A Survey on Video Action Recognition in Sports: Datasets, Methods and Applications

Fei Wu, Qingzhong Wang, Jiang Bian, *Member, IEEE*, Ning Ding, Feixiang Lu, Jun Cheng, Dejing Dou, *Senior Member, IEEE*, and Haoyi Xiong, *Senior Member, IEEE*

***Abstract*—To understand human behaviors, action recognition based on videos is a common approach. Compared with image-based action recognition, videos provide much more information, reducing the ambiguity of actions. In the last decade, many works focus on datasets, novel models and learning approaches have improved video action recognition to a higher level. However, there are challenges and unsolved problems, in particular in sports analytics where data collection and labeling are more sophisticated, requiring people with domain knowledge and even sport professionals to annotate data. In addition, the actions could be extremely fast and it becomes difficult to recognize them. Moreover, in team sports like football and basketball, one action could involve multiple players, and to correctly recognize them, we need to analyze all players, which is relatively complicated. In this paper, we present a survey on video action recognition for sports analytics. We introduce more than ten types of sports, including team sports, such as football, basketball, volleyball, hockey and individual sports, such as figure skating, gymnastics, table tennis, tennis, diving and badminton. Then we compare numerous existing frameworks for sports analysis to present status quo of video action recognition in both team sports and individual sports. Finally, we discuss the challenges and unsolved problems in this area and to facilitate sports analytics, we develop a toolbox using PaddlePaddle [1], which supports football, basketball, table tennis and figure skating action recognition.**

***Index Terms*—Action recognition, video analysis, sports, computer vision, deep learning, survey**

I. Introduction

The number of videos is rapidly increasing and there is a massive demand of analyzing them, namely video understanding, such as understanding the behaviors of people, tracking objects, recognizing abnormal behaviors, and content-based video retrieval. Thanks to the development of video understanding technologies, there are many applications in our everyday life, *e.g.*, surveillance systems. Action recognition lies at the heart of video understanding, which is an elementary module for analyzing videos. Researchers have put much effort on action recognition, labeling a large number of videos [1], [4]–[9] and proposing many impressive models to improve the recognition accuracy [10]–[15]. However, the popular datasets like ActivityNet [5] and Kinetics-400 [16] only consider the activities in our daily life, such as walking, driving cars and riding bikes. Although, some datasets contains sports-related activities, the labels are coarse and it is difficult to directly use them for specific sports analysis. In addition, to achieve the goal of fine-grained sports action recognition, we need to label videos that focus on specific sports, such as football and basketball. Moreover, the fine-grained annotations normally require domain knowledge and professional players should be involved in video labeling. Figure 1 shows the

Fig. 1. The comparison among common actions in our daily life, actions in individual sports and actions in team sports. Top: common actions in UCF101 [1], which is a coarse annotated dataset for action recognition. Middle: figure skating actions in FSD-10 [2], which is a fine-grained annotated figure skating dataset. Bottom: activities in volleyball, basketball and football [3], where each action could involve multiple players.

F. Wu, Q. Wang, and J. Bian contributed equally to this work.

Q. Wang and H. Xiong are the corresponding authors. Please contact them via qingzwang@outlook.com and haoyi.xiong.fr@ieee.org.

Fei Wu and Ning Ding are with the Department of Physical Education, Peking University, Beijing, China. Emails: wufei@pku.edu.cn and dn620@stu.pku.edu.cn.

Qingzhong Wang, Jiang Bian, Feixiang Lu, Jun Cheng, and Haoyi Xiong are with Baidu Inc., Beijing, China. Email: qingzwang@outlook.com, bianjiang03@baidu.com, lufeixiang@baidu.com, chengjun@baidu.com, haoyi.xiong.fr@ieee.org.

Dejing Dou is with Boston Consulting Group (Greater China), Beijing, China. Email: dejingdou@gmail.com.

[1]The toolbox can be found at https://github.com/PaddlePaddle/PaddleVideo.

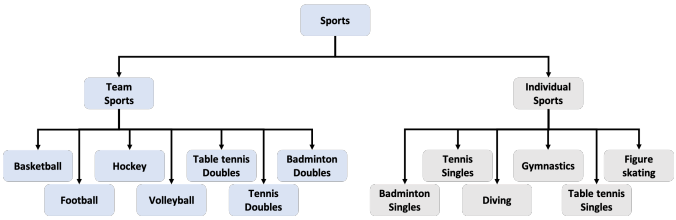


Fig. 2. An example of sports categories based on references [37]–[39].

comparison between common actions in our daily life and actions in specific sports, such as figure skating and basketball. Obviously, to annotate the professional action, such as 3Axel or 3Flip, domain knowledge is required. However, in many cases, it is not easy to find many annotators with domain knowledge for a specific sport, so a normal way is to hire a few people with domain knowledge to train more annotators for annotation. It could be difficult for annotators to discriminant some actions despite we train them, resulting in noisy labels.

Recently, researchers in the communities of computer vision and sports pay much attention to sports video analysis, including building datasets and proposing novel methodologies [2], [17]–[30]. In most existing works on sports video analysis, recognizing the actions of players in videos is crucial. On one hand, recognizing the group activities is able to assist coaches to make better decisions and players to understand their performances on fulfilling the coaches' strategies. On the other hand, recognizing the individual actions can benefit training players via correcting the small action errors [31], [32]. Another wide application of sports action recognition is in sports TV programs, where there is a massive demand of highlights generation and action recognition can significantly improves the localization accuracy [33]–[36].

However, there are many types of sports and each type of sport requires a specific model. Normally, we can roughly classify sports into team sports – individuals are organized into opposing teams that compete to win and individual sports – participants compete as individuals. In Figure 2, we present an example of sports categories. The analytics of team sports like football and individual sports such as diving is different. For team sports, each action could involve multiple players (see Figure 1) and each player has a specific action, such as dive and screen in basketball. In addition, the trajectory of the ball and the interaction between the ball and players are important in team sports analysis, hence, to accurate recognize the actions in team sports, we need to track the ball, multiple players and model the interactions [3]. While in individual sports, we can just pay attention to only one player to recognize the actions in most cases. Though in team sports, there could be only one player who possesses the ball, referring to individual ball possession, and we can track the player to analyze the individual actions, the trajectories and actions of other teammates and the interactions among players are also important for team sports and we can use the trajectories, actions and interactions to analyze the strategy of a team, such as offside trap, all-out attacking and total football. Normally, for team sports, tacking individual players and the ball is the first step and more effort is put into modeling the interactions in the following steps (please refer to section IV for more details), which is different from individual sports.

In this paper, we focus on video action recognition in various sports. One of the most related works is proposed by Y. Zhu *et al.* [40] – a study of deep video action recognition, but it does not pay much attention to sports. Similarly, Z. Sun *et al.* [41] propose a review of human action recognition in the perspective of data modalities such as RGB images, point cloud and WiFi signals, which does not focus on sports either. While D. Tan *et al.* [42] review video-based action recognition approaches in badminton, such as recognizing the actions of service and smashing, while team sports and other individual sports are not considered and the popular datasets used for action recognition are not introduced. Although J. Gudmundsson *et al.* [43], R. Bonidia *et al.* [44] and R. Beal *et al.* [45] review multiple sports, they pay much attention on sports data mining instead of video action recognition. M. Manafifard *et al.* [46] proposes a survey on player tracking in soccer videos, which also reviews video technologies like object tracking and detection, however, only soccer is taken into account. H. Shih [47] proposes a survey on video technologies in content-aware sports analysis, such as object and video event detection, while we focus on action recognition in sports and provide a deep learning toolbox that supports figure skating, football, basketball and table tennis action recognition, which is publicly available.

To sum up, the contributions of the survey are in three folds.

- First, we focus on the key part of sports video understanding – action recognition and introduce more than ten sports, including team sports like football, basketball, volleyball, hockey and individual sports such as diving, tennis, gymnastics and table tennis.
- Second, we provide a sports genre classification and road maps of action recognition methods in different types of sports. In addition, we present a summary of sports-related datasets for action recognition.
- Third, we present the current state of video action recognition in different types of sports and the challenges that should be paid attention to in the future. Moreover, to facilitate research in sports video action recognition, we provide a deep learning toolbox that supports video action recognition in multiple sports, which is publicly available at https://github.com/PaddlePaddle/PaddleVideo [2].

The rest of the paper is organized as follows. In section II, we introduce the sports-related datasets used for action recognition. We present the survey of methodologies for individual action recognition in section III, while in section IV, we review the methodologies for team activity recognition. In section V, we summarize the applications of video action recognition in sports, such as education and coaching. Section VI summarizes the challenges that should be paid more attention to in the future. Last but not least, we make conclusions in section VII.

TABLE I
A LIST OF SPORTS-RELATED DATASETS USED IN THE PUBLISHED PAPERS. NOTE THAT SOME OF THEM ARE NOT PUBLICLY AVAILABLE AND "MULTIPLE" MEANS THAT THE DATASET CONTAINS VARIOUS SPORTS INSTEAD OF ONLY ONE SPECIFIC TYPE OF SPORTS. "DET.", "CLS.", "TRA.", "ASS.", "SEG.", "LOC." STAND FOR PLAYER/BALL DETECTION, ACTION CLASSIFICATION, PLAYER/BALL TRACKING, ACTION QUALITY ASSESSMENT, OBJECT SEGMENTATION AND TEMPORAL ACTION LOCALIZATION, RESPECTIVELY. MORE DETAILS OF THE DATASET CAN BE FOUND IN SECTION II.

Datasets	Sports	Years	Modalities	Tasks	# Videos	Avg. length	# Categories	Publicly Available
CVBASE Handball [48]	handball	2006	RGB	CLS.	3	10m	-	Yes
CVBASE Squash [48]	squash	2006	RGB	CLS.	2	10m	-	Yes
UCF sports [49]	multiple	2008	RGB	CLS.	150	6.39s	10	Yes
APIDIS [50], [51]	basketball	2008	RGB	DET.& CLS.	-	-	-	Yes
Soccer-ISSIA [52]	football	2009	RGB	TRA.	-	-	-	Yes
MSR Action3D [53]	multiple	2010	RGB, depth	CLS.	567	-	20	Yes
Olympic [54]	multiple	2010	RGB	CLS.	800	-	16	Yes
Hockey Fight [55]	hockey	2011	RGB	CLS.	1,000	-	2	Yes
ACASVA [56]	tennis	2011	RGB	CLS.	6	-	4	Yes
THETIS [57]	tennis	2013	RGB, depth, skeleton	CLS.	1,980	-	12	Yes
Sports 1M [58]	multiple	2014	RGB	CLS.	1M	36s	487	Yes
OlympicSports [59]	multiple	2014	RGB	ASS.	309	-	2	Yes
SVW [60]	multiple	2015	RGB	DET.& CLS.	4,100	11.6s	44	Yes
Basket-1,2 [3]	basketball	2016	RGB	DET.& CLS.	-	-	4	No
Volleyball-1,2 [3]	volleyball	2016	RGB	DET.& CLS.	-	-	-	No
HierVolleyball [61]	volleyball	2016	RGB	DET.& CLS.	-	-	-	Yes
HierVolleyball-v2 [62]	volleyball	2016	RGB	DET.& CLS.	-	-	-	Yes
NCAA [63]	basketball	2016	RGB	CLS.& LOC.	14,548	4s	11	Yes
Football Action [64]	football	2017	RGB	CLS.	3,281	-	5	No
TenniSet [65]	tennis	2017	RGB, texts	LOC.& CLS.	5	-	6	Yes
OlympicScoring [66]	multiple	2017	RGB	ASS.	716	-	3	Yes
Soccer Player [67]	football	2017	RGB	TRA.& DET.	-	-	-	Yes
SPIROUDOME [68]	basketball	2017	RGB	DET.	-	-	-	Yes
SpaceJam [69]	basketball	2018	RGB	CLS.	15	1.5h	10	Yes
Diving48 [70]	diving	2018	RGB	CLS.	18,404	-	48	Yes
ComprehensiveSoccer [71]	football	2018	RGB	DET.& CLS.	220	0.77h	-	Yes
TTStroke-21 [72]	table tennis	2018	RGB	CLS.	129	43m	21	Yes
SoccerNet [24]	football	2018	RGB, audio	LOC.& CLS.	500	1.5h	3	Yes
Badminton Olympic [73]	badminton	2018	RGB	LOC.& CLS.	10	1h	12	Yes
SPIN [74]	table tennis	2019	RGB	TRA.& CLS.	-	-	-	No
GolfDB [75]	golf	2019	RGB	CLS.	1,400	-	8	Yes
AQA [76]	multiple	2019	RGB	ASS.	1,189	-	7	Yes
MTL-AQA [77]	diving	2019	RGB	ASS.	1,412	-	-	Yes
OpenTTGames [78]	table tennis	2020	RGB	SEG.& DET.	12	-	-	Yes
FineGym [79]	gymnastics	2020	RGB	CLS.& LOC.	-	-	288	Yes
SSET [80]	football	2020	RGB	TRA.& DET.	350	0.8h	30	Yes
SoccerDB [81]	football	2020	RGB	CLS.& LOC.	346	1.5h	11	Yes
FineBasketball [82]	basketball	2020	RGB	CLS.	3,399,	-	26	Yes
FSD-10 [2]	figure skating	2020	RGB	ASS.& CLS.	-	-	10	Yes
FineSkating [83]	figure skating	2020	RGB	ASS.& CLS.	46	1h	-	Yes
MCFS [84]	figure skating	2021	RGB	LOC.& CLS.	11,656	-	130	Yes
Stroke Recognition [85]	table tennis	2021	RGB	CLS.	22,111	-	11	Yes
MultiSports [28]	multiple	2021	RGB	LOC.& CLS.	3,200	20.9s	66	Yes
Player Tracklet [86]	hockey	2021	RGB	TRA.	84	36s	-	Yes
NPUBasketball [87]	basketball	2021	RGB, depth, skeleton	CLS.	2,169	-	12	Yes
SoccerNet-v2 [88]	football	2021	RGB, audio	LOC.& CLS.	500	1.5h	17	Yes
Win-Fail [89]	multiple	2022	RGB	CLS.	1,634	3.3	2	Yes
Stroke Forecasting [90]	badminton	2022	RGB	CLS.	43,191	-	10	Yes
FenceNet [91]	fencing	2022	RGB	CLS.	652	-	6	Yes

II. SPORTS-RELATED DATASETS

Datasets are required to facilitate model training and evaluation, in particular in the era of deep learning since deep models are normally data-hungry. Researchers have put much effort into developing new sports-related datasets. Generally, to construct a dataset for sports video action recognition, we need to (1) define the type of sports that we want to investigate and the categories of actions in the specific sport, (2) collect videos from multiple sources, such as the internet and self-recorded videos, (3) process the collected videos like trimming and then annotate the processed videos. The annotations could vary based on the goal of the proposed dataset, but it should provide trimmed videos and the corresponding labels or untrimmed videos with the start and end time of each action and the action category. In some datasets, the annotation process could be more complicated. *E.g.*, apart from annotating action labels and temporal positions, bounding boxes of objects that impose the actions are also annotated in AVA dataset [92]. In this section, we provide a comprehensive review of sports-related datasets and the list of datasets is shown in table I.

A. Football

Football is one of the most popular sports in the world and researchers pay much attention to football activity recognition, developing numerous datasets with different scales.

Soccer-ISSIA [52] is a relatively small dataset, composed of 18,000 high resolution frames recorded by 6 static cameras. The recorded videos are first automatically processed to extract blobs that indicate moving players and then the annotated bounding boxes are validated by humans. **Soccer-ISSIA** [52] are normally used for player tracking, detection and team activity recognition. Similarly, **Soccer Player** [67] is developed for player detection and tracking, comprising of

[2]The default language is Chinese and the English version can be found at https://github.com/PaddlePaddle/PaddleVideo/blob/develop/README_en.md. More details in English, including supported datasets, configuration, model zoo, installation and usage can be found at https://github.com/PaddlePaddle/PaddleVideo/tree/develop/docs/en.

2,019 annotated frames with 22,586 player bounding boxes. The limitation of this dataset is the scale is small.

Football Action [64] is a private dataset composed of self-recorded videos that are captured using 14 synchronized and calibrated Full HD cameras and the position of each player is annotated using a bounding box. There are five categories of activities: pass, shoot, loose clearance and dribble. Though the dataset is composed of videos recorded by multiple cameras, it is not publicly available.

ComprehensiveSoccer [71] is composed of 222 broadcast videos and 170 video hours in total. The dataset is annotated in 3 levels: positions of players using bounding boxes, event and story annotation at a coarse granularity and temporal annotations of shots. Totally, there are 11 categories of events, 15 types of stories and 5 types of shots, such as free kick&goal, corner and solo drive. The dataset can be used for various tasks in football video analysis, such as action classification, localization and player detection. The advantage of this dataset is dense and multi-level annotations, however, the video quality is low (360P and 720P) and the data distribution is imbalanced.

SoccerNet [24] is a large-scale dataset for football action recognition and localization. There are 500 complete soccer match videos collected from European leagues during 2014-2017. The total number of temporal annotations is 6,637 and the label of each temporal annotation is one of three categories: goal, substitution and yellow or red card. The actions are relatively sparse in **SoccerNet**, *i.e.*, there are only 8.7 actions per hour on average. One limitation of this dataset is that there are only 3 categories.

SSET [80] is three times smaller than **SoccerNet**, comprising of 350 football match videos, totaling 282 video hours. Similar to **ComprehensiveSoccer**, the annotations are in three levels: bounding boxes of players, event/story categories and shot categories, but **SSET** is larger than **ComprehensiveSoccer** dataset. Also, similar to **ComprehensiveSoccer**, the data distribution is imbalanced.

SoccerDB [81] is in the same scale as **SoccerNet**, which is composed of 171,191 video segments trimmed from 346 soccer match videos and the total length of the videos is 668.6 hours. **SoccerDB** also annotates the positions of players using bounding boxes, which contain 702.096 bounding boxes. 11 labels are taken into account for activity annotation, including goal, foul, injured, red/yellow card, shot, substitution, free kick, corner kick, saves, penalty kick and background. Each segment belongs to one category and has a time boundary. In addition, 17,115 highlights in soccer match videos are also annotated, therefore, the dataset can be used for player detection, activity recognition, activity localization and highlight detection.

SoccerNet-v2 [88] extends **SoccerNet** [24] via re-labeling the 500 untrimmed videos. In **SoccerNet**, there are only 3 categories, while **SoccerNet-v2** has 17 categories, such as throw-in, foul, indirect free kick, corner, shots on target, shots off target, direct free kick, clearance, substitution, kick-off, offside, yellow card, red card, goal, penalty, yellow-to-red card and ball out of play. Moreover, the actions in **SoccerNet-v2** are much denser than these in **SoccerNet**, *e.g.*, there is one action every 25 seconds in **SoccerNet-v2**, whereas, there is only 8.7 actions per hour in **SoccerNet**. Similar to **SoccerNet**, **SoccerNet-v2** can be employed for action recognition and localization.

Basically, large-scale datasets+deep models dominate the field of soccer video action recognition in recent years, increasing the popularity of **SoccerNet** [24] and **SoccerNet-v2** [88]. While **SoccerDB** [81], **SSET** [80] and **ComprehensiveSoccer** [71] are more feasible for the tasks that require player detection.

B. Basketball

Basketball has drawn much attention from researchers owing to its popularity in the world and numerous basketball datasets at different scales have been developed.

APIDIS [50], [51] is composed of seven videos of the same basketball match, which is recorded by seven calibrated cameras located in different positions on the basketball court. The positions of players and balls are annotated using bounding boxes. Clock and non-clock actions are also annotated, such as throw, violation, foul, pass, positioning and rebound. Each action has a time boundary and a label, thus, **APIDIS** can be used for both player detection and basketball action recognition. The dataset is challenging since the contrast between the background and players is low [67]. However, the small scale limits its applications.

Basket-1,2 [3] contains two basketball frame sequences – one has 4000 frames captured by 6 cameras and another has 3000 frames captured by 7 cameras. The cameras are synchronized and each can capture 25 frames per second. There are four action categories in the dataset: possessed ball, passed ball, flying ball, and ball out of play. **Basket-1,2** can be used for basketball action recognition and ball detection.

NCAA [63] is a relatively large dataset for basketball action recognition, composed of 257 untrimmed NCAA game videos and the video length are normally 1.5 hours. After processing, the dataset comprises 14,548 video segments with time boundary, each of which contains an action that belongs to one of 14 categories, such as 3-pint success, 3-point fail, steal, slam dunk success and slam dunk fail. In addition, **NCAA** also provides 9,000 frames with bounding boxes of players, therefore, people can also use it for player detection. There is no annotation of the ball, hence, it can not be used to model the interaction between the ball and players.

SPIROUDOME [68] is similar to **APIDIS**, where the videos are captured using 8 cameras. The positions of players are annotated using bounding boxes, therefore, **SPIROUDOME** is generally employed for player detection.

SpaceJam [69] comprises 10 categories of basketball actions, including step, race, block, dribble, ball in hand, shooting, position, walk, defensive position and no action. **SpaceJam** collects 15 videos of the NBA championship and the Italian championship from YouTube and the length of each video is 1.5 hours. Besides RGB images, the estimated poses of players are also provided. Normally, **SpaceJam** can be used to develop skeleton-based action recognition models. This dataset is small-scale, limiting its applications.

FineBasketball [82] is developed for fine-grained basketball action recognition, containing three broad categories – dribbling, passing and shooting, and 26 fine-grained categories, such as behind-the-back dribbling, cross-over dribbling, hand-off, one-handed side passing, lay up shot, one-handed dunk and block shot. There are 3,399 video segments in total and each category contains roughly 130 video segments on average. **FineBasketball** is challenging since the dataset is imbalanced, *e.g.*, there are 717 video segments belonging to crossover dribbling, while the class of follow-up shot only contains 12 video segments.

NPUBasketball [87] is composed of 2,169 self-recorded video clips of basketball actions performed by professional players and each video belongs to one of 12 categories: standing dribble, front dribble, moving dribble, cross-leg dribble, behind-the-back dribble, turning around, squat, run with the ball, overhead pass (catch or shoot), one-hand shoot, chest pass (catch or shoot), and side throw. Different from **FineBasketball** and **SpaceJam**, **NPUBasketball** provides not only RGB frames, but also depth maps and skeletons of players, thus, it can be used for developing various types of action recognition models. Since this dataset is composed of self-recorded videos, it is difficult to transfer the models trained on it to broadcasting videos.

C. Volleyball

Though volleyball is a relatively popular sport in the world, there are only a few volleyball datasets and most of them are on small scales.

Volleyball-1,2 [3] contains two sequences – one comprises 10,000 frames and another is composed of 19,500 frames. The positions of the ball are manually annotated using bounding boxes, however, detecting the ball is challenging since it moves fast and blurred after striking.

HierVolleyball [61] is developed for team activity recognition, containing 1,525 annotated frames from 15 YouTube volleyball videos. Each player has an action label defined as waiting, setting, digging, falling, spiking, blocking and others, and some players perform a group activity, such as set, spike and pASS.

HierVolleyball-v2 [62] extends **HierVolleyball**, comprising 4,830 annotated frames from 55 YouTube volleyball videos. There are 9 categories of players' actions: waiting, setting, digging, failing, spiking, blocking, jumping, moving and standing, and winpoint is also considered a team activity category. The positions of players are also annotated using bounding boxes, and they can be used for both player detection and action recognition.

Though the mentioned volleyball datasets are composed of dense annotations like player bounding boxes, the scale is relatively small and the action categories are coarse.

D. Hockey

Hockey Fight [55] is a proposed for binary classification: fight and non-fight in hockey games, composed of 1,000 video clips from National Hockey League (NHL) games. Each clip contains 50 frames and has a label indicating fight or non-fight.

Player Tracklet [86] comprises 84 video clips from broadcast NHL games and the average length of the videos is 36s. The positions of players and referees in each frame are annotated with bounding boxes and identity labels like players' names and numbers. **Player Tracklet** can be applied for player tracking and identification.

It lacks datasets for fine-grained hockey action recognition. There are only two categories in **Hockey Fight** and **Player Tracklet** is only for player detection. In addition, the scale of the two datasets is small.

E. Tennis

Tennis is an individual sport, attracting tens of millions of people and researchers have constructed various datasets for tennis video analysis.

ACASVA [56] is developed for tennis action recognition, in particular for evaluating primitive players' actions in tennis games, where there are six broadcast videos of tennis games and three categories of actions: hit, non-hit and serve. The positions of players and time boundaries of actions are labeled, however, the dataset only provides the extracted features of video clips instead of the original videos.

THETIS [57] is composed of 1,980 self-recorded videos belonging to 12 tennis actions: four backhand shots (backhand, backhand with two hands, backhand slice, backhand volley), four forehand shots (forehand flat, forehand slice, forehand volley, forehand open stands), three service shots (service flat, service kick, service slice) and smash. Besides RGB frames, **THETIS** also provides 1,980 depth videos, 1,217 2D skeleton videos and 1,217 3D skeleton videos, so it can be used for developing multiple types of action recognition models.

TenniSet [65] comprises five tennis videos of the 2012 London Olympic matches from YouTube and six categories of events are considered, such as set, hit and serve. The time boundary of each event is labeled, therefore, it can be used for both recognition and localization. Interestingly, **TenniSet** also provides textural descriptions of actions, such as "quick serve is an ace", so it can also be used for action retrieval.

The limitation of the existing tennis datasets is that the scale is small and annotations of **ACASVA** are coarse. Nevertheless, they provide multiple modalities, such as RGB frames, textual descriptions and depth maps, which benefit the research on multimodal learning.

F. Table Tennis

Similar to tennis, strokes in table tennis are important and multiple datasets have been developed for table tennis stroke recognition.

TTStroke-21 [72] is composed of 129 self-recorded videos of 94-hour games in the egocentric perspective. There are 1,378 annotated actions, each of which belongs to one of 21 categories, such as serve backhand spin, forehand push, backhand block and forehand loop. Though the strokes in table tennis games are relatively fast, **TTStroke-21** is not a challenging dataset and one possible reason is that the videos have a high frame rate (120 FPS).

SPIN [74] also comprises self-recorded videos captured by two high-speed cameras (150 FPS), totaling 53 hours and 7.5 million high-resolution (1024×1280) frames. The positions of the ball are annotated using bounding boxes and 30 locations of players' joints are also labeled using heatmaps (15 joints for each player) in each frame. The dataset can be used for multiple tasks like ball tracking, pose estimation and spin prediction based on the trajectory of the ball and the player's poses.

OpenTTGames [78] consists of 12 HD videos of table tennis games (5 videos for training and 7 short videos for testing). Ball coordinates are annotated in each frame and 4,271 events are labeled, each of which has a label – ball bounces, net hits or empty events. In addition, 4 frames before each event and 12 frames after are labeled using segmentation masks, including human, table and scoreboard, hence, **OpenTTGames** can be used for semantic segmentation, ball tracking and event classification.

Stroke Recognition [85] is similar to **TTStroke-21** but much larger, composed of 22,111 trimmed videos and each video contains a stroke belongs to one of 11 categories. The dataset is less challenging, *e.g.*, a random forest with 21 trees achieves an accuracy of 96.20% [85].

P^{2}A [93] is one of the largest datasets for table tennis analysis, composed of 2,721 untrimmed broadcasting videos, and the total length is 272 hours. The authors annotate each stroke in videos, including the category of the stroke and the indices of the starting and end frames. Plus, the stroke labels are confirmed by professional players, including Olympic table tennis players.

Since all datasets except for **P^{2}A** are composed of self-recorded videos, limiting the applications of these datasets. Though **P^2** uses broadcasting videos, the data is imbalanced and the annotations are noisy.

G. Gymnastics

There are few datasets for gymnastics and one recent work named **FineGym** [79] is developed for gymnastic action recognition and localization, consisting of 303 videos with around 708-hour length. **FineGym** is annotated in a hierarchical manner, *e.g.*, there are four high-level event labels, 15 categories of action sets for 4 events and 530 categories of element actions. The time boundaries of actions and sub-actions are labeled, therefore, **Gymnastics** can be used for fine-grained action recognition and localization. The task of event/set-level action recognition and localization is relatively easy, while element-level action recognition and localization are much more challenging.

H. Badminton

Badminton Olympic [73] is composed of 10 videos of "singles" badminton matches from YouTube and each video is generally within one hour. There are multiple types of annotations in the dataset. First, the positions of players in 1,500 frames are annotated using bounding boxes. Second, 751 temporal locations of when a player wins a point are annotated. Third, the time boundaries and labels of strokes are annotated, where there are 12 categories of strokes, such as serve and lob. With three types of annotations, **Badminton Olympic** can be used for multiple tasks – player detection, point localization, action recognition and localization.

Stroke Forecasting [90] is a most recent dataset, consisting of 43,191 trimmed video clips and each video clip has a stroke that belongs to one of 10 categories – smash, push, clear, defensive shot, net shot, drive, drop, lob, long service and short service. In addition to badminton action recognition, the dataset can also be used for stroke forecasting, *i.e.*, given previous strokes in a rally, the model should predict what the next stroke is.

I. Figure skating

There are three dataset proposed for figure skating action recognition in recent years – **FSD-10** [2], **FineSkating** [83] and **MCFS** [84].

FSD-10 [2] comprises ten categories of figure skating actions (Change Combination Spin 4, Fly Camel Spin 4, Choreo Sequence 1, Step Sequence 3, Double Axel, Triple Axel, Triple Flip, Triple Loop, Triple Lutz, Triple Lutz-Triple Toeloop) and each action has 91-233 video clips, ranging from 3s to 30s. In addition to action labels, **FSD-10** also provides scores of actions for action quality assessment.

FineSkating [83] is composed of 46 videos of figure skating competitions in 2018 and 2019, each of which is around 1 hour long. The labels are designed in a hierarchical manner, *i.e.*, event labels and action labels. There are seven event labels, such as jump and spin, and each event has multiple actions, *e.g.*, the event of jump contains 7 actions: Axel, Flip, Toeloop, Loop, Lutz, Salchow and Euler. Moreover, the start time, end time and score of each action are also labeled, hence, it can be used for both action recognition and action quality assessment.

MCFS [84] consists of 11,656 video segments from 38 figure skating competitions, totaling 17.3 hours and 1.7 million frames. Similar to **FineGym** [79], **MCFS** has three-level annotations: 4 set (jump, spin, sequence, none), 22 subsets (Camel spin, Axel,$\cdots$) and 130 element actions (double Axel, double Flip, triple Axel, $\cdots$). The time boundaries of actions are also annotated, so **MCFS** can be applied for action recognition and localization.

J. Diving

Diving48 [70] contains 16,067 diving video segments for training and 2,337 for testing, totaling 18,404 video segments and covering 48 fine-grained categories of diving. Each class of action is composed of multiple elements, such as backward take-off and a half twist. Compared with existing datasets for action recognition, **Diving48** has a relatively low bias, which is fairer for model evaluation.

By contrast, **MTL-AQA** [77] is developed for diving action quality assessment, consisting of 1,412 samples and each sample is annotated with an action quality score, action class and textural commentary, therefore it can be used for multiple tasks, including action quality assessment and recognition.

K. Multiple Types of Sports

There are several datasets supporting multiple sports classification, where each video has a label indicating the category of sports, such as football, basketball and gymnastics, and a model is supposed to classify the videos. Generally, these datasets are used for coarse classification.

UCF sports [49] is proposed in 2008, composed of 150 video clips with 10FPS. The length of videos ranges from 2.02s to 14.40s and there are 10 categories, including diving, golf swing, kicking, lifting, riding a horse, running, skateboarding, swing bench, swing side and walking.

Two years later, W. Li *et al.* [53] develop **MSR Action3D**, which contains 576 sequences of depth maps instead of RGB frames and people can use it to recognize sports actions, such as tennis serve, tennis swing and golf swing. The videos are in **MSR Action3D** are self-recorded.

Olympic [54] is a relatively large dataset, including 800 videos for 16 categories like long jump, high jump, tennis serve, diving and vault, and each category has 50 videos. The videos in Olympics are from Youtube instead of self-recorded, therefore, occlusions and camera movements are involved in videos, being more challenging.

Sports 1M [58] is a much larger dataset, containing around one million videos that are from YouTube and 487 categories. There are 1,000-3,000 videos from each category so the distribution of videos is relatively balanced. Moreover, the labels are designed in a hierarchical manner, *i.e.*, the high-level nodes like team sports, ball sports, winter sports are used for coarse classification and the leaf nodes, such as eight-ball, nine-ball and blackball of billiards can be used for fine-grained classification. To some extent, using this million-scale dataset, we can alleviate the problem of data-hungry in deep learning.

SVW [60] is a dataset for both action classification and detection, composed of 4,100 videos and 44 action categories belonging to 30 types of sports, such as soccer, swimming, tennis and volleyball. One property of this dataset is that the videos are captured by smartphones from the view of coaches and the quality of the videos is normally lower than the broadcasting videos, resulting in challenges for action recognition.

THUMOS [94] is a challenge on untrimmed video action recognition and in THUMOS'15, the training dataset is composed of 13,000 trimmed videos from UCF101 [1] action classes and the validation and test datasets are composed of untrimmed videos, so it can be used for two tasks: action classification and temporal action localization. **Multi-THUMOS** [95] extends THUMOS'14 dataset using dense, multi-label, and frame-level action annotations, which is composed of 400 videos with 38,690 annotations of 65 action classes.

Recently, **MultiSports** [28] is proposed for multi-person sports, which is more challenging since each activity can involve multiple players who can perform different actions. The dataset covers four team sports – aerobic gymnastics, football, basketball and volleyball, and 66 categories of actions. There are 3,200 videos and 37,701 action instances. Apart from annotating video segments (temporal labels), **MultiSports** also provides bounding boxes of players involved in the activities, therefore, it can be used for action recognition, temporal and spatial localization.

Besides recognizing the actions in sports, some other datasets are proposed for action assessment, *i.e.*, a model should not only recognize the actions, but also provide a score that indicates the quality of the action. **OlympicSports** [59] is proposed to evaluate the quality of diving and figure skating actions, comprising 159 diving videos and 150 figure skating videos from Youtube, while **OlympicScoring** [66] extends it by collecting more videos and introducing more types of sports, which is composed of 370 diving videos, 170 figure skating videos and 176 vault videos. However, the number of videos in **OlympicScoring** is still limited for deep learning based methods. In contrast, **AQA** [76] dataset includes seven categories of sports: synchronous diving–10m platform, singles diving–10m platform, synchronous diving–3m springboard, gymnastic vault, skiing, snowboarding and trampoline. There are 1,189 videos in total.

Interestingly, **Win-Fail** [89] is proposed for recognizing win or fail of actions. Though actions could be very complex, the results of actions, *i.e.*, win/fail can be recognized via reasoning on the movements of objects. **Win-Fail** is composed of 817 win-fail video pairs collected from multiple domains like trick shots and internet win-fails.

L. Others

CVBASE Handball [48] is developed for handball action recognition, comprising three synchronized videos and each video is 10-minus long. The trajectories of seven players, team activities like offensive, defensive and individual actions like passes, shot are annotated. Similar to **CVBASE Handball**, **CVBASE Squash** [48] composed of two 10-minus videos of different matches also provides trajectories of players and categories of strokes, such as lob, drop and cross.

GolfDB [75] is proposed to facilitate the analysis of golf swings, consisting of 1,400 high-quality golf swing video segments belonging to eight swing categories, such as toe-up, top, impact and so on. In addition to action labels, **GolfDB** also provides bounding boxes of players, player name and sex.

FenceNet [91] is composed of 652 videos belonging to 6 categories – rapid lunge, incremental speed lunge, with waiting for lunge, jumping sliding lunge, step forward, and step backward. The actions are performed by expert-level fencers. In addition to RGB frames, the dataset also provides 3D skeleton data and depth data.

III. Individual Action Recognition

In this section, we dive in to the review of individual action recognition, *i.e.*, each action involves only one person.

A. Traditional Models

Generally, an action recognition model consists of at least two modules: (1) video feature extraction and (2) classifier, which is shown in Fig. 3. Hand-crafted features dominate traditional models. One simple approach is extracting low/middle-level features of each frame using GIST [96]

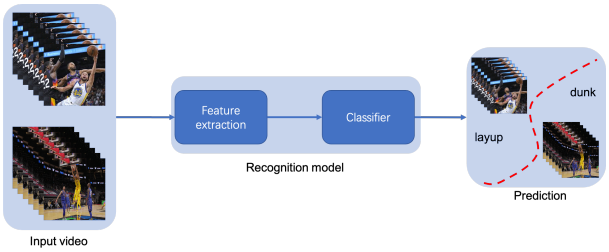


Fig. 3. An illustration of action recognition models. Generally, a feature extraction module and a classifier are required for action recognition.

or *Histogram of Oriented Gradients* (HOGs) [97] and then averaging the frame features over time for classification [4]. H. Kuehne *et al.* [4] evaluate multiple feature extraction approaches on various datasets, such as **UCF Sports** [49], showing that using GIST features achieves better performance (60.0%) than using HOGs (58.6%) on **UCF Sports** since the features are biased to the background, *e.g.*, the sports of ball normally occur on grass field.

Instead of using 2D HOGs, E. Ijjina [98] applies HOG3D [99] to extract video features and a *multi-layer perceptron* (MLP) as classifier. In contrast, T. Campos *et al.* [56] employ HOG3D features + *kernelized Fisher discriminant analysis* (KFDA) for tennis action recognition, achieving AUC of 84.5% on ACASVA [56].

Action bank is proposed by S. Sandanand and J. Corso [100], which is a high-level representation for action recognition, Action bank employs a template-based action detector, which is invariant to appearance changes. The detector is also applied to multi-scale and multi-view videos to be more robust to scales and viewpoints. After that, template actions are selected. Generally, an action bank with N action detectors and M samples yields a $N \times M \times 73$-D feature space. Using an action bank for feature extraction achieves the accuracy of 95% on **UCF sports**.

It is believed that motion plays an important role in action recognition, and various approaches are proposed to use motion information for action recognition, such as *Motion Boundary Histogram* (MBH) [101], *Histograms of Optical Flow* (HOF) [102] and dense trajectories [103], all of which are based on optical flow. MBH is more robust to camera motion, achieving better performance. H. Wang *et al.* [104] propose improved trajectories for action recognition, where camera motion is taken into account, and the model is able to concentrate on the moving objects, achieving much better performance, *e.g.*, using the original trajectories achieves the accuracy of 62.4% on **Olympic** dataset and MBH achieves 82.4%, whereas using the improved trajectories finally obtains 91.1% on **Olympic** [54].

In addition to HOG, *Scale-Invariant Feature Transform* (SIFT) [105] is also widely applied to action recognition. M. Chan *et al.* [106] propose motion SIFT (MoSIFT) to extract video features, where both spatial and temporal are considered, *i.e.*, first, MoSIFT employs histogram of gradients to extract spatial appearance and then employs histogram of optical flow to extract motion features. MoSIFT achieves 89.5% accuracy on **Hockey Fight** [55], outperforming *Space-Time Interest Points* (STIP) [107] (59.0%).

TABLE II
TRADITIONAL MODELS FOR ACTION RECOGNITION.

Method	Venue	UCF Sports	Olympic
Kovashka *et al.* [108]	CVPR-2010	87.27	-
Wang *et al.* [104]	CVPR-2011	88.20	-
Klaser *et al.* [109]	THESIS-2010	86.70	-
Wu *et al.* [110]	CVPR-2011	91.30	-
Sadanand *et al.* [100]	CVPR-2012	88.20	-
Wang *et al.* [111]	BMVC-2009	-	92.10
Laptev *et al.* [112]	CVPR-2008	-	91.80
Wong *et al.* [113]	CVPR-2007	-	86.70
Schuldt *et al.* [114]	ICPR-2004	-	71.50
Kim *et al.* [115]	CVPR-2008	-	95.00
Niebles *et al.* [54]	ECCV-2010	-	72.10

Though spatial-temporal features extracted using HOG, HOF and SIFT can achieve relatively good performance on sports action recognition datasets like **UCF Sports** and **Olympic** (see Table II), it is normally time-consuming to calculate hand-crafted spatial-temporal features. Moreover, traditional models cannot be trained in an end-to-end manner, *i.e.*, the feature extraction module and classifier are learned separately. Recently, researchers pay more attention to deep learning models, proposing many approaches to sports video action recognition and boosting the accuracy of recognition to a higher level.

B. Deep Models

Currently, deep models dominate video action recognition. Traditional methods normally require many storage spaces to store the extracted features and they are not appropriate for large-scale datasets, while deep models are more feasible and can be trained in an end-to-end manner via SGD by running many steps and thanks to the development of GPU and distributed parallel computing techniques, which makes deep learning methods appropriate to million-scale video action recognition. Typically, there are four types of deep models: 2D model, 3D model, Two/multi-stream model and skeleton-based model. We show the basic architectures of four typical models in Fig. 4 and more details can be found in the following subsections.

1) ***2D Models****:* 2D models employ 2D convolutional neural networks (CNN) or transformers [154] to process each video frame separately and then fuse the extracted features for prediction.

A. Karpathy *et al.* [58] introduce CNNs into video action recognition, proposing four-time information fusion approaches: (1) single-frame fusion – using a shared CNN to extract features of every single frame and then concatenate the final representations for classification, (2) early fusion – using a 3D kernel with the size of $11 \times 11 \times 3 \times T$ to combine information of frames across a time window, (3) late fusion – using a shared CNN to compute the representations of two separate frames with the distance of 15 frames and a fully connected layer to fuse the single-frame representations (4) slow fusion – implementing a 3D kernel in the first layer and then slowly fusing the information of frames in higher layers of the network. The experiments show that slow fusion is superior to other fusion approaches, *e.g.*, slow fusion obtains 60.9% accuracy on **Sports 1M** [58], while single-frame fusion,

TABLE III
DEEP LEARNING MODELS FOR INDIVIDUAL ACTION RECOGNITION.

Type	Method*	Venue	Pre-train	Backbone	Generic			Sports				
					Kinetics400	UCF101	HMDB51	Sports1M	FineGym	FSD-10	P^2A	Diving48
2D	Slow fusion [58]	CVPR-2014	-	-	-	-	-	60.9	-	-	-	-
	CNN-LSTM [116]	CVPR-2015	ImageNet	GoogLeNet	-	88.6	-	73.1	-	-	-	-
	LRCN [117]	CVPR-2015	ImageNet	AlexNet	-	82.7	-	-	-	-	-	-
	Composite LSTM [118]	ICML-2015	ImageNet, Sports1M	VGG-16	-	75.8	44.0	-	-	-	-	-
	LENN [119]	CVPR-2016	-	VGG-16	-	76.3	-	-	-	-	-	-
	TSN [120]	TPAMI	ImageNet	ResNet50, BN-Inception	-	87.3	-	-	61.4	59.3	72.1	-
	Attention-LSTM [121]	CVPR-2018	ImageNet	Inception-ResNet-v2, ResNet152	79.4	94.6	69.2	-	-	-	73.3	-
	TSM [13]	ICCV-2019	ImageNet	ResNet50	74.1	95.9	73.5	-	70.6	-	76.8	-
	KTSN [2]	arxiv	-	-	-	-	-	-	-	63.3	-	-
	TimeSformer [122]	ICML-2021	ImageNet	ViT-base	78.0	-	-	-	-	-	77.4	81.0
	VidTr [123]	ICCV-2021	ImageNet	ViT-base	80.5	96.7	74.4	-	-	-	-	-
	VTN [124]	ICCVW-2021	ImageNet	ViT-base	79.8	-	-	-	-	-	-	-
	RViT [125]	CVPR-2022	ImageNet	wide ViT-base	81.5	-	-	-	-	-	-	-
3D	C3D [126]	ICCV-2015	Sports1M	VGG16	59.5	82.3	56.8	61.1	-	-	-	-
	I3D [7]	CVPR-2017	ImageNet, Kinetics	BN-Inception	71.1	95.6	74.8	-	63.2	-	-	-
	P3D [127]	ICCV-2017	Sports-1M	ResNet50	71.6	88.6	-	-	-	-	-	-
	R(2+1)D-RGB [128]	CVPR-2018	Sports1M, Kinetics	R3D-34	72.0	96.8	74.5	73.0	-	-	-	-
	S3D [129]	ECCV-2018	ImageNet, Kinetics	BN-Inception	74.7	96.8	-	-	-	-	-	-
	CSN [130]	ICCV-2019	IG-65M [131]	R3D-152	82.6	-	-	75.5	-	-	-	-
	SlowFast [12]	ICCV-2019	-	ResNet101	79.8	-	-	-	-	-	77.4	77.6
	STM [132]	ICCV-2019	ImageNet, Kinetics	ResNet50	73.7	96.2	72.2	-	-	-	-	-
	X3D [133]	CVPR-2020	-	-	79.1	-	-	-	-	-	-	-
	TPN [134]	CVPR-2020	-	ResNet101	79.8	-	-	-	-	-	-	-
	ViViT [14]	CVPR-2021	-	ViViT-large	81.3	-	-	-	-	-	-	-
	MViT [135]	CVPR-2021	-	MViT-base	81.2	-	-	-	-	-	-	-
	MoViNet [136]	CVPR-2021	-	MoViNet-v6	81.5	-	-	-	-	-	74.1	-
	Mformer [137]	NeurIPS-2021	ImageNet	ViT-base	81.1	-	-	-	-	-	-	-
	ViSwin [138]	arXiv	ImageNet	ViSwin-large	84.9	-	-	-	-	-	81.0	-
	ORViT TimeSformer [139]	arXiv	ImageNet	ORViT	-	-	-	-	-	-	-	88
	BEVT [140]	arXiv	ImageNet, Kinetics	ViSwin-base	80.6	-	-	-	-	-	-	86.7
	MaskFeat [141]	arxiv	Kinetics	MViT-large	87.0	-	-	-	-	-	-	-
	VIMPAC [142]	arXiv	HowTo100M [143]	BERT-L [144]	77.4	92.7	65.9	-	-	-	-	85.5
	TFCNet [145]	arXiv	ImageNet	R3D-50	-	-	-	-	-	-	-	88.3
Two-stream	Two-Stream ConvNet [10]	NIPS-2014	-	-	-	88.0	59.4	-	-	-	-	-
	Two-Stream Fusion [146]	CVPR-2016	-	VGG-16	-	92.5	65.4	-	-	-	-	-
	TSN-Two-Stream [120]	ECCV-2016	ImageNet	ResNet50, BN-Inception	73.9	94.0	68.5	-	76.4	76.0	72.1	-
	R(2+1)D-Two-Stream [128]	CVPR-2018	Sports1M, Kinetics	R3D-34	75.4	97.3	78.7	73.3	-	-	-	-
	TRN-Two-Stream [147]	ECCV-2018	ImageNet	BN-Inception	63.3	83.8	-	-	79.8	-	-	-
	TSM-Two-Stream [13]	ICCV-2019	ImageNet	ResNet50	-	-	-	-	81.2	-	-	-
	KTSN-Two-Stream [2]	arXiv	ImageNet	ResNet50, BN-Inception	-	94.9	82.1	-	-	82.6	-	-
	G-Blend [148]	CVPR-2020	IG-65M	R3D-50	83.3	-	-	62.8	-	-	-	-
Skeleton	ST-GCN [149]	AAAI-2018	-	GCN	30.7†	-	-	-	25.2	60.5	-	-
	AGCN [150]	TIP	-	GCN	36.1†	-	-	-	-	65.9	-	-
	EfficientGCN [151]	MM-2020	-	GCN	-	-	-	-	-	65.5	-	-
	CTR-GCN [152]	ICCV-2021	-	GCN	-	-	-	-	-	66.2	-	-
	PoseC3D [153]	CVPR-2022	-	C3D	47.7†	-	-	-	94.3	68.8	-	-

*All the reported methods have been evaluated on at least one sports video dataset or related.
†The performances of all the skeleton-based algorithms are conducted on the Kinetics-Skeleton-400 dataset.

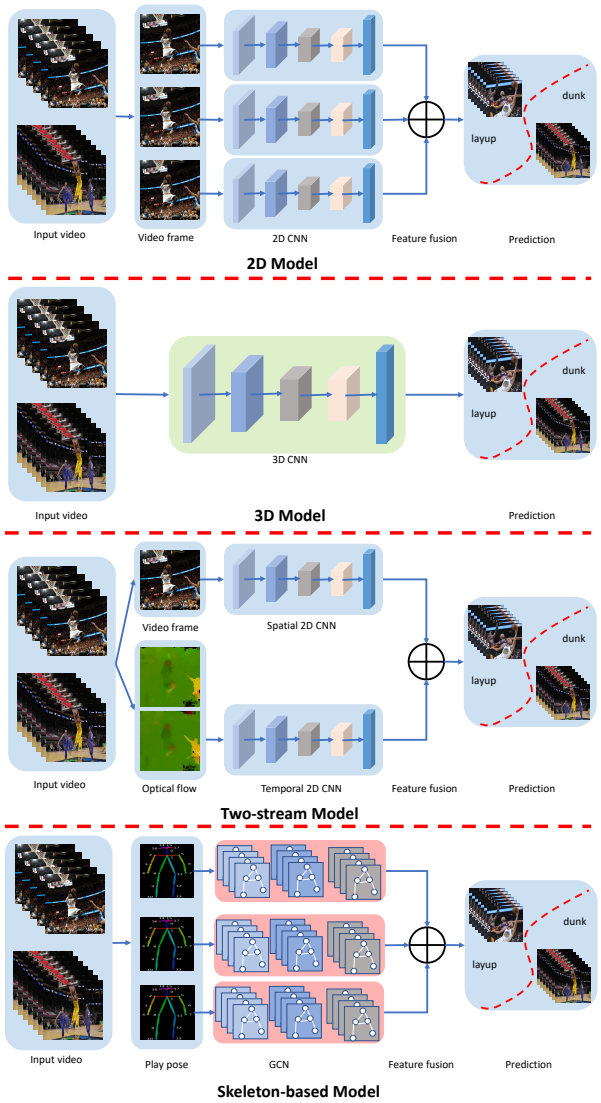


Fig. 4. An illustration of deep models for action recognition. We present 4 types of deep models: **2D model**, **3D model**, **two-stream model** and **skeleton-based model**. Note that we only present the basic frameworks and there could be some other variants (more details can be found in section III-B).

early fusion and late fusion achieve 59.3%, 57.7% and 59.3%, respectively. Interestingly, using hand-crafted features like HOG only achieves 55.3% accuracy, which is considerably lower than using CNNs, indicating that deep models are promising for sports video action recognition and inspiring researchers to develop more deep models.

Another family of 2D deep models is directly using *Long-short Term Memory* (LSTM) [155] networks to capture temporal information, which is relatively popular in early deep models. In 2015, Y. Ng *et al.* [116] propose an approach that combines 2D CNNs and LSTMs, *i.e.*, first, using a shared 2D CNN to obtain spatial representations of frames and then applying a multi-layer LSTM to fuse the spatial representations. Also, J. Donahue *et al.* [117] propose a similar model which uses a two-layer LSTM, termed *Long-term Recurrent Convolutional Networks* (LRCN). While N. Srivastava *et al.* [118] employ an LSTM-based auto-encoder to learn better video representations trained in an unsupervised manner. Latter, C. Gan *et al.* [119] propose a *Lead-exceed Neural Network* (LENN) which is similar to the model in [116], but LENN uses web images to fine-tune the lead network to filter out irrelevant video frames.

As mentioned above, temporal information fusion is crucial in 2D models. Alternatively, L. Wang *et al.* [120] propose a *Temporal Segment Network* (TSN) for video action recognition, which is composed of a spatial CNN and a temporal CNN. First, an input video is divided into some segments and the short snippets composed of RGB frames, optical flow and RGB differences are randomly sampled from segments. After that, the snippets are fed into spatial and temporal networks to make predictions. Finally, we can obtain a prediction by aggregating the snippet prediction scores. TSN uses temporal information in two ways: (1) it directly introduces optical flow into the framework, (2) similar to late fusion in [58], TSN aggregates the snippet predictions. Finally, the 2D TSN that only using RGB frames obtains impressive performance, *e.g.*, 61.4% accuracy on **FineGym** [79] and 87.3% on the generic action recognition dataset – **UCF101** [1]. Another variant of TSN is using key video frames instead of random sampling, namely KTSN [2]. Applying key video frames achieves better performance on **FSD-10**, *i.e.*, 63.3% vs. 59.3%.

Instead of using simple aggregation approaches, such as concatenation and linear combination, B. Zhou *et al.* [147] propose a *Temporal Relational Network* (TRN) to capture the temporal relations among frames, where the relations are computed using an MLP and can be plugged into any existing frameworks. TRN remarkably improves the performance on **FineGym** [79], obtaining 68.7% accuracy.

However, using MLPs in TRN is time-consuming when considering many frames and cannot well capture useful low-level features. To address this issue, J. Lin *et al.* [13] propose a simple yet efficient module, namely *Temporal Shift Module* (TSM) to capture temporal information for action recognition, where spatial features are extracted using 2D CNNs on video frames and then inserting TSM into 2D convolutional blocks. TSM achieves 70.6% accuracy on **FineGym** [79], outperforming 2D TSN, 2D TRN and some 3D approaches like I3D [7] but having lower computational complexity.

In recent 2 years, vision transformers (ViT) [154] become increasingly popular for computer vision tasks, where multi-head self-attention [156] is employed to replace convolutional kernels. G. Bertasius *et al.* [122] investigate different combinations of spatial self-attention and temporal self-attention (space-only, joint space-time, divided space-time, sparse local-global and axial attention), where spatial attention is performed over patches belong to the same video frame and temporal attention is applied to patches across frames, yielding a model termed TimeSformer. Experiments show that using divided space-time attention outperforms other architectures, achieving 81.0% accuracy on **Diving48** [70]. Similarly, VidTr [123] employs separable attention (temporal attention first and then spatial attention) to reduce computational complexity, achieving 80.5% accuracy on **Kinetics-400**. While *Vision Transformer Network* (VTN) [124] employs a temporal transformer to fuse frame representations, obtaining 79.8% accuracy on **Kinetics-400**. Instead of using temporal attention to fuse information of different frames, RViT [125] employs recurrent mechanism,

which takes less memory and obtains competitive performance on **Kinetics-400**, *i.e.*, 81.5% accuracy.

In summary, for 2D deep models, we can find that both spatial and temporal modules are shifting to transformers since transformers are much more powerful to model sequences and to extract frame features, however, transformers have more learnable parameters, requiring more computational resources. In addition, training a large model is non-trivial due to the difficulty of convergence. Another trend is adopting pre-training, *i.e.*, using large-scale image datasets like ImageNet [157] to pre-train the spatial networks.

*2) **3D Models**:* Compared with 2D models, 3D models normally treat a sequence of frames as a whole and apply 3D convolutional neural networks or cube-based transformers to simultaneously capture spatial and temporal information.

3D CNN for action recognition [158] is a pioneer work proposed by S. Ji *et al.*, which is composed of a hardwired layer, two 3D convolutional layers, two subsampling layers, one 2D convolutional layer and a fully-connected layer. Though the proposed network is relatively small and only evaluated on small datasets, this work presents a prototype of 3D CNNs for action recognition and achieves better performance than using 2D CNNs.

Later, in 2015, D. Tran *et al.* [126] design a modern and deep 3D architecture – C3D for large-scale action recognition, where eight 3D convolutional layers with $3 \times 3 \times 3$ kernel size are adopted. C3D obtains 61.1% accuracy on **Sports 1M** [58], which is relatively competitive. Likewise, J. Carreira and A. Zisserman [7] propose a *Inflated 3D CNN* (I3D), where a 2D kernel with $N \times N$ size is expanded into a $N \times N \times N$ 3D kernel and the parameters of 3D kernels are also from pre-trained 2D kernels via bootstrapping. Compared with C3D, I3D is much deeper, stacking 9 3D inception modules [159] and 4 individual 3D convolutional layers. With these modern designs, I3D obtains much better performances on multiple datasets, *e.g.*, 95.6% vs. 82.3% on **UCF101** [1].

Directly expanding $N \times N$ 2D convolution into $N \times N \times N$ 3D convolution can significantly increase the number of parameters, improving the capacity of deep models but also raising computational complexity and the risk of overfitting. To mitigate the problem, Z. Qiu *et al.* [127] propose a *Pseudo 3D* (P3D) network, where 3D convolution is substituted by stacking a 2D convolution and a 1D convolution. Similarly, D. Tran *et al.* [128] explores different architectures (2D, 3D and (2+1)D), finding that stacking a 2D convolution with $1 \times N \times N$ kernel size and a $t \times 1 \times 1$ 1D convolution is superior to other architectures. While S3D [129] replaces part of 3D inception modules in I3D [7] with 2D inception modules to balance the performance and computational complexity. Later, D. Tran *et al.* [130] propose a set of architectures, termed – *3D Channel-Separated Networks* (CSN), to further reduce FLOPs, where group convolution, depth convolution and different combinations of them are explored. CSN achieves much better performance than 3D CNNs with only one-third FLOPs of 3D CNNs.

SlowFast [12] is composed of two branches – one is the slow branch with a low frame rate and another is the fast branch with a high frame rate. The slow branch with a low frame rate can pay more attention to spatial semantics, while the fast branch pays more attention to object motion. To achieve this, the network of the slow branch is designed only using 2D convolution in the bottom layers and using (1+2)D convolution in the top layers, whereas the fast branch uses (1+2)D convolution in each layer. Note that the fast branch is designed to capture object motion instead of high-level semantics, thus it can be a lightweight neural network. In addition, SlowFast adopts lateral connections to fuse slow and fast features. With elaborate designs of a slow branch, fast branch and lateral connections, SlowFast achieves state-of-the-art performance on several popular action recognition datasets.

To model long video sequences, S. Zhang [145] introduces *Temporal Fully Connected Operation* into SlowFast, proposing TFCNet, where the features of all frames are combined by an FC layer. With a simple operation, TFCNet boosts the performance on **Diving48** to 88.3%, nearly 11% higher than that achieved by SlowFast.

STM [132] adopts two modules – *Channel-wise Spatial-Temporal Module* (CSTM) and *Channel-wise Motion Module* (CMM), where CSTM employs (2+1)D convolution to fuse spatial and temporal features, while CMM only uses 2D convolution but concatenates the features of three successive frames. Compared with P3D [127] and R3D [128], STM performs better.

X3D [133] expand 2D CNNs in four manners – space, time, depth and width, which explores a number of architectures, finding that high spatial-temporal networks are superior to other models. X3D is inferior to SlowFast on **Kinetics-400** (79.1% vs. 79.8%), but X3D has fewer parameters and takes less time during training and inference. To further reduce the number of parameters and FLOPs, D. Kondratyuk *et al.* [136] propose *Mobile Video Networks* (MoViNets) that are able to process streaming videos. Tow core techniques are applied in MoViNets – the first one is *Neural Architecture Search* (NAS) [160] for efficient architectures generation and the second one is stream buffer technique that equips 3D CNNs to tackle streaming videos with arbitrary length. With these two techniques, MoViNets only requires 20% FLOPs of X3D, but achieves better performance.

SlowFast [12] shows that introducing different temporal resolutions benefits action recognition, however, it applies an individual network to each resolution, which is time-consuming. In contrast, TPN [134] applies one backbone network and uses a temporal pyramid to 3D features in different levels, *i.e.*, low frame rate for the high-level features to capture semantics and high frame rate in low-level features to capture motion information. TPN achieves the same performance on **Kinetics-400** but only adopts one branch.

After 2020, the number of transformers using 3D modules is rising. Compared with 2D transformer-based models like TimeSformer [122] which separately uses spatial and temporal self-attention, 3D transformer-based models execute self-attention over non-overlap cubes, which is more similar to 3D convolution. ViViT [14] expand ViT into video action recognition via using tubelet embedding. Also, ViViT explores different architectures of transformers – spatial-temporal transformer, factorised encoder, factorised self-attention and fac-

torised dot-product, finding that spatial-temporal transformer performs the best on large datasets but overfits small datasets and needs much more FLOPs than other architectures since spatial-temporal transformer executes self-attention over all tokens with a computational complexity of N_t^2, where N_t^2 denotes the number of tokens.

MViT [136] mimic the multi-scale architectures of CNNs, introducing multi-head pooling attention into ViT [154], *i.e.*, high resolution for low-level features and low resolution for high-level features. In terms of action recognition, 3D pooling attention is applied. Though MViT executes self-attention over all spatial-temporal tokens, the number of tokens drops when it goes deeper and the dimension of token embedding is low in shallow layers, hence, the FLOPs of MViT are around 1/5 of ViViT FLOPs. Compared with ViViT, MViT with fewer parameters and less computational cost achieves similar performance on **Kinetics-400**.

Similar to MViT, *Video Swin Transformer* (ViSwin) [138] uses different resolutions in different levels, but it only reduces the spatial resolution in each level and keeps the temporal resolution. One important property of ViSwin is using 3D shifted window based self-attention, which reduces the computational complexity and increases the receptive field via stacking multiple layers. Finally, ViSwin-large achieves 84.9% accuracy on **Kinetics-400** with ImageNet-21K pre-trained parameters and a high spatial resolution (384×384).

Mformer [137] also uses cuboid embedding like MViT [136] and ViViT [14], but it applies separate space and time positional encoding like TimeSformer [122]. The key difference between Mformer and ViViT is the trajectory attention module. Different from joint space-time attention and divided space-time attention, trajectory attention models the probabilistic path of a token among frames, where the similarity between each pair of tokens is calculated, but self-attention is performed along the time dimension to compute trajectories. Trajectory attention has the same computational complexity as joint space-time attention, *i.e.*, quadratic complexity in both space and time, taking more time than divided space-time attention. To speed up the calculation, Mformer introduces an approximation approach, achieving 81.1% accuracy on **Kinetics-400**.

As we have mentioned above, transformer-based models normally split frames into 2D non-overlap patches or 3D non-overlap cubes, thus, the objects in videos could be divided into different patches or cubes. missing object-centric information. ORViT, short for *Object-Region Vision Transformer* [139] introduces object-dynamic module and object-region attention into vision transformers. In the object-dynamic module, object bounding box coordinates are encoded using the box position encoder, while in the object-region attention module, object representations obtained by RoIAlign [161] are employed to generate key and value vectors. With these two modules, ORViT pays more attention to objects and achieves 88% accuracy on **Diving48**, 8% higher than the baseline. Though introducing object features can benefit the model to capture more semantics, it requires multi-object tracking to obtain the bounding boxes of objects.

Similar to *Masked Language Models* (MLM) [144], researchers also develop a number of masked video models. BEVT [140] expands BEIT [162] to video domain. Briefly, BEVT predicts the representations of masked patches, where the presentations are obtained by VQ-VAE [163]. Likewise, VIMPAC [142] predicts patch representations obtained by VQ-VAE, but uses a 24-layer BERT-like backbone instead of ViSwin [138] and applies contrastive learning during training – discriminating positive video clip pairs from negative ones. Though VIMPAC employs both patch representation prediction and contrastive learning, it is inferior to BEVT and one possible reason is that ViSwin is more powerful and the parameters of the image Swin are shared with ViSwin, hence, it can well model spatial information. Alternatively, MaskFeat [141] employs MViT [136] as the backbone and explores predicting the features of the masked patches obtained by different approaches, such as HOG, VQ-VAE and DINO [164], finding that predicting HOG is slightly worse than using DINO but DINO requires a pre-trained model.

Through the numbers in Table III, we can make the conclusion that 3D models are normally superior to 2D models, but 3D models could be time-consuming and cost more computational resources. Also, we can find that the pre-train-fine-tune paradigm is increasingly popular for 3D models, in particular for 3D transformer-based models since it is straightforward to introduce the tricks of MLM into video models.

3) ***Two-stream Models****:* Two-stream models normally take RGB frames and optical flow as input and each stream employs a deep neural network (see Fig. 4). RGB frames provide both spatial and temporal information, while optical flow mainly provides information on motion. Obviously, we can easily expand the above 2D/3D models that only take RGB frames as input into two-stream models, resulting in their two-stream variants, such as TSN-Two-Stream [120], TSM-Two-Stream [13] and TRN-Two-Stream [147]. Compared with their one-stream versions that only use video frames, two-stream models achieve better performance but require calculating optical flow first and an additional neural network to obtain deep representations of motion.

Another problem with two-stream models is how to combine the representations of frames and optical flow. An early work Two-Stream ConvNet [10] proposed by K. Simonyan *et al.*directly averages the prediction of each stream, while C. Feichtenhofer *et al.* [146] explores different fusing approaches, including max-pooling, concatenation, bilinear, sum and convolution in different layers of the two-stream networks.

Recently, researchers observe that some advanced one-stream models outperform their two-stream counterparts since tow-stream networks have higher capacity, easily overfitting the dataset. In addition, the generalizability of using video frames and optical flow are different, so training a two-stream network with one strategy is sub-optimal. W. Wang *et al.* [148] endeavors to address the issues, proposing *Gradient Blending* (G-Blend) where the weights of different loss functions are estimated during training, hence, it assigns a weight to each stream.

4) ***Skeleton-based Models****:* 2D, 3D and two-stream deep models take RGB frames as input, while skeleton-based models take players' skeleton graphs as input (see Fig. 4).

Normally, *Graph Convolutional Networks* (GCN) [165] are used to model the skeleton graph composed of joints.

S. Yan *et al.* [149] propose a *Spatial-Temporal GCN* (ST-GCN) for action recognition, which is similar to 3D convolutional networks but executed on skeleton graph, achieving 30.7% accuracy on **Kinetics-400**. Compared with frame based models like 2D and 3D models, the performance of ST-GCN is much worse since it cannot capture the appearance information, however, convolution on graphs is much faster.

AGCN [150] introduces an attention mechanism into GCN. Three types of attention are employed in AGCN – spatial attention, temporal attention and channel attention. With these types of attention, AGCN achieves higher accuracy scores. Similarly, C. Si *et al.* [166] propose an *Attention Enhanced Graph Convolutional LSTM Network* (AGC-LSTM), where the temporal information is captured using an LSTM and the spatial information is captured using a GCN with attention.

Y. Song *et al.*improve GCNs with a bag of advanced techniques, such as batch normalization [167], yielding an EfficientGCN [151] that achieves competitive performance on **FSD-10**, but takes less time for training and is more explainable.

The topology of graphs is crucial for action recognition and Y. Chen *et al.*propose a *Channel-wise Topology Refinement GCN* (CTR-GCN) [152] to effectively model the topology. Specifically, CTR-GCN employs a channel-wise topology modeling block to compute the channel-wise correlation and then models the relationship among graph nodes in different channels. Finally, CTR-GCN achieves 66.2% accuracy on **FSD-10**, better than ST-GCN and AGCN.

The drawback of using skeleton graphs composed of joints is that we need to detect the joints first and normally the predicted graphs are noisy, leading to worse performance on existing datasets. Alternatively, PoseC3D [153] applies the heatmaps of joints and limbs instead of graphs, which are more robust than directly using skeleton graphs. Pose3D treats the heatmaps as frames, hence, traditional 3D convolutional networks can be adopted. Through Table III, we can find that Pose3D is superior to other skeleton-based models, but still inferior to two-stream models.

As we have mentioned above, skeleton-based models require detecting the joints first, resulting in extra computation cost and prediction noise. Though using heatmaps can mitigate the problem of noise, the performance is still worse than other types of models.

5) ***Others****:* In addition to 2D, 3D, two-stream and skeleton-based models, hybrid models that are composed of multiple model types are also applied for video action recognition. One recent work – *Temporal Query Networks* (TQN) [171] combines 3D CNNs and transformers. Specifically, 3D CNNs are used as the backbone to extract video features and transformers are adopted as decoders, *i.e.*, given a query, the transformers output a response, where the queries are texts like `the number of flips` for diving and the responses are the corresponding attributes, such as a number or a label. The transformer-based decoder models the relevance among visual features, queries and responses. In terms of fine-grained action recognition, TQN requires pre-defined action labels and each label has a set of attributes for classification, hence, we can classify the actions based on the responses. Compared with its 3D counterparts, TQN shows its superiority, achieving 89.6% on **FineGym** and 81.8% on **Diving48**.

TABLE IV
CURRENT STATE OF INDIVIDUAL SPORTS VIDEO ACTION RECOGNITION. HERE WE ONLY LIST THE PERFORMANCE ON THE SPORTS-RELATED DATASETS NOT IN TABLE III.

Datasets	Models	Years	Performance
Tennis			
ACASVA [56]	HOG3D+CNN [98]	2020	93.78
THETIS [57]	Lightweight 3D [168]	2022	90.9
TenniSet [65]	Two-stream [65]	2017	81.0
Table tennis			
TTStroke-21 [72]	Two-stream [25]	2020	91.4
SPIN [74]	Multi-stream [74]	2019	72.8
Stroke Recogn. [85]	TCN [169]	2021	99.37
Badminton			
Badm. Olymp. [73]	TCN [169]	2018	71.49
Basketball			
NCAA [63]	CNN+LSTM [63]	2016	51.6
FineBasketball [82]	TSN-Two-Stream [120]	2020	29.78
NPUBasketball [87]	Skeleton-based [87]	2020	80.9
Football			
SoccerNet [24]	3D [24]	2018	65.2
Others			
Hockey Fight [55]	Two-stream [170]	2017	97.0
GolfDB [75]	CNN+LSTM [75]	2019	79.2
FenceNet [91]	TCN [169]	2022	87.6

Note that videos are composed of not only frames but also audio, and they are family of models that adopt multiple modalities. Similar to two-stream models, multimodal models consist of several branches. One recent work is AudioSlowFast [172] proposed by F. Xiao *et al.*, where acoustic information is introduced into the original SlowFast [12] model using an audio branch, hence, AudioSlwoFast has 3 branches – slow, fast and audio. While Y. Bian *et al.* [173] propose an ensemble model that adopts video frames, optical flow and audio. In our developed toolbox [3], we also adopt acoustic information to classify football actions, where there are 8 categories, such as red card, corner and free kick. Using multiple modalities is able to improve the capacity of deep models and the redundant information could make the model more robust, however, it is difficult to combine different modalities and training multimodal models is non-trivial [148]. In addition, using more branches leads to large models, so overfitting can easily occur.

In Table IV, we present current state of action recognition in different types of sports. We can see that 3D and two-stream models are relatively popular and the recent advanced models like MoViNet [136] are rarely used in sports. One possible reason is that some sports-related datasets lack challenges and two-stream models can achieve high accuracy, *e.g.*, 91.4% on **TTStroke-21** [72]. While some other datasets like **NCAA** [63] and **FineBasketball** [82] are still challenging, requiring more advanced models.

Note that many models can be applied to both common action recognition and sports action recognition, however, the performance of some methods is unsatisfying, *e.g.*, skeleton-based models perform much worse than 3D and 2D models

[3]https://github.com/PaddlePaddle/PaddleVideo

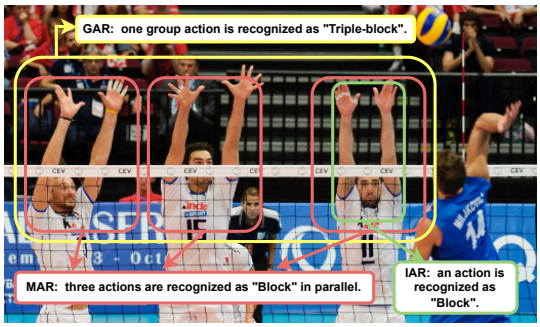


Fig. 5. An example of individual, group, and multi-player activity recognition in a frame of a volleyball competition video.

on FineGym dataset, since it is difficult to detect players and estimate their poses due to motion blur (see reference [79] for more details). In this case, we need to design more robust pose estimation methods to tackle the problem of motion blur, in particular for the frames with intense actions. Also, for broadcasting videos with multiple views and shot transitions, we need specialized methods to handle these problems.

IV. Group/Team Activity Recognition

Group/team activity recognition is one branch of the human activity recognition problem which targets the collective behavior of a group of people, resulting from the individual actions of the persons and their interactions. It is a basic task for automatic human behavior analysis in many areas, such as **sports**, health care and surveillance. Note that, although group/team activity is conceptually an activity performed by multiple people or objects, the group/team activity recognition (GAR) is quite different from another common task – the multi-player activity recognition (MAR) [174]. The former is the process of recognizing the activities of multiple players, where a single group activity is a function of the action of each and every player within the group [175]. The activity of a group can be observed as spontaneous emergent action, conducted by the activities and interactions of individuals within it. While the latter is the recognition of separate actions of multiple players in parallel, where two or more players participates. Figure 5 shows the differences among individual action recognition (IAR), GAR, and MAR respectively. The GAR example (yellow box) shows that where without knowledge of all of the players on the opposite of the net, it is improbable that the algorithm will infer the accurate actions (e.g., if one of the players does not participate in the blocking, the activity is "double-block" indeed). Only observing all subjects provides enough evidence for the correct recognition. Therefore, GAR is more challenging than individual action recognition, requiring a combination of multiple computer vision techniques, such as player detection, pose estimation and ball tracking. Fig. 6 presents a typical framework for GAR.

Basically, individual networks are various and we can use 2D, 3D or skeleton-based models to extract individual features, whereas two types of group networks dominate this field: LSTM-based and graph-based models. An early work on group activity recognition is proposed by W. Choi *et al.*in 2009 [176].

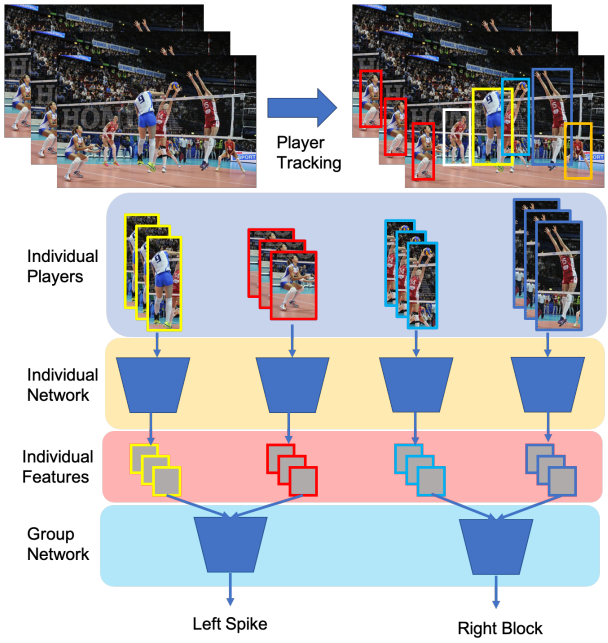


Fig. 6. A typical framework for group activity recognition (GAR). Compared with models for individual action recognition shown in Fig. 4, GAR models normally require player tracking, individual player feature extraction and group feature combination, which is more complicated.

The proposed framework is composed of people detection, tracking, pose estimation, spatial-temporal local descriptor and classifier, where hand-crafted features – HOG is adopted. Though there is not a group network in this model and it is only tested on a private dataset, it inspires the following approaches.

A. *LSTM-based Models*

M. Ibrahim *et al.* [61] proposed a hierarchical deep model for GAR, where each player is detected first and the dynamics of each player are modeled using an LSTM, finally, then a group-level LSTM is adopted to aggregate all players' dynamics and makes a prediction. The hierarchical deep model achieves 51.1% on **HierVolleyball** dataset and 81.9% on **HierVolleyball-v2**.

Interestingly, though T. Shu *et al.* [177] use a graph to model group activities and propose a *Confidence-Energy Recurrent Network* (CERN), LSTMs are applied to perform message passing. Specifically, CERN first employs a tracker to obtain the trajectories of players and then constructs a graph, where each node represents an individual player position in a video frame and each edge represents the relationship between two nodes. Two types of LSTMs are applied – node LSTM and edge LSTM to compute deep features of graph nodes and edges. CERN achieves 83.6% on **HierVolleyball-v2**.

Similarly, T. Bagautdinov *et al.* [178] proposed an end-to-end approach for GAR, where player detection and action recognition adopt a shared fully-connected CNN. The detection branch applies *Markov Random Field* (MRF) to refine the predicted player positions and the classification branch uses a matching *Recurrent Neural Network* (RNN) to predict individual's action and their group activity. Without extra tracking models, the proposed model takes less time

TABLE V
DEEP LEARNING MODEL FOR GROUP ACTIVITY RECOGNITION IN SPORTS.

Model	Venue	HierVolleyball-v2
M. Ibrahim *et al.* [61]	CVPR-2016	81.9
CERN [177]	CVPR-2017	83.6
T. Bagautdinov *et al.* [178]	CVPR-2017	87.1
RCRG [23]	ECCV-2018	89.5
StageNet [29]	TCSVT	89.3
POGARS [179]	Arxiv	93.9
Anchor-Transformer [180]	CVPR-2020	94.4
DIN [181]	CVPR-2021	93.1
Pose3D [153]	CVPR-2022	91.3

for training and inference. In terms of performance, it obtains 87.1% accuracy on **HierVolleyball-v2**.

StageNet [29] is composed of 4 stages: player detection, semantic graph construction, temporal information integration and spatial-temporal attention. Player detection and semantic graph construction are similar to RCRG [23], *i.e.*, each node of the graph represents a player position and the edges represent the relationships determined by the spatial distance and temporal correlations among players. In terms of temporal information integration, structural RNNs – node RNN and edge RNN are applied and finally the aggregated information is fed into spatial-temporal module. Using the spatial-temporal attention makes StageNet more explainable.

B. *Graph-based Models*

A. Maksai *et al.* [3] propose an approach to model the interaction between players and the ball for GAR. The proposed approach employs graphical models to track the ball and detect players, resulting in a player graph and a ball graph. In the player graph, each node represents a play location. With massage passing over the two graphs, the proposed approach can model the interaction between the ball and players. However, the main purpose of this work is ball tracking and the settings of GAR lack challenge, *e.g.*, there are only 4 classes of the ball state – flying, passed, possessed and out of play.

RCRG [23] extend the two-stage framework in [177], [178] via introducing a hierarchical relational network to replace LSTMs, which is similar to graph neural networks, *i.e.*, the new representation of a node is obtained by aggregating the information of its neighbors.

H. Yuan *et al.* [181] introduces dynamic relation (DR) and dynamic walk (DW) into GAR models, proposing a *Dynamic Inference Network* (DIN), where the detected players are constructed into a spatial-temporal graph and then DR is used to predict the relationships among players and DW is used to predict the dynamic walk offset to allow global interaction over the entire spatial-temporal graph. Using DR and DW, DIN obtains 93.1% on **HierVolleyball-v2**.

C. *Others*

Recently, the poses of players are introduced into GAR. H. Thilakarathne *et al.* [179] propose a *Pose Only Group Activity Recognition System* (POGARS), which consists of two key modules – player tracking and pose estimation and each player is represented by 16 2D keypoints. After that, POGARS stacks multiple temporal and spatial convolutional layers to obtain high-level player representations. In addition, POGARS investigates different person-level fusion approaches, including early fusion and late fusion. Finally, POGARS achieves 93.2% accuracy on **HierVolleyball-v2** and the performance can be further improved to 93.9% by using both player poses and the ball tracklets. While Pose3D [153] adopts skeleton heatmaps instead of the 2D coordinates and the feature extraction model is a 3D CNN, achieving 91.3% accuracy. A more advanced model – GIRN is presented in [182]. Similar to POGARS, GIRN first estimates the poses of players, but it introduces 3 relational modules to model intra-person, inter-person and person-object relationships, *i.e.*, message passing is conducted among the joints of the same person and different persons. It achieves 92.2% accuracy on **HierVolleyball-v2** by using attention mechanism.

K. Gavrilyuk *et al.* [180] propose a transformer based model – Anchor-Transformer, where the representations of different players are fused via a transformer instead of an LSTM. Similarly, Anchor-Transformer first employs a player detection model to obtain the individuals and then fuses the individual embeddings using a transformer for classification. It achieves 94.4% on **HierVolleyball-v2** using both pose and optical flow.

Apart from volleyball, GAR in football is also investigated. T. Tsunoda *et al.* [64] propose a hierarchical LSTM model to recognize football team activities, which is similar to the model in [61], but the videos in the football dataset are captured by multiple synchronized cameras.

Also, we present the performances of different models in Table V. Note that most models conduct experiments on **HierVolleyball-v2**, thus, we only report the performance on this dataset. And the proposed models are flexible and can be transferred into other team sports like football and basketball.

V. APPLICATIONS

As aforementioned, video action recognition in sports spawns a wide sort of applications in our daily life. We categorize the applications into the following aspects.

Training Aids. Since the sports video corpus contains a large number of historical records of competition and training clips, it is a good source of information for sports coaches and players to analyze and extract useful tactics. As one of the most common approaches, video action recognition can provide a straightforward way to obtain the actions/events (i.e., the basic unit of sports). Then, the action sequences/combinations could be correlated with the winning strategies, which can either guide the training of players or help with designing the game plan. *E.g.*, [183] introduces an action recognition hourglass network (ARHN) to interpret player's actions in ice hockey videos, where the recognized hockey players' poses and hockey actions are valuable pieces of information that potentially can help coaches in assessing player's performance. Another well-known case for training aid is the sports AI coach system [184], which can provide personalized athletic training experiences based on video sequences. Action recognition is one of the key steps in the AI coach system to support complex visual information extraction and summarization.

Game Assistance (Video Judge). The video-based game judge has been widely involved in modern sports video analysis systems, where most of the systems adopt action recognition as the elementary module. [185] proposes a virtual referring network to evaluate the execution of a diving performance. This assessment is based on visual clues as well as the body actions in sequences. Upon the same sports (diving), [77] comes up with an idea to learn spatio-temporal features that explain the related tasks such as fine-grained action recognition, so as to improve the action quality assessment. Rather than judge the performance of the athlete via action recognition, [186] develops a sports referee training system, which intends to recognize whether a trainee makes the proper judging signals. In this work, a deep belief network is adopted to capture high-quality features for hand gesture recognition.

Video Highlights. Highlights segmentation and summarization in sports videos are with a wide viewership and a great amount of commercial potential. While the foundation for accomplishing this goal is the action recognition step in processing the sports video. As a typical example, [187] proposes an automatic highlight detection method to recognize the spatio-temporal pose in skating videos. Through an accurate action recognition module, the proposed method is capable of locating and stitching the target figure skating poses. Since the jumps in figure skating sports are one of the most eye-catching actions/poses, it appears commonly in the highlight clips of figure skating sports, where [188] dedicates to recognizing the 3D jump actions and recovering the poor-visualising actions. Another work [189] treats the video highlights as a combinatorial optimization problem, and regards the diversity of recognized action as one of the constraints. To maximize the diversity and lower the recognition error, the overall quality of the highlights video is improved drastically.

Automatic Sports News Generation (ASNG). There is a large demand for sports news generation. Existing ASNG systems normally adopt the statistical numbers in matches, such as the number of shots, corners and free kicks in a football match and then use texts to describe the numbers [190], [191]. However, in many cases, the numbers are provided by a human instead of automatically recognized in videos, which is time-consuming and a massive workload. While video action recognition techniques can automatically generate these numbers and only require a few people to verify the final results, saving time and reducing workload. Plus, thanks to the technique of visual captioning, *i.e.*, using texts to describe images [192]–[194] and videos [195], [196], we can also directly generate textural descriptions from videos. Nevertheless, recognizing the actions of players is still required, since better recognition results can significantly improve the naturalness, fluency and accuracy of the final texts.

General Research Purposes. As one of the main branches of video analysis, action recognition is never stopped being studied. We can observe that the sports videos account for a significant portion of the target video categories [197]–[201]. Not surprisingly, sports video analysis has been a very popular research topic, due to the variety of application areas, ranging from analysis of athletes' performances and rehabilitation to multimedia intelligent devices with user-tailored digests. Datasets (videos) [48]–[56], [56], [57] focused on sports activities or datasets including a large amount of sports activity classes are now available and many research contributions benchmark on those datasets. A large amount of work is also devoted to fine-grained action recognition through the analysis of sports gestures/poses using motion capture systems. On the other hand, the ability to analyze the actions which occur in a video is essential for the automatic understanding of sports. The action recognition techniques can efficiently collect and classify the actions/events in sports videos, and consequently help a lot with the sports statistics analysis which is the basis to understand the sports [47], [202]–[206].

All in all, the application of video action recognition in sports is widely spread in different purposes and draws more attention from either sports domains or computer vision domains. In the next section, we will go through the possible challenges when applying action recognition in realistic sports videos.

VI. Challenges and Future Work

In this section, we summarize the challenges when applying those action recognition baselines on sports videos in practice. **Data Collection and Annotation**. As one of the crucial steps for establishing a dataset for further research, data collection and annotation draw more attention and their qualities directly affect the performance of the action recognition task [7], [207], [208]. However, the main difference between sports datasets compared to other human action recognition datasets (e.g., ActivityNet, Kinetics400, and UCF101) in terms of collections and annotations are 1) Accessibility: Most of the representative sports videos come from untrimmed live broadcasting clips, which is access-restricted due to the authorship or the copyright of the clips. While the self-recorded sports videos are of comparably lower quality either in footage resolution (without the best angle) or the content itself (e.g., the target players are amateurish), such datasets can lead to the inefficient training of the action recognition algorithms, which generates models with poor generalization ability in the practical task; 2) Expertise: Since the sports videos normally focus on specific sports category (e.g., hockey, volleyball, and figure skating), the annotation requires higher expertise than the regular human actions (e.g., walk, run, and sit). The more professional the annotators are especially in the target sports domain, the better the quality of the annotations is, which leads to the promising performance of action recognition algorithms in real inference tasks. One possible direction is using active learning approaches [209]–[211] to reduce the workload of annotation; 3) Multi-purpose: As a general trend, the video dataset for actions recognition is rarely with only one purpose, so are sports datasets. Some of the video datasets [212], [213] also are designed to accomplish temporal action localization, spatio-temporal action localization, and complex event understanding. To serve multiple purposes, the author of the dataset needs to prepare a variety of labeling content and auxiliary feature information, which is even more challenging for sports videos due to the specific nature of the

actions. *E.g.*, extracting the skeleton feature from a table tennis video is difficult due to the dense and fast-moving nature of the stroke actions. Compared to the general human actions recognition datasets, sports action recognition datasets usually take more effort to be established and developed.

Dense and Fast-moving Actions. On the one hand, the traditional action recognition baselines [13], [119]–[122] are designed to tackle those actions around $4 \sim 20$ (or over 20s as an event) seconds on average, where some of the actions in sports video are out of this range. *E.g.*, the stroke action in a table tennis competition commonly tasks only $0.4 \sim 2$ seconds via a conventional broadcasting camera. Fast-moving characteristics require the action recognition algorithms to capture relatively short-lived events from the video stream and tolerate the background changes which is easy to confuse the judgment in such a scenario [214], [215]. On the other hand, as the nature of table tennis sports itself, two players take action to stoke the ball in turns until one of the players wins a point, where the stoke actions are in a super dense distribution compared to other sports (e.g., soccer and basketball). There could be 8 to 10 stroke actions in less than 6 seconds, which means the action recognition algorithms should be more sensitive to the boundary of two actions and it is proved to be a challenging task for some of the state-of-the-art models [58], [116]–[118]. Although we can fine-tune the baselines carefully on the video datasets with dense and fast-moving actions, the performance is still far less than expectation [131], [216] compared to those regular action recognition tasks. Thus, sports with fast-moving and dense actions are the potential to be further explored in the action recognition domain and could be a basis for developing more robust recognition algorithms.

Camera Motion, Cut, Occlusion and Low Quality. The main difference between video datasets and still image datasets are the motion of the target object, where the quality of the motion features may affect the action recognition performance [217]–[219]. The traditional way to form motion trajectories heavily relies on the extraction of optical flow [220], [221], where most of them are based on the video recorded by the fixed camera with the complete and clear view of objects. However, in recent sports videos/streaming, the camera motion is no longer fixed and tends to be variant since the highlights of the video keep changing (e.g., the zoom-in and zoom-out highlights). This naturally leads to the cut of view and more or less occlusion in the recorded videos/streaming, which causes challenges to those well-established action recognition benchmarks [2], [10], [13], [120], [128], [146]–[148] (e.g., those algorithms are barely tolerable to the data sample from different camera motions, with cut and occluded objects). Although there exists work [103], [104], [222] to take the camera motion into consideration when designing the motion descriptor for action recognition task, the cut and occluded objects are still a problem which makes the feature space inconsistent. Several works [223]–[225] intend to solve the occlusion problem individually by modifying the structure and attention of the motion descriptor, where it is limited to a single target and we know that sports videos commonly involve multiple players, which increases the complexity when applying these occlusion-handling methods. Another challenge is low-quality videos, *e.g.*, low-resolution [226] video action recognition. Though video super-resolution is able to alleviate the low-resolution problem, biases are introduced by super-resolution models.

Long-tailed Distribution and Imbalanced Data. Before applying action recognition algorithms on the video datasets, we normally check the statistics of the dataset in case of any undesirable situation such as the long-tailed distribution of the target actions. As we know, the long-tailed learning [227]–[229] is one of the most challenging problems in visual recognition, aims to train well-performing models from a large number of frames that follow a long-tailed class distribution. Unfortunately, sports datasets such as soccer, basketball, and table tennis suffer a lot from such long-tailed class distribution and imbalance, which degrades the model performance drastically [230]–[233]. This common status quo in sports video datasets motivates us to either adopt a proper data augmentation method prior to training or design a robust action recognition algorithm to mitigate the negative effects of long-tailed distribution. As shown in Figure 7, we briefly compare the distribution of classes in general video recognition versus the distribution in long-tailed video recognition. Further we showcase two representative datasets, which are table tennis videos (P^2A [93] dataset) and the sports video in wild (SVW [93] dataset). The middle and bottom figures demonstrate the class of action in untrimmed sports videos commonly follow a long-tailed distribution and naturally form imbalanced datasets.

Multi-camera and Multi-view Action Recognition. As we mentioned in the Applications section, action recognition techniques are widely used in web or TV streaming for the purpose of Video Highlights. While the videos are normally recorded via multiple cameras and are in different views [2], [49], [58], [65], [70], [79], [93], this requires the robustness and adaptability of the corresponding action recognition algorithms. Via a thorough investigation in this paper, most of the benchmarks [13], [58], [116]–[122] of action recognition on video datasets focus on single-camera or single-view actions, where it does not conform with the format of sports videos. Although some of the action recognition algorithms [234]–[236] intend to split the task into several sub-tasks (i.e., training separately on each view) and combine the results for a performance promotion, it is still challenging to detect and switch the sub-models between each view when handling a complete sports video.

Transfer, Few-shot and Zero-shot Learning. To ensure the accuracy of action recognition, there frequently needs to collect a large number of video clips, extract frames from clips, and annotate frames with fine-grained labels (such as temporal labels and/or skeletons). The data collection and annotation thus become extremely expensive, when sports of multiple categories are desired. Yet another way to lower the cost of action recognition from sports videos is to pre-train backbone models using videos collected from a wide spectrum of sports categories in a self-supervised manner [237]–[239] and then fine-tune [240]–[243] the pre-trained model using few labeled samples for the target sport analytic tasks, so as to transfer the knowledge of video understanding to specific

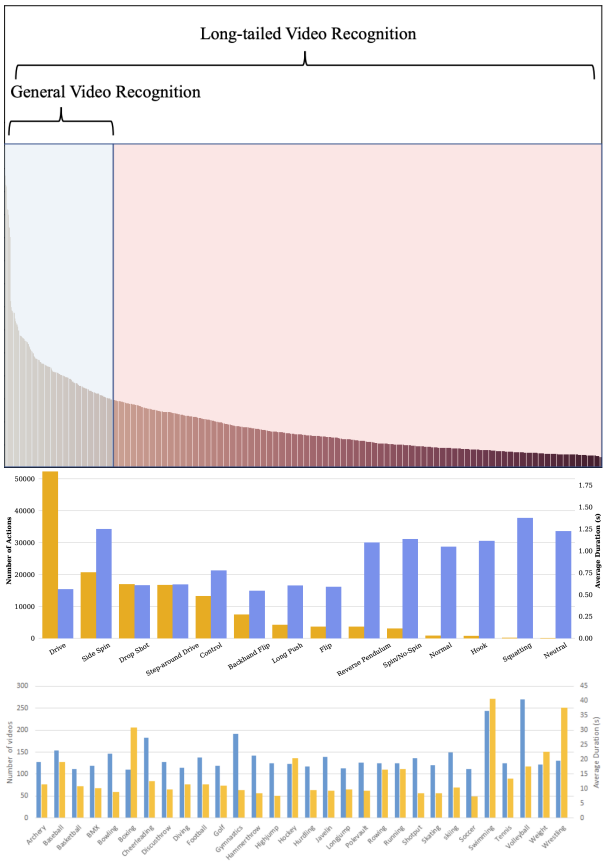


Fig. 7. Example of Long-tailed Distribution and Imbalanced Data. Top: Long-tailed vs General [230]; Middle: The long-tailed distribution of classes in P^2A dataset [93]; Bottom: The imbalanced classes in SVW dataset [60].

sport action recognition tasks. Thus, few-shot and even zero-shot learning [244]–[247] are requested to generalize action recognition tasks by incorporating labeled samples and/or explicit domain knowledge [248].

Note that common action recognition could share the same challenges as sports action recognition. Some problems are more significant in traditional action recognition, such as camera motion and low-quality videos since many datasets are constructed based on online and self-recorded videos, the quality of which varies. While sports datasets are normally based on broadcasting videos recorded by professional equipment. On the other hand, there are some relatively significant challenges for sports action recognition like dense action and shot transition, which are more common in sports.

To address the above challenges, we think the following possible directions should be considered in the future. First, to tackle dense and fast-moving actions, two kinds of solutions can be considered. On one hand, we can use high-speed cameras to capture the motions, which is able to provide much more detail and mitigate the problem of motion blur. However, it takes much longer time and more computational resources to process the videos recorded by high-speed cameras. Alternatively, we can introduce other types of data, such as gyroscope data to augment video action recognition. On the other hand, models which are more robust on motion blur should be developed. As mentioned in reference [79], it is difficult to detect players and estimate their poses for the frames with intense motions.

Second, for the problem of camera motion, cut and occlusion, which are relatively common in sports videos, in particular in broadcasting videos, we think specialized modules should be proposed to detect camera motion and shot transition. It is difficult to avoid occlusion, in particular for team sports, but using multiple cameras is able to alleviate the problem. In this case, we need to develop models which is able to combine the information of different views. In addition, if the parameters of cameras are provided, 3D vision techniques can be applied to benefit action recognition and scene understanding [249], [250]. In addition, calibration-free approaches draw much attention in recent years and these techniques can be applied to multi-view action recognition without camera calibration [251], [252].

In terms of transfer, few-shot and zero-shot learning. It is believed that the pre-train and fine-tune paradigm are able to handle this challenge and some models like MaskFeat [141] have employed it, achieving much better performance (see table III for more details). Thanks to the development of large-scale datasets, distributed parallel computing and big models [253]–[257], we can achieve powerful models using pre-train and fine-tune paradigm to tackle the problems of few-shot, zero-shot and long-tailed distribution. However, most big models are developed for natural language processing and in the field of sports analytics, it lacks both specific big models and large-scale datasets. It should be promising to develop big models for sports analytics.

Another promising direction should be action detection in long and untrimmed videos. Many existing works concentrate on recognising actions in trimmed videos, while sports videos are normally long.There are some works, such as BSN [258], BMN [259], TCANet [260], MLAD [261] and STALE [262] focusing on action detection. BSN, BMN and TCANet employ CNNs to extract video features, while MLDA designs two attention blocks to model action dependencies and STALE uses vision transformers and textual descriptions for zero-shot action detection. Another crucial difference among these models is that, BSN, BMN and TCANet are two-stage models, *i.e.*, they first localize the actions and then trim the videos for classification, introducing localization error in the 2nd stage, whereas STALE is a one-stage model, which can be trained in an end-to-end manner. One-stage models with fast training and satisfying performance should be considered for analyzing untrimmed sports videos. Also, R. Modi *et al.* [263] points out the limitations of existing datasets and methods for action detection, which can inspire future work.

VII. Conclusion

In this paper, we review and survey the works on video action recognition for sports analytics. We cover dozens of sports, categorized into two streams *(1) team sports*, such as football, basketball, volleyball, hockey and *(2) individual sports*, such as figure skating, gymnastics, table tennis, tennis, diving and badminton. Specifically, we present numerous existing solutions, such as statistical learning-based methods for traditional computer visions, deep learning-based methods

with 2D and 3D neural models, and skeleton-based methods using auxiliary information, all for sports video analytics. We compare the performance of these methods using literature reviews and experiments, where we clearly illustrate the status quo on the performance of video action recognition for both team sports and individual sports. Finally, we discuss the open issues, including technical challenges and interesting problems, in this area and conclude the survey. To facilitate the research in this field, we release a toolbox for sport video analytics for public research.

ACKNOWLEDGEMENT

This work was support in part by the National Key R&D Program of China (No. 2021ZD0110303), and the Humanities and Social Science Research Grant (No. 20YJA890024) from Ministry of Education of China.

REFERENCES

[1] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[2] S. Liu, X. Liu, G. Huang, L. Feng, L. Hu, D. Jiang, A. Zhang, Y. Liu, and H. Qiao, "Fsd-10: a dataset for competitive sports content analysis," *arXiv preprint arXiv:2002.03312*, 2020.

[3] A. Maksai, X. Wang, and P. Fua, "What players do with the ball: A physically constrained interaction modeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 972–981.

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[5] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[6] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[7] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[8] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056.

[9] 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.

[10] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[11] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[12] 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.

[13] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[14] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[15] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "Tea: Temporal excitation and aggregation for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 909–918.

[16] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[17] B. Giles, S. Kovalchik, and M. Reid, "A machine learning approach for automatic detection and classification of changes of direction from player tracking data in professional tennis," *Journal of sports sciences*, vol. 38, no. 1, pp. 106–113, 2020.

[18] 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.

[19] D. Hendry, K. Chai, A. Campbell, L. Hopper, P. O'Sullivan, and L. Straker, "Development of a human activity recognition system for ballet tasks," *Sports medicine-open*, vol. 6, no. 1, pp. 1–10, 2020.

[20] C. Pickering and J. Kiely, "The development of a personalised training framework: Implementation of emerging technologies for performance," *Journal of Functional Morphology and Kinesiology*, vol. 4, no. 2, p. 25, 2019.

[21] B. Russell, A. McDaid, W. Toscano, and P. Hume, "Moving the lab into the mountains: A pilot study of human activity recognition in unstructured environments," *Sensors*, vol. 21, no. 2, p. 654, 2021.

[22] K. Rangasamy, M. A. As' ari, N. A. Rahmad, N. F. Ghazali, and S. Ismail, "Deep learning in sport video analysis: a review," *Telkomnika*, vol. 18, no. 4, pp. 1926–1933, 2020.

[23] M. S. Ibrahim and G. Mori, "Hierarchical relational networks for group activity recognition and retrieval," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 721–736.

[24] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1711–1721.

[25] P.-E. Martin, J. Benois-Pineau, R. Péteri, and J. Morlier, "Fine grained sport action recognition with twin spatio-temporal convolutional neural networks," *Multimedia Tools and Applications*, vol. 79, no. 27, pp. 20 429–20 447, 2020.

[26] P.-E. Martin, J. Calandre, B. Mansencal, J. Benois-Pineau, R. Péteri, L. Mascarilla, and J. Morlier, "Sports video: Fine-grained action detection and classification of table tennis strokes from videos for mediaeval 2021," *arXiv preprint arXiv:2112.11384*, 2021.

[27] A. Cioppa, A. Deliege, S. Giancola, B. Ghanem, M. V. Droogenbroeck, R. Gade, and T. B. Moeslund, "A context-aware loss function for action spotting in soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 126–13 136.

[28] Y. Li, L. Chen, R. He, Z. Wang, G. Wu, and L. Wang, "Multisports: A multi-person video dataset of spatio-temporally localized sports actions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 536–13 545.

[29] M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. Van Gool, "stagnet: an attentive semantic rnn for group activity and individual action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 549–565, 2019.

[30] B. Mahaseni, E. R. M. Faizal, and R. G. Raj, "Spotting football events using two-stream convolutional neural network and dilated recurrent neural network," *IEEE Access*, vol. 9, pp. 61 929–61 942, 2021.

[31] G. Bertasius, H. Soo Park, S. X. Yu, and J. Shi, "Am i a baller? basketball performance assessment from first-person videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2177–2185.

[32] P. Sri-Iesaranusorn, F. C. Garcia, F. Tiausas, S. Wattanakriengkrai, K. Ikeda, and J. Yoshimoto, "Toward the perfect stroke: A multimodal approach for table tennis stroke evaluation," in *2021 Thirteenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*. IEEE, 2021, pp. 1–5.

[33] A. Tejero-de Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, "Summarization of user-generated sports video by using deep action recognition features," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2000–2011, 2018.

[34] P. Shukla, H. Sadana, A. Bansal, D. Verma, C. Elmadjian, B. Raman, and M. Turk, "Automatic cricket highlight generation using event-driven and excitement-based features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1800–1808.

[35] K. Zhao, T. Osogami, and T. Morimura, "Visual analytics for team-based invasion sports with significant events and markov reward process," *arXiv preprint arXiv:1907.01221*, 2019.

[36] C. Yan, X. Li, and G. Li, "A new action recognition framework for video highlights summarization in sporting events," in *2021 16th*

International Conference on Computer Science & Education (ICCSE). IEEE, 2021, pp. 653–666.

[37] H. Kajbafnezhad, H. Ahadi, A. R. Heidarie, P. Askari, and M. Enayati, "Difference between team and individual sports with respect to psychological skills, overall emotional intelligence and athletic success motivation in shiraz city athletes," *Journal of Physical Education and Sport*, vol. 11, no. 3, p. 249, 2011.

[38] J.-F. Grehaigne, P. Godbout, and D. Bouthier, "Performance assessment in team sports," *Journal of teaching in Physical Education*, vol. 16, no. 4, pp. 500–516, 1997.

[39] M. B. Evans, M. A. Eys, and M. W. Bruner, "Seeing the "we" in "me" sports: The need to consider individual sport team environments." *Canadian Psychology/Psychologie Canadienne*, vol. 53, no. 4, p. 301, 2012.

[40] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, "A comprehensive study of deep video action recognition," *arXiv preprint arXiv:2012.06567*, 2020.

[41] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[42] N. A. Rahmad, M. A. As'Ari, N. F. Ghazali, N. Shahar, and N. A. J. Sufri, "A survey of video based action recognition in sports," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 3, pp. 987–993, 2018.

[43] J. Gudmundsson and M. Horton, "Spatio-temporal analysis of team sports," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–34, 2017.

[44] R. P. Bonidia, L. A. Rodrigues, A. P. Avila-Santos, D. S. Sanches, and J. D. Brancher, "Computational intelligence in sports: A systematic literature review," *Advances in Human-Computer Interaction*, vol. 2018, 2018.

[45] R. Beal, T. J. Norman, and S. D. Ramchurn, "Artificial intelligence for team sports: a survey," *The Knowledge Engineering Review*, vol. 34, 2019.

[46] M. Manafifard, H. Ebadi, and H. A. Moghaddam, "A survey on player tracking in soccer videos," *Computer Vision and Image Understanding*, vol. 159, pp. 19–46, 2017.

[47] H.-C. Shih, "A survey of content-aware video analysis for sports," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1212–1231, 2017.

[48] J. Pers, "Cvbase 06 dataset: a dataset for development and testing of computer vision based methods in sport environments," *SN, Ljubljana*, 2005.

[49] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[50] C. De Vleeschouwer, F. Chen, D. Delannay, C. Parisot, C. Chaudy, E. Martrou, A. Cavallaro *et al.*, "Distributed video acquisition and annotation for sport-event summarization," *NEM summit*, vol. 8, 2008.

[51] P. Parisot and C. De Vleeschouwer, "Consensus-based trajectory estimation for ball detection in calibrated cameras systems," *Journal of Real-Time Image Processing*, vol. 16, no. 5, pp. 1335–1350, 2019.

[52] T. D'Orazio, M. Leo, N. Mosca, P. Spagnolo, and P. L. Mazzeo, "A semi-automatic system for ground truth generation of soccer video sequences," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2009, pp. 559–564.

[53] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 9–14.

[54] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *European conference on computer vision*. Springer, 2010, pp. 392–405.

[55] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *International conference on Computer analysis of images and patterns*. Springer, 2011, pp. 332–339.

[56] T. De Campos, M. Barnard, K. Mikolajczyk, J. Kittler, F. Yan, W. Christmas, and D. Windridge, "An evaluation of bags-of-words and spatio-temporal shapes for action recognition," in *2011 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 2011, pp. 344–351.

[57] S. Gourgari, G. Goudelis, K. Karpouzis, and S. Kollias, "Thetis: Three dimensional tennis shots a human action dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 676–681.

[58] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[59] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *European Conference on Computer Vision*. Springer, 2014, pp. 556–571.

[60] S. M. Safdarnejad, X. Liu, L. Udpa, B. Andrus, J. Wood, and D. Craven, "Sports videos in the wild (svw): A video dataset for sports analysis," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–7.

[61] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1971–1980.

[62] ——, "Hierarchical deep temporal models for group activity recognition," *CoRR*, vol. abs/1607.02643, 2016. [Online]. Available: http://arxiv.org/abs/1607.02643

[63] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei, "Detecting events and key actors in multi-person videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3043–3053.

[64] T. Tsunoda, Y. Komori, M. Matsugu, and T. Harada, "Football action recognition using hierarchical lstm," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 99–107.

[65] H. Faulkner and A. Dick, "Tenniset: A dataset for dense fine-grained event recognition, localisation and description," in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2017, pp. 1–8.

[66] P. Parmar and B. Tran Morris, "Learning to score olympic events," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 20–28.

[67] K. Lu, J. Chen, J. J. Little, and H. He, "Light cascaded convolutional neural networks for accurate player detection," *arXiv preprint arXiv:1709.10230*, 2017.

[68] P. Parisot and C. De Vleeschouwer, "Scene-specific classifier for effective and efficient team sport players detection from a single calibrated camera," *Computer Vision and Image Understanding*, vol. 159, pp. 74–88, 2017.

[69] S. Francia, S. Calderara, and D. F. Lanzi, "Classificazione di azioni cestistiche mediante tecniche di deep learning," *URL: https://www.researchgate. net/publication/330534530_Classificazione_di_Azioni_Cestistiche_mediante_Tecniche_di_Deep_Learning*, 2018.

[70] Y. Li, Y. Li, and N. Vasconcelos, "Resound: Towards action recognition without representation bias," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 513–528.

[71] J. Yu, A. Lei, Z. Song, T. Wang, H. Cai, and N. Feng, "Comprehensive dataset of broadcast soccer videos," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2018, pp. 418–423.

[72] P.-E. Martin, J. Benois-Pineau, R. Péteri, and J. Morlier, "Sport action recognition with siamese spatio-temporal cnns: Application to table tennis," in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2018, pp. 1–6.

[73] A. Ghosh, S. Singh, and C. Jawahar, "Towards structured analysis of broadcast badminton videos," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 296–304.

[74] S. Schwarcz, P. Xu, D. D'Ambrosio, J. Kangaspunta, A. Angelova, H. Phan, and N. Jaitly, "Spin: A high speed, high resolution vision dataset for tracking and action recognition in ping pong," *arXiv preprint arXiv:1912.06640*, 2019.

[75] W. McNally, K. Vats, T. Pinto, C. Dulhanty, J. McPhee, and A. Wong, "Golfdb: A video database for golf swing sequencing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[76] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 1468–1476.

[77] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 304–313.

[78] R. Voeikov, N. Falaleev, and R. Baikulov, "Ttnet: Real-time temporal and spatial video analysis of table tennis," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 884–885.

[79] 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.

[80] N. Feng, Z. Song, J. Yu, Y.-P. P. Chen, Y. Zhao, Y. He, and T. Guan, "Sset: a dataset for shot segmentation, event detection, player tracking in soccer videos," *Multimedia Tools and Applications*, vol. 79, no. 39, pp. 28 971–28 992, 2020.

[81] Y. Jiang, K. Cui, L. Chen, C. Wang, and C. Xu, "Soccerdb: A large-scale database for comprehensive video understanding," in *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports*, 2020, pp. 1–8.

[82] X. Gu, X. Xue, and F. Wang, "Fine-grained action recognition on a novel basketball dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2563–2567.

[83] T. Zhao and S. Liu, "Fineskating: A high-quality figure skating dataset and multi-task approach for sport action," *Peng Cheng Laboratory Commumications*, vol. 1, no. 3, p. 107, 2020.

[84] S. Liu, A. Zhang, Y. Li, J. Zhou, L. Xu, Z. Dong, and R. Zhang, "Temporal segmentation of fine-grained semantic action: A motion-centered figure skating dataset," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2163–2171.

[85] K. M. Kulkarni and S. Shenoy, "Table tennis stroke recognition using two-dimensional human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4576–4584.

[86] K. Vats, P. Walters, M. Fani, D. A. Clausi, and J. Zelek, "Player tracking and identification in ice hockey," *arXiv preprint arXiv:2110.03090*, 2021.

[87] C. Ma, J. Fan, J. Yao, and T. Zhang, "Npu rgbd dataset and a feature-enhanced lstm-dgcn method for action recognition of basketball players," *Applied Sciences*, vol. 11, no. 10, p. 4426, 2021.

[88] A. Deliege, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. Van Droogenbroeck, "Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4508–4519.

[89] P. Parmar and B. Morris, "Win-fail action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 161–171.

[90] W.-Y. Wang, H.-H. Shuai, K.-S. Chang, and W.-C. Peng, "Shuttlenet: Position-aware fusion of rally progress and player styles for stroke forecasting in badminton," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[91] K. Zhu, A. Wong, and J. McPhee, "Fencenet: Fine-grained footwork recognition in fencing," *arXiv preprint arXiv:2204.09434*, 2022.

[92] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, and A. Zisserman, "The ava-kinetics localized human actions video dataset," *arXiv preprint arXiv:2005.00214*, 2020.

[93] J. Bian, Q. Wang, H. Xiong, J. Huang, C. Liu, X. Li, J. Cheng, J. Zhao, F. Lu, and D. Dou, "$p^2\alpha$: A dataset and benchmark for dense action detection from table tennis match broadcasting videos," *arXiv preprint arXiv:2207.12730*, 2022.

[94] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos "in the wild"," *Computer Vision and Image Understanding*, vol. 155, pp. 1–23, 2017.

[95] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei, "Every moment counts: Dense detailed labeling of actions in complex videos," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 375–389, 2018.

[96] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[97] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[98] E. P. Ijjina, "Action recognition in sports videos using stacked auto encoder and hog3d features," in *Proceedings of the Third International Conference on Computational Intelligence and Informatics*. Springer, 2020, pp. 849–856.

[99] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[100] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1234–1241.

[101] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision*. Springer, 2006, pp. 428–441.

[102] J. Perš, V. Sulić, M. Kristan, M. Perše, K. Polanec, and S. Kovačič, "Histograms of optical flow for efficient representation of body motion," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1369–1376, 2010.

[103] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[104] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[105] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[106] M.-y. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," 2009.

[107] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2, pp. 107–123, 2005.

[108] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2046–2053.

[109] A. Klaser, M. Marszałek, I. Laptev, and C. Schmid, "Will person detection help bag-of-features action recognition?" Ph.D. dissertation, INRIA, 2010.

[110] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *CVPR 2011*. IEEE, 2011, pp. 489–496.

[111] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Bmvc 2009-british machine vision conference*. BMVA Press, 2009, pp. 124–1.

[112] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[113] S.-F. Wong, T.-K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural information," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–6.

[114] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 32–36.

[115] T.-K. Kim, S.-F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[116] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[117] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[118] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*. PMLR, 2015, pp. 843–852.

[119] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei, "You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 923–932.

[120] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.

[121] X. Long, C. Gan, G. De Melo, J. Wu, X. Liu, and S. Wen, "Attention clusters: Purely attention based local feature integration for video

classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7834–7843.

[122] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding," *arXiv preprint arXiv:2102.05095*, vol. 2, no. 3, p. 4, 2021.

[123] Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, H. Chen, I. Marsic, and J. Tighe, "Vidtr: Video transformer without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 577–13 587.

[124] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3163–3172.

[125] J. Yang, X. Dong, L. Liu, C. Zhang, J. Shen, and D. Yu, "Recurring the transformer for video action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 063–14 073.

[126] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[127] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[128] 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.

[129] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 305–321.

[130] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5552–5561.

[131] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 046–12 055.

[132] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "Stm: Spatiotemporal and motion encoding for action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2000–2009.

[133] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 203–213.

[134] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 591–600.

[135] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6824–6835.

[136] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "Movinets: Mobile video networks for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 020–16 030.

[137] M. Patrick, D. Campbell, Y. Asano, I. Misra, F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques, "Keeping your eye on the ball: Trajectory attention in video transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 493–12 506, 2021.

[138] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," *arXiv preprint arXiv:2106.13230*, 2021.

[139] R. Herzig, E. Ben-Avraham, K. Mangalam, A. Bar, G. Chechik, A. Rohrbach, T. Darrell, and A. Globerson, "Object-region video transformers," *arXiv preprint arXiv:2110.06915*, 2021.

[140] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y.-G. Jiang, L. Zhou, and L. Yuan, "Bevt: Bert pretraining of video transformers," *arXiv preprint arXiv:2112.01529*, 2021.

[141] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," *arXiv preprint arXiv:2112.09133*, 2021.

[142] H. Tan, J. Lei, T. Wolf, and M. Bansal, "Vimpac: Video pre-training via masked token prediction and contrastive learning," *arXiv preprint arXiv:2106.11250*, 2021.

[143] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2630–2640.

[144] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[145] S. Zhang, "Tfcnet: Temporal fully connected networks for static unbiased temporal reasoning," *arXiv preprint arXiv:2203.05928*, 2022.

[146] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[147] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 803–818.

[148] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 695–12 705.

[149] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[150] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020.

[151] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1625–1633.

[152] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 359–13 368.

[153] H. Duan, Y. Zhao, K. Chen, D. Shao, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," *arXiv preprint arXiv:2104.13586*, 2021.

[154] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[155] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[156] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[157] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[158] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[159] S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, "An efficient deep residual-inception network for multimedia classification," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 373–378.

[160] G. Bender, H. Liu, B. Chen, G. Chu, S. Cheng, P.-J. Kindermans, and Q. V. Le, "Can weight sharing outperform random architecture search? an investigation with tunas," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 323–14 332.

[161] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[162] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[163] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[164] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[165] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[166] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1227–1236.

[167] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[168] T. E. Rasmussen, L. H. Clemmensen, and A. Baum, "Compressing cnn kernels for videos using tucker decompositions: Towards lightweight cnn applications," *arXiv preprint arXiv:2203.07033*, 2022.

[169] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[170] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violent interaction detection in video based on deep learning," in *Journal of physics: conference series*, vol. 844, no. 1. IOP Publishing, 2017, p. 012044.

[171] C. Zhang, A. Gupta, and A. Zisserman, "Temporal query networks for fine-grained video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4486–4496.

[172] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[173] Y. Bian, C. Gan, X. Liu, F. Li, X. Long, Y. Li, H. Qi, J. Zhou, S. Wen, and Y. Lin, "Revisiting the effectiveness of off-the-shelf temporal modeling approaches for large-scale video classification," *arXiv preprint arXiv:1708.03805*, 2017.

[174] D. Gordon, "Group activity recognition using wearable sensing devices," 2014.

[175] C. Direkoğlu and N. E. O'Connor, "Team activity recognition in sports," in *European Conference on Computer Vision*. Springer, 2012, pp. 69–83.

[176] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*. IEEE, 2009, pp. 1282–1289.

[177] T. Shu, S. Todorovic, and S.-C. Zhu, "Cern: confidence-energy recurrent network for group activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5523–5531.

[178] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4315–4324.

[179] H. Thilakarathne, A. Nibali, Z. He, and S. Morgan, "Pose is all you need: The pose only group activity recognition system (pogars)," *arXiv preprint arXiv:2108.04186*, 2021.

[180] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. Snoek, "Actor-transformers for group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 839–848.

[181] H. Yuan, D. Ni, and M. Wang, "Spatio-temporal dynamic inference network for group activity recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7476–7485.

[182] M. Perez, J. Liu, and A. C. Kot, "Skeleton-based relational reasoning for group activity analysis," *Pattern Recognition*, vol. 122, p. 108360, 2022.

[183] M. Fani, H. Neher, D. A. Clausi, A. Wong, and J. Zelek, "Hockey action recognition via integrated stacked hourglass network," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 29–37.

[184] J. Wang, K. Qiu, H. Peng, J. Fu, and J. Zhu, "Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 374–382.

[185] M. Nekoui, F. O. T. Cruz, and L. Cheng, "Falcons: Fast learner-grader for contorted poses in sports," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 900–901.

[186] T.-Y. Pan, W.-L. Tsai, C.-Y. Chang, C.-W. Yeh, and M.-C. Hu, "A hierarchical hand gesture recognition framework for sports referee training-based emg and accelerometer sensors," *IEEE Transactions on Cybernetics*, 2020.

[187] T. Nakano, A. Sakata, and A. Kishimoto, "Estimating blink probability for highlight detection in figure skating videos," *arXiv preprint arXiv:2007.01089*, 2020.

[188] L. Tian, X. Cheng, M. Honda, and T. Ikenaga, "Multi-technology correction based 3d human pose estimation for jump analysis in figure skating," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 49, no. 1, 2020, p. 95.

[189] N. Shroff, P. Turaga, and R. Chellappa, "Video précis: Highlighting diverse aspects of videos," *IEEE Transactions on Multimedia*, vol. 12, no. 8, pp. 853–868, 2010.

[190] J. Kanerva, S. Rönnqvist, R. Kekki, T. Salakoski, and F. Ginter, "Template-free data-to-text generation of finnish sports news," *arXiv preprint arXiv:1910.01863*, 2019.

[191] J. Gong, W. Ren, and P. Zhang, "An automatic generation method of sports news based on knowledge rules," in *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. IEEE, 2017, pp. 499–502.

[192] Q. Wang, J. Wang, A. B. Chan, S. Huang, H. Xiong, X. Li, and D. Dou, "Neighbours matter: Image captioning with similar images," in *31st British Machine Vision Virtual Conference (BMVC 2020)*. British Machine Vision Association, BMVA, 2020.

[193] Q. Wang, J. Wan, and A. B. Chan, "On diversity in image captioning: Metrics and methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[194] J. Wang, W. Xu, Q. Wang, and A. B. Chan, "On distinctive image captioning via comparing and reweighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[195] S. Chen and Y.-G. Jiang, "Motion guided region message passing for video captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1543–1552.

[196] Z. Zhang, Z. Qi, C. Yuan, Y. Shan, B. Li, Y. Deng, and W. Hu, "Open-book video captioning with retrieve-copy-generate network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9837–9846.

[197] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human action recognition with video data: research and evaluation challenges," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 5, pp. 650–663, 2014.

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

[199] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2021.

[200] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, pp. 1–36, 2022.

[201] D. Wu, N. Sharma, and M. Blumenstein, "Recent advances in video-based human action recognition using deep learning: A review," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2865–2872.

[202] X.-H. Meng, H.-Y. Shi, and W.-H. Shang, "Analysis of basketball technical movements based on human-computer interaction with deep learning," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[203] H. Li, S. G. Ali, J. Zhang, B. Sheng, P. Li, Y. Jung, J. Wang, P. Yang, P. Lu, K. Muhammad *et al.*, "Video-based table tennis tracking and trajectory prediction using spatial-temporal cnns based on deep learning," *Fractals*, 2022.

[204] F. Carson, "Utilizing video to facilitate reflective practice: Developing sports coaches," *International Journal of Sports Science & Coaching*, vol. 3, no. 3, pp. 381–390, 2008.

[205] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer vision in sports*. Springer, 2014, pp. 181–208.

[206] G. Liu, D. Zhang, and H. Li, "Research on action recognition of player in broadcast sports video," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 9, no. 10, pp. 297–306, 2014.

[207] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5479–5490, 2016.

[208] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "Rgb-d-based action recognition datasets: A survey," *Pattern Recognition*, vol. 60, pp. 86–105, 2016.

[209] X. Zhan, Q. Wang, K.-h. Huang, H. Xiong, D. Dou, and A. B. Chan, "A comparative survey of deep active learning," *arXiv preprint arXiv:2203.13450*, 2022.

[210] X. Zhan, Q. Li, and A. B. Chan, "Multiple-criteria based active learning with fixed-size determinantal point processes," *arXiv preprint arXiv:2107.01622*, 2021.

[211] X. Zhan, H. Liu, Q. Li, and A. B. Chan, "A comparative survey: Benchmarking for pool-based active learning," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021.

[212] P. Chen, C. Gan, G. Shen, W. Huang, R. Zeng, and M. Tan, "Relation attention for temporal action localization," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2723–2733, 2019.

[213] S. Megrhi, M. Jmal, W. Souidene, and A. Beghdadi, "Spatio-temporal action localization and detection for human action recognition in big dataset," *Journal of visual communication and image representation*, vol. 41, pp. 375–390, 2016.

[214] Z. Hao, Q. Zhang, E. Ezquierdo, and N. Sang, "Human action recognition by fast dense trajectories," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 377–380.

[215] K. Anuradha and N. Sairam, "Spatio-temporal based approaches for human action recognition in static and dynamic background: a survey," *Indian Journal of Science and Technology*, vol. 9, no. 5, pp. 1–12, 2016.

[216] G. Lorre, J. Rabarisoa, A. Orcesi, S. Ainouz, and S. Canu, "Temporal contrastive pretraining for video action recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 662–670.

[217] L. Wang, Y. Qiao, X. Tang *et al.*, "Action recognition and detection by combining motion and appearance features," *THUMOS14 Action Recognition Challenge*, vol. 1, no. 2, p. 2, 2014.

[218] M. Lee, S. Lee, S. Son, G. Park, and N. Kwak, "Motion feature network: Fixed motion filter for action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 387–403.

[219] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[220] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," in *German conference on pattern recognition*. Springer, 2018, pp. 281–297.

[221] A. Piergiovanni and M. S. Ryoo, "Representation flow for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9945–9953.

[222] M. Jain, H. Jégou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2555–2562.

[223] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *European Conference on Computer Vision*. Springer, 2010, pp. 635–648.

[224] F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, "2d pose-based real-time human action recognition with occlusion-handling," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1433–1446, 2019.

[225] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view human action recognition under occlusion based on fuzzy distances and neural networks," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 1129–1133.

[226] U. Demir, Y. S. Rawat, and M. Shah, "Tinyvirat: Low-resolution video action recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 7387–7394.

[227] W. Ouyang, X. Wang, C. Zhang, and X. Yang, "Factors in finetuning deep model for object detection with long-tail distribution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 864–873.

[228] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5409–5418.

[229] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *arXiv preprint arXiv:2110.04596*, 2021.

[230] X. Zhang, Z. Wu, Z. Weng, H. Fu, J. Chen, Y.-G. Jiang, and L. S. Davis, "Videolt: Large-scale long-tailed video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7960–7969.

[231] K. Sozykin, S. Protasov, A. Khan, R. Hussain, and J. Lee, "Multi-label class-imbalanced action recognition in hockey videos via 3d convolutional neural networks," in *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. IEEE, 2018, pp. 146–151.

[232] W. Ding, D.-Y. Huang, Z. Chen, X. Yu, and W. Lin, "Facial action recognition using very deep networks for highly imbalanced class distribution," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1368–1372.

[233] D. Wu, Z. Wang, Y. Chen, and H. Zhao, "Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset," *Neurocomputing*, vol. 190, pp. 35–49, 2016.

[234] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, "Generative multi-view human action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6212–6221.

[235] D. Wang, W. Ouyang, W. Li, and D. Xu, "Dividing and aggregating network for multi-view action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 451–467.

[236] T. Hao, D. Wu, Q. Wang, and J.-S. Sun, "Multi-view representation learning for multi-view action recognition," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 453–460, 2017.

[237] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki, "Self-supervised learning of motion capture," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[238] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1134–1141.

[239] K. Hu, J. Shao, Y. Liu, B. Raj, M. Savvides, and Z. Shen, "Contrast and order representations for video self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7939–7949.

[240] X. Li, H. Xiong, H. Wang, Y. Rao, L. Liu, and J. Huan, "Delta: Deep learning transfer using feature map with attention for convolutional networks," in *International Conference on Learning Representations*, 2019.

[241] R. Wan, H. Xiong, X. Li, Z. Zhu, and J. Huan, "Towards making deep transfer learning never hurt," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 578–587.

[242] X. Li, H. Xiong, H. An, C.-Z. Xu, and D. Dou, "Rifle: Backpropagation in depth for deep transfer learning through re-initializing the fully-connected layer," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6010–6019.

[243] H. Xiong, R. Wan, J. Zhao, Z. Chen, X. Li, Z. Zhu, and J. Huan, "Grod: Deep learning with gradients orthogonal decomposition for knowledge transfer, distillation, and adversarial training," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2022.

[244] X. Xu, T. Hospedales, and S. Gong, "Semantic embedding space for zero-shot action recognition," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 63–67.

[245] A. Mishra, V. K. Verma, M. S. K. Reddy, S. Arulkumar, P. Rai, and A. Mittal, "A generative approach to zero-shot and few-shot action recognition," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 372–380.

[246] Y. Bo, Y. Lu, and W. He, "Few-shot learning of video action recognition only based on video contents," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 595–604.

[247] H. Zhang, L. Zhang, X. Qi, H. Li, P. H. Torr, and P. Koniusz, "Few-shot action recognition with permutation-invariant attention," in *European Conference on Computer Vision*. Springer, 2020, pp. 525–542.

[248] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.

[249] J. Zhang, Y. Cai, S. Yan, J. Feng *et al.*, "Direct multi-view multi-person 3d pose estimation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 153–13 164, 2021.

[250] Q. Zhang and A. B. Chan, "3d crowd counting via geometric attention-guided multi-view fusion," *International Journal of Computer Vision*, pp. 1–17, 2022.

[251] V. Larsson, N. Zobernig, K. Taskin, and M. Pollefeys, "Calibration-free structure-from-motion with calibrated radial trifocal tensors," in *European Conference on Computer Vision*. Springer, 2020, pp. 382–399.

[252] Q. Zhang and A. B. Chan, "Calibration-free multi-view crowd counting." in *European Conference on Computer Vision*, 2022.

[253] S. Chen, W. Li, C. Chen, J. Gu, J. Chu, X. Tao, and Y. Guo, "Seal: A large-scale video dataset of multi-grained spatio-temporally action localization," *arXiv preprint arXiv:2204.02688*, 2022.

[254] Z. Ma, J. He, J. Qiu, H. Cao, Y. Wang, Z. Sun, L. Zheng, H. Wang, S. Tang, T. Zheng *et al.*, "Bagualu: targeting brain scale pretrained models with over 37 million cores," in *Proceedings of the 27th*

ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, 2022, pp. 192–204.

[255] Q. Liu and Y. Jiang, "Dive into big model training," *arXiv preprint arXiv:2207.11912*, 2022.

[256] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu *et al.*, "Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv preprint arXiv:2107.02137*, 2021.

[257] W. Fedus, J. Dean, and B. Zoph, "A review of sparse expert models in deep learning," *arXiv preprint arXiv:2209.01667*, 2022.

[258] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[259] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "Bmn: Boundary-matching network for temporal action proposal generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3889–3898.

[260] Z. Qing, H. Su, W. Gan, D. Wang, W. Wu, X. Wang, Y. Qiao, J. Yan, C. Gao, and N. Sang, "Temporal context aggregation network for temporal action proposal refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 485–494.

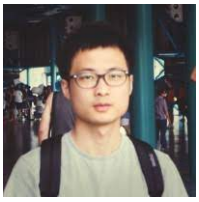
[261] P. Tirupattur, K. Duarte, Y. S. Rawat, and M. Shah, "Modeling multi-label action dependencies for temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1460–1470.

[262] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Zero-shot temporal action detection via vision-language prompting," *arXiv preprint arXiv:2207.08184*, 2022.

[263] R. Modi, A. J. Rana, A. Kumar, P. Tirupattur, S. Vyas, Y. Rawat, and M. Shah, "Video action detection: Analysing limitations and challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4911–4920.

Fei Wu is now a full professor in the Department of Physical Education, Peking University and also the associate head of the same department. She received the Ph.D degree from Peking University in 2018, and Bachelor's Degree and Master's Degree both in Physical Education and Training from Beijing Sport University in 2000 and 2003. She is a Chair Umpire officiating in several international table tennis games, including 2008 Beijing Olympics, 2012 London Olympics and 2020 Tokyo Olympics.

Qingzhong Wang (Member, IEEE) now is a researcher at Baidu Research. He received his Ph.D degree from City University of Hong Kong in 2021. Before that, he received his B.Eng and M.Eng from Harbin Engineering University in 2013 and 2016, respectively. His research interests include computer vision and vision-language learning.



Jiang Bian (Member, IEEE) is a researcher in Baidu Research. He received the Ph.D. degree at the University of Central Florida in 2020. In advance of stepping in doctoral research, he received the B.Eng degree in Logistics Systems Engineering from Huazhong University of Science and Technology, and earned his M.Sc degree in Industrial Systems Engineering from the University of Florida. His research interests include internet of things, sports analytics and ubiquitous computing.

Ning Ding received both Master's and Bachelor's Degrees from Shanghai University of Sport in 2016 and 2020 respectively. She is now pursuing her Master's degree in the Department of Physical Education, Peking University. As a former professional table tennis player, she was the winner of women's singles in World Championships (2011, 2015, 2017), World Cup (2011,2014, 2018), and received Silver Medal (2012) and Gold Medal (2016) from Olympics.

Feixiang Lu is currently a researcher at the Robotics and Autonomous Driving Lab, Baidu Inc. He obtained his Ph.D. degree in Computer Science from Beihang University in 2019. His research interests include 3D reconstruction, scene parsing, and their applications in sports. He received the Best Paper Award at the Computer Graphics International (CGI) in 2018.

Jun Cheng received his Master's and Bachelor's degrees both from Department of Computer Science and Technology, Zhejiang University, Hangzhou, China in 2004 and 2007 respectively. He is now a Staff R&D Engineer affiliated to Baidu, Inc. His research interests include deep learning, computer vision, and video analytics.

Dejing Dou (Senior Member, IEEE) is currently the Chief Data Scientist at Boston Consulting Group (Greater China), Beijing, China. From 2004 to 2022, he was an Assistant/Associate/Full Professor in the Computer and Information Science Department at the University of Oregon, Eugene OR. He was on leave at Big Data Lab, Baidu, Inc., Beijing China during 2019–2022. He received the Ph.D. degree in Artificial Intelligence in 2004 at Yale University and B.E. degree in Electronic Engineering in 1996 at Tsinghua University. His research interests include Artificial Intelligence, Data Mining, Data Integration, Information Extraction, Biomedical and Health Informatics.

Haoyi Xiong (Senior Member, IEEE) received Ph.D in computer science from Télécom SudParis and Université Pierre et Marie Curie, France in 2015. From 2016 to 2018, he was a Tenure-Track Assistant Professor with the Department of Computer Science, Missouri University of Science and Technology, Rolla MO, USA. Before that, he was a Postdoc at University of Virginia, Charlottesville VA from 2015 to 2016. He is currently the tech lead and principal architect at Big Data Lab, Baidu Inc., Beijing, China. His current research interests include AutoDL and ubiquitous computing. He has published more than 70 papers in top computer science conferences and journals, including UbiComp, ICML, ICLR, RTSS, KDD, AAAI, IJCAI, IEEE TMM, TMC, TC, TKDE, TNNLS, IOTJ, and ACM TKDD. He was a co-recipient of the 2020 IEEE TCSC Award for Excellence in Scalable Computing (Early Career Researcher) and the prestigious Science & Technology Advancement Award (First Prize) from Chinese Institute of Electronics in 2019.