and w_r is a learnable weight vector.

Finally, we use the frame-level attention weight and the region-level attention weight to perform weighted averaging on each scale s's graph structure $G_s = (V_s, E_s)$'s nodes and edges, obtaining the spatial-temporal feature at that scale h_s. The formula is as follows:

$$h_s = \frac{1}{N_s}\sum_{i=1}^{N_s} A_f(r_{s,i}) \sum_{j=1}^{N_s} A_r(r_{s,i}, r_{s,j}) r_{s,j} \tag{6}$$

In this way, we use the attention mechanism to determine the importance of each frame and each spatio-temporal region at the feature level, and fuse the information from different scales and different regions, obtaining richer and more meaningful spatial-temporal features.

D. Multi-temporal Kernel Decomposition

Multi-temporal kernel decomposition is a method to improve the network's understanding of temporal data, and it is an important component of our proposed Spatial-Temporal Pyramid Reasoning (STPR) method.

The basic idea of multi-temporal kernel decomposition is to decompose a convolutional kernel into sub-kernels with different lengths, in order to expand the network's temporal receptive field. The temporal receptive field refers to the range of temporal information in the input feature that the network can capture. Expanding the temporal receptive field helps the network learn longer temporal action features, thereby improving the effect of action recognition and strategy analysis.

Specifically, for a convolutional kernel $K \in \mathbb{R}^{C\times T\times H\times W}$, where C is the channel number, T is the temporal length, H and W are the spatial sizes, we decompose it into M sub-kernels $K_m \in \mathbb{R}^{C\times T_m\times H\times W}$, where $T_m < T$ is the temporal length of the m-th sub-kernel. We use these sub-kernels to convolve the input feature, and concatenate the output features to obtain a larger temporal receptive field. The formula is as follows:

$$K * F = [K_1 * F; K_2 * F; ...; K_M * F] \tag{7}$$

where $*$ denotes the convolution operation, and $F \in \mathbb{R}^{C\times T\times H\times W}$ is the input feature.

Our method uses different-length sub-kernels to capture different-scale temporal information, thereby enhancing the network's diversity and flexibility for temporal data. At the same time, our method reduces the computational cost and memory consumption by reducing the number of parameters of the convolutional kernel, thereby improving the network's efficiency and speed. In addition, our method increases the network's output channel number by concatenating different sub-kernel output features, thereby improving the network's expressive ability and classification ability.

To summarize, our approach has three unique advantages : 1) it uses different-length sub-kernels to capture different-scale temporal information; 2) it reduces the computational cost and memory consumption by reducing the number of parameters of the convolutional kernel; 3) it increases the network's output channel number by concatenating different sub-kernel output features. These differences make our method more suitable for complex sports video analysis.

E. Energy-Motion-based Feature Aggregation

Energy-motion feature aggregation is a method to improve the network's understanding of the soccer player's motion in the video sequence, and it is an important component of our proposed Spatial-Temporal Pyramid Reasoning (STPR) method. The basic idea of this method is to fuse the energy representation and motion dynamics of the soccer player's motion in the video sequence, and obtain the fine-grained feature. The fine-grained feature can reflect the behavior and strategy of the soccer player in the video sequence, thereby improving the network's effect of fine-grained action analysis for the soccer player.

Specifically, for the input video sequence $V = \{v_1, v_2, ..., v_T\}$, we use an energy network E and a motion network M to extract its energy representation and motion dynamics respectively. Energy representation reflects the energy consumed or released by the soccer player in each frame image, which can measure the spatial activity and variation of the soccer player. The energy network uses a two-dimensional convolutional layer to process each frame image, and obtains the energy feature $E(V) = \{e_1, e_2, ..., e_T\}$, where $e_t = E(v_t)$ is the energy feature of the t-th frame. Motion dynamics reflects the force or acceleration generated or received by the soccer player in the entire video sequence, which can measure the temporal coherence and change rate of the soccer player. The motion network uses a three-dimensional convolutional layer to process the entire video sequence, and obtains the motion feature $M(V) = m$. These two types of features can complement each other, forming a more comprehensive and meaningful video representation. We fuse the energy feature and the motion feature to obtain the fine-grained feature g. The formula is as follows:

$$g = \frac{1}{T}\sum_{t=1}^{T}(e_t \odot m) \tag{8}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication). We believe that there is an intrinsic relationship between energy representation and motion dynamics, that is, the soccer player's motion dynamics are driven by their energy representation, and their energy representation are influenced by their motion dynamics. Therefore, we use Hadamard product to implement the fusion of two types of features, that is, element-wise multiplication, obtaining a new feature vector. This feature vector can reflect the behavior and strategy of the soccer player in the video sequence, thereby improving the network's effect of fine-grained action analysis for the soccer player.

IV. Experiment

In this section, we evaluate the performance of the proposed STPR on four large-scale and widely used action recognition datasets: Something-Something V1 [50], Something-Something V2 [51], FineGym [52], and the DFL dataset from the Kaggle competition [53]. Firstly, we introduce the datasets and baselines is Section IV-A and Section IV-B. Subsequently, we visualize the effectiveness of our model on these datasets, demonstrate the robustness and generality of our approach across different backbone models in Section IV-C, and compare it with state-of-the-art methods in Section IV-D. Moving forward, we conduct an ablation study on the network design to investigate the influence of different components and factors in Section IV-E. Finally, we present learning curve visualizations to assess model convergence and compare performance across different metrics in Section IV-F. It's important to note that, unless otherwise specified, all experiments are conducted using a single modality (i.e., RGB frames) and evaluated on the validation set.

A. Datasets

Following [54], the four datasets we employ are described below.

Something-Something V1 [50]. This dataset comprises an extensive array of video clips depicting various activities associated with typical human existence and engagement with ordinary items. The emphasis lies on capturing motion rather than identifying associated objects. The dataset encompasses 86,000 training videos and 11,000 validation videos, spanning across 174 distinct action classifications. The duration of each video in the dataset varies between 2 to 6 seconds.

Something-Something V2 [51]. The dataset contains twice the number of videos compared to Something-Something V1. Specifically, it encompasses 220,000 video clips, comprising approximately 169,000 training videos and 25,000 validation videos. These clips were gathered by contributors who followed predefined fundamental actions involving objects. Each video has a duration spanning from 2 to 6 seconds. The primary focus of the dataset is on human activities involving commonplace objects, with the surrounding backgrounds playing a minor role in predicting the final action category. Furthermore, certain action classifications, such as "Pushing something from left to right" and "Pushing something from right to left," necessitate not only spatial semantic understanding but also robust spatial-temporal motion analysis.

FineGym [52]. Constructed from gymnastic videos, this dataset offers a nuanced perspective on actions, operating

at both the overall action level and its finer sub-action components. For our evaluation, we utilize the Gym99 sub-action collection, encompassing 99 distinct action classes and comprising approximately 20,000 training videos and 8,500 validation videos. The videos within this sub-action assortment generally have durations of under 2 seconds, prioritizing the intricate spatial-temporal motion details.

the DFL dataset [53]. The competition dataset includes video recordings of nine football matches, each divided into halves. The task involves identifying three types of player events in these videos, including their timing and nature. The training set is composed of 12 videos for both training and testing purposes, each of which has 60min. It encompasses video footage from eight games. Among these, four games have both halves included, while the other four have only one half incorporated. Additionally, the test data for the public leaderboard comprises complete video recordings from one game and segments from four other games, with the remaining halves of these games included in the training set. The test data also features brief clips from ten extra games. These clips lack event annotations but aid the model in generalizing to environments not covered by the training data.

B. Baselines

In the comparison experiment, we select a total of 8 baselines, all of which are classic and great algorithms for fine-grained action recognition in the sports field that we focus on in this study.

R(2+1)D CNN [55] is a novel spatiotemporal convolutional block named "R (2+1)D." It decomposes 3D convolutional filters into distinct spatial and temporal components, resulting in a significant enhancement in accuracy. Their motivation stems from the following observation: 2D CNNs applied to individual frames of videos continue to perform well in action recognition, while within a residual learning framework, 3D CNNs have demonstrated their superiority in accuracy over 2D CNNs.

ARTNet [56] is a deep learning model designed for video analysis, aiming to enhance the representation of spatiotemporal features through the incorporation of attention mechanisms. This model employs two parallel network branches: one dedicated to extracting spatial features from the video, and the other focused on capturing temporal features. The outputs of these two branches are then combined through attention mechanisms, allowing the model to emphasize the most relevant and informative aspects. This approach enables ARTNet to better capture object motion and contextual information in videos, thereby enhancing its performance in tasks such as action recognition and behavior analysis.

P3D-CTN [57] is a deep learning model designed for video action recognition. It builds upon existing 2D Convolutional Neural Networks (CNNs) with the goal of effectively utilizing temporal information to capture motion patterns within videos. P3D-CTN employs pseudo-3D convolutional operations, introducing convolutional operations along the temporal dimension. This enables the model to capture temporal changes in videos more effectively while maintaining a smaller parameter count. As a result, P3D-CTN achieves strong performance in video action recognition tasks while reducing computational overhead.

I3D-LSTM [58] is a deep learning architecture used for video action recognition. It evolved from 2D Convolutional Neural Networks (CNNs) by extending pre-trained 2D convolutional layers into three dimensions, allowing it to directly process video data. I3D introduces convolutional operations along the temporal dimension, enabling effective capture of spatiotemporal features within videos. Due to its fusion of 2D and 3D convolutions, I3D excels in action recognition tasks, as it can leverage both appearance and motion information. This architectural advantage has led to significant success for I3D in the field of video understanding.

SlowFast [59] combines two network branches: a "slow" branch that processes sparse temporal information, and a "fast" branch that handles more frequent temporal information. The slow branch processes fewer frames to capture more context and action details, while the fast branch can quickly capture dynamic changes. Through this approach, SlowFast achieves improved performance in various video analysis tasks, such as action recognition, object detection, and video segmentation.

SATD [60] is a deep learning architecture designed for modeling time series data. This network combines self-attention mechanisms and depthwise separable convolutions to effectively capture temporal and spatial relationships within the data. SATD employs self-attention mechanisms to focus on the relationships between different time steps in time series data, capturing long-range temporal dependencies. Depthwise separable convolutions help reduce the number of model parameters and improve computational efficiency. By integrating these two techniques, SATD enhances its ability to handle time series data, such as videos and time sequences, resulting in improved performance across various applications like action recognition, time series prediction, and more.

PoseConv3D [61] is a novel approach to skeleton-based action recognition. It relies on a 3D heatmap volume, rather than a sequence of images, as the fundamental representation of human skeletal structure. It can handle multi-person scenarios without incurring additional computational costs. Hierarchical features can be seamlessly integrated with other patterns in the early fusion stages, offering significant design flexibility for enhancing performance.

STA-GCN [62] is a spatiotemporal adaptive graph convolutional network designed to learn adaptive spatiotemporal topology and efficiently aggregate features for skeleton-based action recognition. The proposed network consists of spatial adaptive graph convolution (SA-GC) and temporal adaptive graph convolution (TA-GC), as well as an adaptive topology encoder. SA-GC leverages spatial adaptive topology to extract spatial features for each pose, while TA-GC learns temporal features by adaptively modeling direct distant temporal dependencies.

C. Results of Fine-grained Classification

The initial experiment in this study aims to assess the detection performance of both soccer players and soccer balls.

TABLE I
COMPARATIVE EXPERIMENTAL RESULTS FOR FINE-GRAINED CLASSIFICATION. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	AP	AR	A-F1
R(2+1)D CNN	0.513	0.438	0.482
ARTNet	0.616	0.463	0.551
P3D-CTN	0.728	0.619	0.632
I3D-LSTM	0.701	0.597	0.613
SlowFast	0.539	0.526	0.517
SATD	0.623	0.580	0.602
PoseConv3D	0.689	0.601	0.633
STA-GCN	0.671	0.599	0.635
STPR	**0.794**	**0.742**	**0.759**

TABLE II
ABLATION EXPERIMENT RESULTS. MODULE 1-4 RESPECTIVELY REPRESENT THE FOUR CORE MODULES MENTIONED ABOVE: MULTI-SCALE AGGREGATION STRUCTURE OF THE SPATIAL-TEMPORAL PYRAMID, THE SELF-ATTENTION MECHANISM, THE MULTI-TEMPORAL KERNEL DECOMPOSITION, AND THE ENERGY-MOTION FEATURE AGGREGATION MODULE.

Methods	Module 1	Module 2	Module 3	Module 4	A-F1
STPR-1	✓				0.612
STPR-2	✓	✓			0.694
STPR-3	✓		✓		0.659
STPR-4	✓	✓		✓	0.738

We employed the STPR method to accurately classify and detect fine-grained entities including soccer players, cameras, and referees. This involved an initial target screening process, followed by the simultaneous output of bounding boxes, segmentation masks, and human keypoints in a single iteration. Our evaluation metrics encompassed the Intersection over Union (IOU) between predicted and actual bounding boxes, as well as the processing speed measured in frames per second (FPS) during inference.

The model underwent training on the COCO dataset and was subsequently tested using captivating videos from soccer matches. The decision to train on the COCO dataset stemmed from its wide array of images, covering diverse subjects ranging from individual soccer players and goalkeepers, categorized under "Person," to soccer camera falling under the "sportsobject" classification.

Figure 3 visually presents the outcomes of the detection process applied to these thrilling soccer videos. It's crucial to emphasize that these outcomes are derived from all authentic annotations within the dataset, extending beyond instances of players dribbling or taking shots. The results indicate an impressive 92% accuracy in fine-grained classification and an average accuracy of 89.92% in player detection, all while maintaining an average inference speed of 8.5 FPS.

Of notable significance is the utilization of dribbling videos sourced from Bundesliga matches, employing cameras with a wider field of view to capture scenes removed from players engaged in actions. Despite the presence of camera distortion, its influence on the overall results remains minimal. With this type of data, observers gain a clearer understanding of both the collective strategies and individual states across the entire field.

D. Comparisons of Soccer Players' Fine-grained Action Analysis

We compared the proposed method with a series of existing algorithms, as described in Section IV-B. These algorithms are divided into two categories. The first category is video-based methods. Specifically, we used ResNet(2+1)D as the base model, which uses spatial 2D convolution and then temporal 1D convolution in each layer, and was pre-trained on the Kinetics dataset. We only fine-tuned the last convolutional block and the last fully connected layer on the soccer highlights video dataset. The second category is skeleton-based methods. We used models such as STA-GCN, PoseConv3D, etc. In each task, we divided the video dataset into training, validation, and test sets, and used random sampling to generate the splits. Following [63], we used 5-fold cross-validation to evaluate different algorithms, to achieve fair comparison, and gave the average precision, recall, and F1 score of all player actions. The average precision (AP) is defined as:

$$AP = \frac{1}{N}\sum_{i=1}^{N} P_{precision}(i) \tag{9}$$

The average recall (AR) is defined as:

$$AR = \frac{1}{N}\sum_{i=1}^{N} P_{recall}(i) \tag{10}$$

The average F1 score (A-F1) is defined as:

$$A-F1 = \frac{1}{N}\sum_{i=1}^{N} P_{F1-score}(i) \tag{11}$$

Where, in each formula, i represents each action category, and N represents the total number of action categories.

Table I showcases the outcomes and a comparative assessment between our approach (STPR) and a sequence of video-based algorithms concerning the meticulous scrutiny of players' shooting motions. The results vividly illustrate that our method stands out as the most adept in categorizing players' nuanced shooting actions. Furthermore, it is evident that the noteworthy improvements predominantly stem from a heightened average recall rate.

A notable observation emerges when considering video-based methodologies relying on standalone convolutional neural networks. These approaches tend to exhibit relatively subpar performance in the realm of fine-grained classification and action recognition. A case in point is the comparison between I3D-LSTM and R(2+1)D, wherein the former surpasses the latter in terms of performance. Our analysis posits that these techniques primarily excel in comprehensive, policy-driven, coarse-grained action analysis undertaken within extensive training settings. This specialization proves immensely challenging when applied to the context of fine-grained analysis, marked by limited data and annotations, especially within the context of sports-related scenarios. In contrast, our method adeptly translates intricate temporal and spatial information

Fig. 3. Fine-grained classification results for football videos. We intercepted the classification results of some scenes in the Bundesliga game, including dribbling players, goalkeepers, referees, and players in other positions. Furthermore, we have established confidence levels for different classification outcomes. Using this as a metric, we assess the likelihood of a category belonging to the current location. Based on its positional changes, we conduct a macroscopic analysis of the regular gameplay strategy.

into a singular representation, which the model harnesses to glean fine-grained features. This adept handling of complexity culminates in outstanding outcomes.

E. Ablation Study

In this section, we conduct ablation studies to evaluate how each module's features contribute to the proposed framework (STPR). We run STPR with four different network designs, which are respectively designed for the multi-scale aggregation structure of the spatial-temporal pyramid, the self-attention mechanism, the multi-temporal kernel decomposition, and the energy-motion feature aggregation module, respectively called STPR-1, STPR-2, STPR-3, and STPR-4. Each network removes or replaces a different module or technique. Specifically, STPR-1 directly uses a fixed single-scale computation without using the aggregation module, STPR-2 directly removes the self-attention mechanism for temporal adjustment and uses a fixed length, STPR-3 replaces the multi-temporal kernel with a general convolution kernel, and STPR-4 simply uses a naive feature fusion module instead of the energy-motion feature aggregation module. The naive feature fusion module tested consists of a series of fully connected layers with normalization layers, activation layers and DropOut layers. Table II shows the results of the ablation experiments. It can be seen that different modules have different degrees of improvement for the model performance.

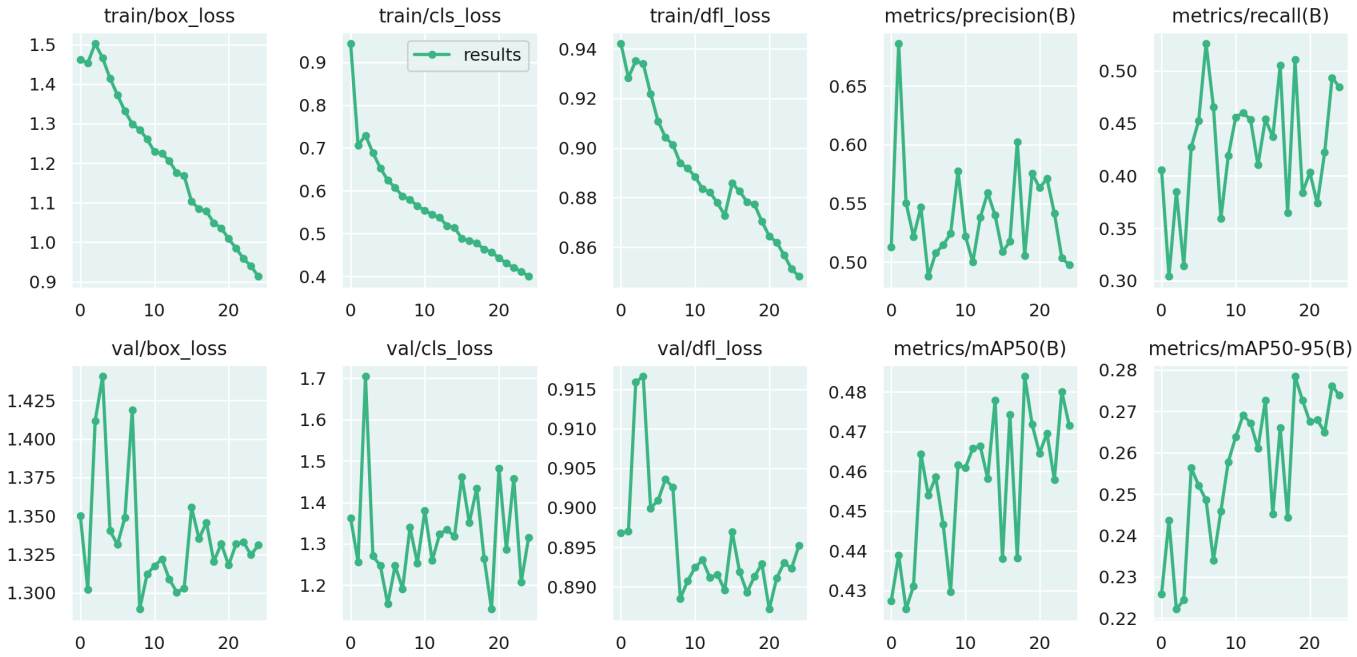

Fig. 4. Visualization of model convergence. The loss images of the model include training loss, validation loss, accuracy, recall, and a series of other metrics.

F. Convergence and Visualization

In order to show the model's loss function changes and performance improvements more intuitively during the training process, we record and display various metrics of the model in real time. We trained the model on the DFL dataset, and intercepted some of the results for visual display to analyze the convergence of the model more intuitively. Figure 4 shows the loss images of the model on these datasets, including training loss, validation loss, accuracy, recall, and a series of other metrics. It can be seen from the figures that our model can effectively reduce the loss function value and improve the classification accuracy and recall during the training process. It is worth noting that these metrics only represent a part of the results in the training process, to demonstrate our model's fast convergence characteristics. In order to evaluate our model's performance more comprehensively, we also test it on a separate test set for each dataset and compare it with several state-of-the-art models using different evaluation criteria such as map50. The results show that our model outperforms the existing models in most cases and achieves competitive performance in others.

V. Conclusive

In this paper, we propose a lightweight spatial-temporal pyramid reasoning approach designed to tackle three main challenges in automated sports video analysis: the complexities of action recognition and fine-grained classification, the non-uniformity of data distribution and limited temporal feature distinguishability, as well as coarse-grained interpretational capacity. This method employs a novel multi-scale spatial-temporal pyramid framework to infer spatial-temporal relationships. It accomplishes this by enabling the model to focus more precisely on critical spatiotemporal information using self-attention mechanisms, rather than solely channel features. Simultaneously, it leverages multi-temporal kernel decomposition to expand the network's temporal receptive field, enhancing feature distinctiveness while reducing computational costs. Additionally, the method employs dense sampling and trains video networks of varying lengths, fulfilling the demands of fine-grained queries. Experimental results demonstrate significant performance improvements of this approach across various video benchmark datasets, including datasets of exciting football matches. The article presents an effective and innovative solution for the field of sports video analysis.

References

[1] W. J. Baker, *Sports in the western world*. University of Illinois Press, 1988, vol. 160.

[2] V. Sarlis and C. Tjortjis, "Sports analytics—evaluation of basketball players and team performance," *Information Systems*, vol. 93, p. 101562, 2020.

[3] E. Morgulev, O. H. Azar, and R. Lidor, "Sports analytics and the big-data era," *International Journal of Data Science and Analytics*, vol. 5, pp. 213–222, 2018.

[4] K. Apostolou and C. Tjortjis, "Sports analytics algorithms for performance prediction," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 2019, pp. 1–4.

[5] A. Lees, "Technique analysis in sports: a critical review," *Journal of sports sciences*, vol. 20, no. 10, pp. 813–828, 2002.

[6] T. A. Herberger and C. Litke, "The impact of big data and sports analytics on professional football: A systematic literature review," *Digitalization, Digital Transformation and Sustainability in the Global Economy: Risks and Opportunities*, pp. 147–171, 2021.

[7] J. J. Coakley and E. Pike, "Sports in society: Issues and controversies," 2009.

[8] M. Svensson and B. Drust, "Testing soccer players," *Journal of sports sciences*, vol. 23, no. 6, pp. 601–618, 2005.

[9] T. Stølen, K. Chamari, C. Castagna, and U. Wisløff, "Physiology of soccer: an update," *Sports medicine*, vol. 35, pp. 501–536, 2005.

[10] D. Lusher, G. Robins, and P. Kremer, "The application of social network analysis to team sports," *Measurement in physical education and exercise science*, vol. 14, no. 4, pp. 211–224, 2010.

[11] K. A. Ericsson, "Training history, deliberate practice and elite sports performance: an analysis in response to tucker and collins review—what makes champions?" pp. 533–535, 2013.

[12] G. S. Schiftan, L. A. Ross, and A. J. Hahne, "The effectiveness of proprioceptive training in preventing ankle sprains in sporting populations: a systematic review and meta-analysis," *Journal of science and medicine in sport*, vol. 18, no. 3, pp. 238–244, 2015.

[13] V. Armatas, A. Yiannakos, and P. Sileloglou, "Relationship between time and goal scoring in soccer games: Analysis of three world cups," *International Journal of Performance Analysis in Sport*, vol. 7, no. 2, pp. 48–58, 2007.

[14] S. Li, B. Zhang, P. Fei, P. M. Shakeel, and R. D. J. Samuel, "Computational efficient wearable sensor network health monitoring system for sports athletics using iot," *Aggression and Violent Behavior*, p. 101541, 2020.

[15] F. J. Dian, R. Vahidnia, and A. Rahmati, "Wearables and the internet of things (iot), applications, opportunities, and challenges: A survey," *IEEE access*, vol. 8, pp. 69 200–69 211, 2020.

[16] A. Sabban, "Small new wearable antennas for iot, medical and sport applications," in *2019 13th European Conference on Antennas and Propagation (EuCAP)*. IEEE, 2019, pp. 1–5.

[17] D. Connaghan, P. Kelly, N. E. O'Connor, M. Gaffney, M. Walsh, and C. O'Mathuna, "Multi-sensor classification of tennis strokes," in *SENSORS, 2011 IEEE*. IEEE, 2011, pp. 1437–1440.

[18] T. Steels, B. Van Herbruggen, J. Fontaine, T. De Pessemier, D. Plets, and E. De Poorter, "Badminton activity recognition using accelerometer data," *Sensors*, vol. 20, no. 17, p. 4685, 2020.

[19] C. T. Kiang, C. K. Yoong, and A. Spowage, "Local sensor system for badminton smash analysis," in *2009 IEEE Instrumentation and Measurement Technology Conference*. IEEE, 2009, pp. 883–888.

[20] M. Bhatia, "Iot-inspired framework for athlete performance assessment in smart sport industry," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9523–9530, 2020.

[21] K. Rangasamy, M. A. As'ari, N. A. Rahmad, N. F. Ghazali, and S. Ismail, "Deep learning in sport video analysis: a review," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 4, pp. 1926–1933, 2020.

[22] W. Jingyao, Y. Naigong, and E. Firdaous, "Gesture recognition matching based on dynamic skeleton," in *2021 33rd Chinese Control and Decision Conference (CCDC)*. IEEE, 2021, pp. 1680–1685.

[23] E. E. Cust, A. J. Sweeting, K. Ball, and S. Robertson, "Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance," *Journal of sports sciences*, vol. 37, no. 5, pp. 568–600, 2019.

[24] N. A. Rahmad, M. A. As' ari, M. F. Ibrahim, N. A. J. Sufri, and K. Rangasamy, "Vision based automated badminton action recognition using the new local convolutional neural network extractor," in *Enhancing Health and Sports Performance by Design: Proceedings of the 2019 Movement, Health & Exercise (MoHE) and International Sports Science Conference (ISSC)*. Springer, 2020, pp. 290–298.

[25] Z. Cai, H. Neher, K. Vats, D. A. Clausi, and J. Zelek, "Temporal hockey action recognition via pose and optical flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[26] W. J. McNally, K. Vats, T. Pinto, C. Dulhanty, J. McPhee, and A. Wong, "Golfdb: A video database for golf swing sequencing," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 2553–2562. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2019/html/CVSports/McNally_GolfDB_A_Video_Database_for_Golf_Swing_Sequencing_CVPRW_2019_paper.html

[27] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[28] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[29] H. A. Qazi, U. Jahangir, B. M. Yousuf, and A. Noor, "Human action recognition using sift and hog method," in *2017 International Conference on Information and Communication Technologies (ICICT)*. IEEE, 2017, pp. 6–10.

[30] J. Wang and N. Yu, "Top-down meets bottom-up for multi-person pose estimation," in *2022 34th Chinese Control and Decision Conference (CCDC)*. IEEE, 2022, pp. 6086–6091.

[31] M. Pietikäinen, "Local binary patterns," *Scholarpedia*, vol. 5, no. 3, p. 9775, 2010.

[32] K. Seemanthini and S. Manjunath, "Human detection and tracking using hog for action recognition," *Procedia computer science*, vol. 132, pp. 1317–1326, 2018.

[33] G. Yao, T. Lei, and J. Zhong, "A review of convolutional-neural-network-based action recognition," *Pattern Recognition Letters*, vol. 118, pp. 14–22, 2019.

[34] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3218–3226.

[35] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition," *IEEE signal processing letters*, vol. 24, no. 4, pp. 510–514, 2016.

[36] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream cnn: Learning representations based on human-related regions for action recognition," *Pattern Recognition*, vol. 79, pp. 32–43, 2018.

[37] Y.-L. Chang, T.-H. Tan, W.-H. Lee, L. Chang, Y.-N. Chen, K.-C. Fan, and M. Alkhaleefah, "Consolidated convolutional neural network for hyperspectral image classification," *Remote Sensing*, vol. 14, no. 7, p. 1571, 2022.

[38] Y. Chen, J. Liu, X. Qi, X. Zhang, J. Sun, and J. Jia, "Scaling up kernels in 3d cnns," *arXiv preprint arXiv:2206.10555*, 2022.

[39] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE access*, vol. 6, pp. 1155–1166, 2017.

[40] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li, "Skeleton-based action recognition using lstm and cnn," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017, pp. 585–590.

[41] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 420–435.

[42] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 130–13 137.

[43] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[44] L. Liu, Z. Qiu, G. Li, Q. Wang, W. Ouyang, and L. Lin, "Contextualized spatial–temporal network for taxi origin-destination demand prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3875–3887, 2019.

[45] G. Luo, H. Zhang, Q. Yuan, J. Li, and F.-Y. Wang, "Estnet: embedded spatial-temporal network for modeling traffic flow dynamics," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 10, pp. 19 201–19 212, 2022.

[46] J. Wang, X. Ruan, and J. Huang, "Hdpp: High-dimensional dynamic path planning based on multi-scale positioning and waypoint refinement," *Applied Sciences*, vol. 12, no. 9, p. 4695, 2022.

[47] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 914–921.

[48] M. Aljoufie, M. Zuidgeest, M. Brussel, and M. van Maarseveen, "Spatial–temporal analysis of urban growth and transportation in jeddah city, saudi arabia," *Cities*, vol. 31, pp. 57–68, 2013.

[49] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, "Deep multi-view spatial-temporal network for taxi demand prediction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[50] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[51] F. Mahdisoltani, G. Berger, W. Gharbieh, D. Fleet, and R. Memisevic, "On the effectiveness of task granularity for transfer learning," *arXiv preprint arXiv:1804.09235*, 2018.

[52] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2616–2625.

[53] M. J. M. M. M. R. H. U. D. Jakub Michalczyk, Maggie, "Dfl - bundesliga data shootout," 2022. [Online]. Available: https://kaggle.com/competitions/dfl-bundesliga-data-shootout

[54] T. Geng, F. Zheng, X. Hou, K. Lu, G.-J. Qi, and L. Shao, "Spatial-temporal pyramid graph reasoning for action recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 5484–5497, 2022.

[55] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[56] M. Ibrar, A. Akbar, S. R. U. Jan, M. A. Jan, L. Wang, H. Song, and N. Shah, "Artnet: Ai-based resource allocation and task offloading in a reconfigurable internet of vehicular networks," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 1, pp. 67–77, 2020.

[57] J. Wei, H. Wang, Y. Yi, Q. Li, and D. Huang, "P3d-ctn: Pseudo-3d convolutional tube network for spatio-temporal action detection in videos," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 300–304.

[58] X. Wang, Z. Miao, R. Zhang, and S. Hao, "I3d-lstm: A new model for human action recognition," in *IOP Conference Series: Materials Science and Engineering*, vol. 569, no. 3. IOP Publishing, 2019, p. 032035.

[59] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[60] J. Zhang, G. Ye, Z. Tu, Y. Qin, Q. Qin, J. Zhang, and J. Liu, "A spatial attentive and temporal dilated (satd) gcn for skeleton-based action recognition," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 1, pp. 46–55, 2022.

[61] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2969–2978.

[62] R. Hang and M. Li, "Spatial-temporal adaptive graph convolutional network for skeleton-based action recognition," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 1265–1281.

[63] R. Li and B. Bhanu, "Energy-motion features aggregation network for players' fine-grained action analysis in soccer videos," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.