Comprehensive Soccer Video Understanding: Towards Human-comparable Video Understanding System in Constrained Environment

Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, Chen Wang, Hui Liu, Changliang Xu
Xinhua Zhiyun Technology Co., Ltd.
{jiangyudong, cuikaixu, chenleilei, wangcanjin, wangchen, liuhui, xuchangliang}@shuwen.com

Abstract

Comprehensive video understanding, a challenging task in computer vision to understand videos like humans, has been explored in ways including object detection and tracking, action classification. However, most works for video understanding mainly focus on isolated aspects of video analysis, yet ignore the inner correlation between those tasks. Sports games videos can serve as a perfect research object of restrictive conditions, while complex and challenging enough to study the core problems of computer vision comprehensively. In this paper, we propose a new soccer video database named SoccerDB with the benchmark of object detection, action recognition, temporal action detection, and highlight detection. We further survey a collection of strong baselines on SoccerDB, which have demonstrated state-of-the-art performance on each independent task in recent years. We believe that the release of SoccerDB will tremendously advance researches of combining different tasks in closed form of the comprehensive video understanding problem. Our dataset and code will be published after the paper accepted.

1 Introduction

Variable-controlling, a critical method, is widely adopted by modern scientific research including artificial intelligence. Early in 1968, Terry Winograd developed a blocks world QA system named SHRDLU. The user can interact with the system on moving objects, naming collections and querying the state of the simplified world by English language [Russell and Norvig, 2016]. In modern AI research, there are so many surprising outcomes of such manner like IBM Watson QA system on television quiz to show Jeopardy!, AlphaGo on Chinese chess etc. In the field of video understanding, we lack such a challenging "game" like Jeopardy!, with clear rules and restrictive conditions, to conduct rigorous quantitative analysis on the factors we are interested in. In this paper, we choose soccer match as the research object, and build a dataset with multiple visual understanding tasks providing many different analysis aspect, to open the door of developing a human comparable video understanding system in soccer video analysis like IBM Watson.

1.1 Soccer Video Understanding

Soccer video understanding is a valuable topic not only in academic communities but also in industry world. European soccer market generates an annual revenue of $28.7 billion [Giancola et al., 2018]. Automatic soccer video analysis is valuable in soccer content production, which can help editors generate match summaries, visualize crucial players performance for tactical analysis, and so on. Some pioneering companies like GameFace1, SportLogiq2 adopt this technology on match statistics for analyzing strategies and players’ performance. However, automatic video analysis still cannot reach professional needs in some situation. The CEO of Wyscout3, claims that they employ 400 people on soccer data analytics, which taking over 8 hours to provide up to 2000 annotations per game.

1.2 Object Detection

Imagine, if you can only see players, ball and goal, could you still understand what happening in the match? Look at the left picture in Figure 1, maybe you can guess right: that's the moment of a shooting. Object detection, a fast-growing technology in last few years, gain human level performance in many applications like face detection, pedestrian detection. It is a foundational task in computer vision for dealing with localizing instances of semantic objects in images. As a very basic visual task, it can also provide extremely meaningful

1http://flixsense.com/soccer/
2https://sportlogiq.com/en/technology
3http://wyscout.cn/
knowledge for the other video understanding tasks like action recognition [Wang and Gupta, 2018] etc. For soccer video analysis, the detection system can help us recognize the position of ball, players, goal in the game. We produce superior attractive and vivid effect for football tactics visualization by using the position information as shown in Figure 2. Although the advance detection system can output reliable results from most of time, there are still many challenges when the object is small, fast moving, or blur. In this work, we construct a soccer game object detection dataset and benchmark two state-of-the-art detection models: RetinaNet [Lin et al., 2017], a “one-stage” detection algorithm, and Faster R-CNN [Ren et al., 2015], a “two-stage” detection algorithm.

### 1.3 Action Recognition

Action recognition, a core video understanding problem, has been made a lot of progress recently. Some large-scale dataset is published such as Kinetics [Carreira and Zisserman, 2017], Sports-1M [Karpathy et al., 2014], YouTube-8M [Abu-El-Haija et al., 2016]. Many state-of-the-art deep learning based algorithms like I3D [Carreira and Zisserman, 2017], Non-local Neural Networks [Wang et al., 2018], slowFast Network [Feichtenhofer et al., 2019], are proposed to this task. While supervised learning shows powerful ability on large scale recognition datasets, it failed when lacking of training data. In soccer game, some events such as goal, penalty kick, are rare. That gives rise to state-of-the-art recognition models cannot output convincing results. We hope this problem could be further investigated by joint considering multiple object relationships on the dataset.

### 1.4 Temporal Action Localization

Temporal action localization is a significant but harder problem than action recognition in video understanding, since it requires recognizing both action categories and the time boundary of the event at the same time. The definition of temporal boundary of event is ambiguous which leading to some famous database like Charades and MultiTHUMOS are not consistent between different human annotator [Sigurdsson et al., 2017]. That increase the difficulty of our dataset labeling. To overcome this ambiguous dilemma, we re-define the soccer events base on their actual meaning in soccer rules for clear emphasizes on the time boundaries. For example, we define "red/yellow card” as starting from the referee shows the card, and ending with resuming the game. The new definition helps us get more consistency action localization annotations.

### 1.5 Highlight Detection

The goal of highlight detection is to distill interesting content from a long video. Because of the subjectivity problem, constructing a highlight detection dataset usually needs multi-person label for the same video. That largely increases the cost and limit the scale of the dataset [Song et al., 2015]. We observe that in soccer TV broadcast, editors usually play the video segments containing brilliant events many times, which can be taken as an important clue for soccer video highlight detection. The SoccerDB provides this playback label. To our knowledge, this is the first work to detect highlight segments of match by playback clues, we begin the research with video classification models.

#### Contributions

- We introduce the new database and benchmark on comprehensive soccer video understanding. Four tasks, which are crucial to understanding videos, can be joint investigated in constraint environment by closed form.
- We provide strong baseline systems on each task, which are not only meaningful for academic research, but also can be applied to real world application.
- We discuss the inner connections between different visual understanding tasks based on SoccerDB.

### 2 Related Work

In this section, we focus on sports analytics, and the video understanding datasets. We discuss the progress on those areas recently which have close connection with our work.

#### 2.1 Sports Analytics

Automate sports analytics especially football, basketball, which are popular around the world, have been researched seriously since last few years in the computer vision community. Vignesh Ramanathan et al. bring a new automatically attention mechanism on RNN to identifying who is the key players of the event in basketball games [Ramanathan et al., 2016]. Silvio Giancola focus on temporal soccer events detection for finding highlight moments in soccer TV broadcast videos [Giancola et al., 2018]. Rajkumar Theagarajan et al. present an approach that generates visual analytics and player statistics for solving the talent identification problem in soccer match videos [Theagarajan et al., 2018]. The above works are only the tip of the iceberg among magnanimous research achievements in sports analytics area.

#### 2.2 Datasets

Lots of datasets are contributed to sports video understanding. Vignesh Ramanathan et al. provide 257 basketball games with 14K event annotations corresponding to 10 event classes for event classification and detection [Ramanathan et al., 2016]. Karpathy et al. collect one million sports videos from Youtube belonging to 487 classes of sports promoting...
Table 1: Bounding box statistics for object detection dataset. The scale of the bounding box, small, medium and large, following the definition of COCO dataset.

| Classes   | #Small | #Medium | #Large |
|-----------|--------|---------|--------|
| Player    | 196553 | 573816  | 12423  |
| Ball      | 52183  | 2399    | 82     |
| Goal      | 150    | 3091    | 12830  |
| Total     | 248886 | 579306  | 25335  |

depth learning research on action recognition greatly [Karpathy et al., 2014]. Datasets for video classification in the wild plays a vital role in research. Two famous large-scale datasets, Youtube-8M [Abu-El-Haija et al., 2016] and Kinetics [Carreira and Zisserman, 2017] are widely investigated inspiring most of the state-of-the-art methods in last few years. Google proposes AVA dataset for tackle dense activity understanding problem, which contains 57,600 clips of 3 seconds duration taken from featured films [Gu et al., 2018]. ActivityNet tackle general activity understanding by providing 849 video hours of 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video [Caba Heilbron et al., 2015]. Although ActivityNet considers video standing from multiple aspects including semantic ontology, trimmed and untrimmed video classification, spatial-temporal action localization, we argue that it stills too far away from general activity understanding as human being in unconstrained environment. Part of the source videos in our dataset is collected from SoccerNet [Giancola et al., 2018], a benchmark with a total of 6,637 temporal annotations on 500 complete soccer games from six main European leagues. A comparison between different databases is shown in Table 3.

3 Creating SoccerDB

This section introduces how do we construct the dataset, and give statistic, analysis and comparisons about it. The first section describes the setup of image object detection dataset. The rest of the following section only talk about video part of SoccerDB expect for special illustration.

3.1 Object Detection Dataset Collection

For image object detection task, we crawl around 55,290 images from internet. We label the bounding boxes of the player, the ball, and the goal. As shown in Table 1, the total number of bounding box labels are 853527, with 782792 player boxes, 54664 ball boxes, and 16071 goal boxes. We also calculate the scale of the boxes by COCO definition [Lin et al., 2014]. The dataset is divided into 49716 for training, and 5529 for testing.

3.2 Video Dataset Collection

We adopt 346 high qualities full soccer match videos, including 270 matches from SoccerNet [Giancola et al., 2018] covering six main European leagues ranging from 2014 to 2017 three seasons. 76 matches videos from China Football Association Super League in 2017 to 2018, and the 18th, 19th, 20th FIFA World Cup. The whole dataset consume 1.4 TB storages, with a total duration of 668.6 hours. We split the games into 226 for training, 63 for validation, and 57 for testing randomly.

3.3 Event Annotations

We define 10 different soccer events’ boundaries as clear as possible and annotate all of them densely in long soccer videos. The annotation system records the start/end time of an event, the categories of the event and if the event is playback. An annotator takes about three hours for labeling one match, and another experienced annotator review those annotations for ensuring the quality of outcomes.

3.4 Video Segments Processing

We split the whole soccer video into 3 to 30 seconds segments for easily processing. We make sure an event not be divided into two segments, and keep the event temporal boundary in a segment. The video without event is separated randomly into 145473 video clips with time duration from 3 to 20 seconds. All of the processed segments are checked again by human being for eliminating annotation mistakes. Some confused segments are discarded during this process. Finally, we get the total number of 25719 video segments with event annotation (core dataset), and 145473 background segments. There are 1.47 labels per segment on the core dataset.

3.5 Dataset Analysis

The detail of SoccerDB statistics is shown in Table 2. There are 14358 segments have “shot” label which achieves 38.07% among all events except for background, in contrast, we only collect 156 segments for “penalty kick”, 1160 for “red and yellow card”, which take the proportion of 0.41% and 3.07%. Since the dataset have extremely class imbalance problem, it is difficult for current state-of-the-art supervised methods of producing convincing results. We also statistic playback distribution on events which shows relevance to event type since 100% goal has playback, contrasting with 1.6% proportion of “substitution”. In section 5.4 we prove this relevance. As shown in section 2.2, we also provide comparisons of many aspects between other popular datasets and

| Events            | #Segments | Dur(min) | #Playback |
|-------------------|-----------|----------|-----------|
| Background        | 145473    | 25499.3  | 0         |
| Injured           | 1472      | 304.32   | 660       |
| Red/Yellow Card   | 1160      | 244.08   | 219       |
| Shot              | 14358     | 2125.35  | 8490      |
| Substitution      | 867       | 298.92   | 14        |
| Free Kick         | 3119      | 400.53   | 843       |
| Corner            | 3275      | 424.08   | 668       |
| Saves             | 5467      | 735.95   | 2517      |
| Penalty Kick      | 156       | 28.25    | 130       |
| Foul              | 5276      | 766.33   | 1015      |
| Goal              | 2559      | 532.03   | 2559      |
| Total             | 183182    | 31359.14 | 17115     |
ours. Our dataset supports more variety of tasks and more detail soccer class labels on constrained video understanding.

4 The Baseline System
To evaluate the capability of current video understanding technologies, and also to understand the challenge to the dataset, we develop many algorithms having produced strong performances on various datasets for each task which can provide strong baselines for future work to compare with. The whole picture of the baseline system are presented in Figure 3. In our baseline system, the action recognition submodule plays the essential role by providing basic visual representation to both temporal action detection and highlight detection tasks.

4.1 Object Detection
We adopt two representative object detection algorithms as baselines. One is Faster R-CNN, developed by Shaoqing Ren et al. [Ren et al., 2015]. The algorithm and its variant are widely used in many detection systems in recent years. Faster R-CNN belongs to the two-stage detector: The model use RPN proposes a set of regions of interests (RoI), then a classifier and a regressor only process the region candidates to get the category of the RoI and precise coordinate of bounding box. Another one is RetinaNet, which is well known as an one-stage detector. The authors Tsung-Yi Lin et al. discover that the extreme foreground-background class imbalance encountered is the central cause and introduced focal loss for solving this problem [Lin et al., 2017].

4.2 Action Recognition
The label of a video is encoded into binary vector corresponding to the number of classes. Cross entropy loss is used for each model. Two state-of-the-art action recognition algorithms are explored, the slowFast Networks and the Non-local Neural Networks. The slowFast networks contain two pathways: a slow pathway, simple with low frame rate, to capture spatial semantics, and a fast pathway, opposed to the slow pathway, operating at high frame rate, to capture motion pattern. We use ResNet-50 as the backbone of the network. The Non-local Neural Networks proposed by Xiaolong Wang et. al [Wang et al., 2018], that can capture long-range dependencies on video sequence. The non-local operator as a generic building blocks can be plugged into many deep architectures. We adopt I3D with ResNet-50 backbone, and insert non-local operators.

4.3 Temporal Action Detection
We explore temporal action detection by two-stage based method. First, a class-agnostic algorithm generates potential event proposals, then apply classifyng proposals approach for final temporal boundary localization. On the first stage, we utilize Boundary-Matching Network (BMN), a “bottom-up” temporal action proposal generation method, for generating high quality propose [Lin et al., 2019]. The BMN model is composed by three modules: (1) Base module processes the extracted feature sequence of the origin video, and output another video embedding shared by Temporal Evaluation Module (TEM) and Proposal Evaluation Module (PEM). (2) TEM evaluates starting and ending probabilities of each location in video to generate boundary probability sequences. (3) PEM transfers the feature to boundary-matching feature map which contains confidence scores of proposals. On the second stage, an action recognition models mentioned in section 4.2 predict the classification score of each top K proposals. We choose the highest prediction score of each class as the final detection result.

4.4 Highlight Detection
In this section, we formalize the highlight detection task as a binary classification problem, that recognizing which video is the "playback" video. We select SF-32 network (slowFast framework by sampling 32 frames per video segments) as the basic classifier, then we consider four scenarios:

- **Fully-connected only (fc-only) approach** involves extracting features from the final fc layer of a pre-trained model which is trained by action recognition task as shown in section 4.2. Then we train a logistic regressor for highlight detection. This approach evaluates the strength of the representation learned by action recogn-
Datasets | Context | #Video | #Instance | Dur(hrs) | #Classes | Support Tasks
---|---|---|---|---|---|---
YouTube-8M | General | 6100000 | 18300000 | 350000 | 3862 | Video Classification
Kinetics-600 | General | 495547 | 495547 | 1377 | 600 | Video Classification
AVA dataset | Movies | 57600 | 210000 | - | 80 | Video Classification
ActivityNet | General | 19994 | 30791 | 648 | 200 | Video Classification
Sports-1M | Sports | 1133158 | - | - | 487 | Video Classification
SoccerNet | Soccer | 1000 | 6637 | 764 | 4 | Video Classification
NCAA Basketball | Basketball | 257 | 14000 | 385.5 | 11 | Video Classification
Ours | Soccer | 171192 | 37709 | 668.6 | 11 | Object Detection, Video Classification, Temporal Detection, Highlight Detection

Table 3: The comparison of different datasets on video understanding. The background is taken as a class in classes number statistics.

- **Fully-finetuning (full-ft) approach** finetuning a binary classification network by initializing weights from action recognition model.
- **Multi-task (mt) approach** we train a multi-label classification network for both action recognition and highlight detection tasks. We adopt a per-label sigmoid output followed by logistic loss at the end of slowFast-32 network. This approach takes highlight segments as another action in action recognition framework. The advantage of this setting is forcing the network learning the relevance to different tasks, while the disadvantage is new label may introduce noise confusing the learning procedure.

- **Multi-task with highlight detection branch (mt-hl-branch) approach** we add a new two layers 3x3x3 convolution branch for playback recognition, which sharing the same backbone with the recognition task. We only train the highlight detection branch by freezing action recognition pre-trained model initialized parameters at first, then finetuning all parameters for multi-task learning.

The structures of the highlight detection models are presented in Figure 4.

5 Experiments

In this section, we focus on the performance of our baseline system on SoccerDB for object detection, action recognition, temporal action detection and highlight detection tasks.

5.1 Object Detection

We choose ResNeXt-101 with FPN as the backbone of both RetinaNet and Faster R-CNN. We train the models by 8 2080ti GPU, with the initial learning rate of 0.01 for RetinaNet, and 0.02 for Faster R-CNN. MS-COCO style evaluation method is applied for models’ benchmark. We present

| Methods   | small | medium | large | all  |
|------------|-------|--------|-------|------|
| RetinaNet  | 0.467 | 0.624  | 0.697 | 0.637 |
| Faster R-CNN | 0.450 | 0.618  | 0.672 | 0.630 |

Table 4: Object detection result with COCO evaluation tools. The metric is $AP_{0.5:0.95}$ on different scale.

| Methods   | mAP   | player | ball  | goal |
|------------|-------|--------|-------|------|
| RetinaNet  | 0.637 | 0.752  | 0.434 | 0.725 |
| Faster R-CNN | 0.630 | 0.756  | 0.416 | 0.717 |

Table 5: Classes $AP_{0.5:0.95}$ of both methods.

AP with IoU=0.5:0.95 and multi-scale in table 4, and also report the AP of each class as shown in table 5. RetinaNet performs slightly better than Faster R-CNN, and large-scale object is easier for both methods than small object. The ball detection result is lower than player and goal dual to the small scale and motion blur issue. All of the detection experiments are powered by mmdetection software [Chen et al., 2019].

5.2 Action Recognition

We setup the experiments by open source tool PySlowFast\(^5\), and boost all recognition network from Kinetics pre-training model. Since some labels are rare in dataset, we adjust the distribution of different labels appearing in training batch in order to balance the proportion of labels. We crop 224x224 pixels from a video, or its horizontal flip, with a shorter side in range 256 to 320 pixels randomly. On inference stage, we resize the frame to 256 shorter side with the origin image aspect ratio, and crop the long side to 256 by center symmetry. We compare 32 and 64 sample frame number for investigating the sample rate influence. For each class, the average precision (AP) scores are demonstrated on Table 6. Dense frame sample rate surpasses sparse sample rate for both methods. The classes with more instance like “shot” perform better than

\(^4\)http://cocodataset.org/#detection-eval
\(^5\)https://github.com/facebookresearch/SlowFast
In this section, we evaluate temporal action proposal generation and detection, and give quantized analysis on how does action recognition task affect temporal action localization. In practical, the action detection works after action recognition, only segments containing events should be processed. For a fair comparison of different action detection algorithms, we benchmark our baseline system on core dataset instead of a fix length D by zero padding or average pooling with D=100. To evaluate proposal quality, Average Recall (AR) under multiple IoU thresholds \{0.5:0.05:0.95\} is calculated. We report AR under different Average Number of proposals (AN) as AR@AN, and the area under the AR-AN curve (AUC) as ActivityNet-1.3 metrics, where AN is varied from 0 to 100. To show the different feature extractor influence on detection task, we compare two slowFast-32 pre-trained models, one is trained on SoccerDB action recognition task described in section 4.2, another is trained on Kinetics. Table 7 demonstrate the contrast results of those two extractors. The feature extractor trained on our dataset exceed Kinetics extractor by 0.7% on AUC metric. The results mean we gain some benefit from training feature encoder on the same dataset of temporal action proposal generation task, but the gain is limited. We use the same SF-32 classifier to produce the final detection results based on temporal proposals, and the detection metric is mAP with IoU thresholds \{0.3:0.1:0.7\}. For Kinetics proposals the mAP is 52.35%, ours mAP is 54.30%.

### 5.4 Highlight Detection

We set the experiments on whole SoccerDB dataset. The average precision results of our four baseline models are shown in Table 8. The fc-only model gets 63.14% AP demonstrates the action recognition model can provide strong representation to highlight detection task, and also indicate a close relationship between our defined events and the highlight segments. The mt model decrease the AP of the full-ft model by 2.33% which means highlight segments is very different from action recognition when sharing the same features. The mt-hl-branch model gives the highest AP by better utilizing the correlation of the two tasks while distinguish their difference. We also find the mt model is harmful to recognition classification task which decreases the mAP by 1.85 comparing to the slowFast-32 action recognition model. The mt-hl-branch can increase the action recognition mAP by 1.46%, when providing the highest highlight detection score. The detail action recognition mAP for the three models is shown in Table 9. Better way to utilize the connection between action recognition and highlight detection can boost both of the performances.

### 6 Conclusion

In this paper, we introduce SoccerDB, a new benchmark for comprehensive video understanding. It helps us discuss object detection, action recognition, temporal action detection, and video highlight detection in closed form with restricted but challenging soccer match environment. We explore many state-of-the-art methods on different tasks and discuss the relationship between those tasks. The quantized results show that there are very close connections between different visual understanding tasks, and algorithms can benefit a lot by considering the connections. We are releasing the benchmark

---

**Table 6**: Average Precision of different recognition models on each classes. SF-32: SlowFast Network with 32 sample rates. SF-64: SlowFast Network with 64 sample rates. NL-32: Non-local Network with 32 sample rates. NL-64: Non-local Network with 64 sample rates.

| Events          | SF-32   | NL-32   | SF-64   | NL-64   |
|-----------------|---------|---------|---------|---------|
| Background      | 0.9908  | 0.9916  | 0.9932  | 0.9926  |
| Injured         | 0.2303  | 0.3606  | 0.2256  | 0.3770  |
| R/Y Card        | 0.2862  | 0.3674  | 0.4662  | 0.4883  |
| Shot            | 0.8298  | 0.8532  | 0.8825  | 0.8517  |
| Substitution    | 0.9234  | 0.9060  | 0.9344  | 0.9030  |
| Free Kick       | 0.7333  | 0.7292  | 0.7734  | 0.7430  |
| Corner          | 0.9176  | 0.9182  | 0.9316  | 0.9192  |
| Saves           | 0.3891  | 0.4077  | 0.5224  | 0.4217  |
| Penalty Kick    | 0.6302  | 0.4851  | 0.7348  | 0.5336  |
| Foul            | 0.6475  | 0.6575  | 0.6778  | 0.6801  |
| Goal            | 0.3189  | 0.3192  | 0.4744  | 0.3994  |
| mAP             | 0.6270  | 0.6360  | 0.6924  | 0.6645  |

**Table 7**: Temporal action proposal AR@AN and AUC results.

| Extractor       | @1     | @10    | @50    | @100   | AUC    |
|-----------------|--------|--------|--------|--------|--------|
| Kinetics        | 0.5836 | 0.8335 | 0.8701 | 0.8821 | 0.8521 |
| Ours            | 0.6122 | 0.8401 | 0.8770 | 0.8882 | 0.8591 |

**Table 8**: Highlight detection models’ average precision results.

| Methods         | mAP    |
|-----------------|--------|
| fc-only         | 0.6314 |
| full-ft         | 0.7699 |
| mt              | 0.7465 |
| mt-hl-branch    | 0.7850 |

**Table 9**: Highlight detection multi-task learning models’ result on action recognition. SF-32 is the basic classification model described in section 4.2.
to video understanding community in the hope of driving researchers building towards human-comparable video understanding system.

References

[Abu-El-Haija et al., 2016] Sami Abu-El-Haija, Nisarg Kothari, Jooonseok Lee, Paul Natese, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675, 2016.

[Caba Heilbron et al., 2015] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 961–970, 2015.

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.

[Chen et al., 2019] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiаrui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155, 2019.

[Feichtenhofer et al., 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In Proceedings of the IEEE International Conference on Computer Vision, pages 6202–6211, 2019.

[Giancola et al., 2018] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 1711–1721, 2018.

[Gu et al., 2018] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul 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, pages 6047–6056, 2018.

[Karpathy et al., 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

[Lin et al., 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[Lin et al., 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017.

[Lin et al., 2019] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In Proceedings of the IEEE International Conference on Computer Vision, pages 3889–3898, 2019.

[Ramanathan et al., 2016] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorbán, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3043–3053, 2016.

[Ren et al., 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015.

[Russell and Norvig, 2016] Stuart J Russell and Peter Norvig. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited., 2016.

[Sigurdsson et al., 2017] Gunnar A Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In Proceedings of the IEEE International Conference on Computer Vision, pages 2137–2146, 2017.

[Song et al., 2015] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsun: Summarizing web videos using titles. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5179–5187, 2015.

[Theagarajan et al., 2018] Rajkumar Theagarajan, Federico Pala, Xiу Zhang, and Bir Bhanu. Soccer: Who has the ball? generating visual analytics and player statistics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 1749–1757, 2018.

[Wang and Gupta, 2018] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In Proceedings of the European Conference on Computer Vision (ECCV), pages 399–417, 2018.

[Wang et al., 2018] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7794–7803, 2018.