

Advanced Vision Techniques in Soccer Match Analysis: From Detection to Classification

Jakub Eichner¹^a, Jan Nowak^{1,2}^b, Bartłomiej Grzelak^{1,3}^c, Tomasz Górecki¹^d,
Tomasz Piłka¹^e and Krzysztof Dyczkowski¹^f

¹Adam Mickiewicz University, Poznań, Poland

²Poznan Supercomputing and Networking Center, Poznań, Poland

³KKS Lech Poznań, Poznań, Poland

Keywords: Computer Vision in Football, Object Detection and Tracking, Team Classification in Sports, Deep Learning for Sports Analytics, Football Match Analysis.

Abstract: This paper introduces an integrated pipeline for detecting, classifying, and tracking key objects within soccer match footage. Our research uses datasets from KKS Lech Poznań, SoccerDB, and SoccerNet, considering various stadium environments and technical conditions, such as equipment quality and recording clarity. These factors mirror the real-world scenarios encountered in competitions, training sessions, and observations. We assessed the effectiveness of cutting-edge object detection models, focusing on several R-CNN frameworks and the YOLOv8 methodology. Additionally, for assigning players to their respective teams, we compared the performance of the *K*-means algorithm with that of the Multi-Modal Vision Transformer CogVLM model. Despite challenges like suboptimal video resolution and fluctuating weather conditions, our proposed solutions have successfully demonstrated high precision in detecting and classifying key elements such as players and the ball within soccer match footage. These findings establish a robust basis for further video analysis in soccer, which could enhance tactical strategies and the automation of match summarization.


1 INTRODUCTION


The intricate flow of player movements, strategic formations, and dynamic decision-making within a soccer match creates a complex data landscape for analysis. The accurate interpretation of this data has immense potential for coaches seeking to optimize player positioning and team strategy. Traditional analysis methods rely on manual observation and limited data points. However, advances in computer vision offer a compelling opportunity to unlock a new level of objectivity and detail.


This paper presents the groundwork for creating an unsupervised pipeline using computer vision techniques to provide comprehensive insights into on-field events and tactical analysis. Our solution aims to


automatically identify individual players, track their movements across the field, classify team affiliation, and extract key tactical trends. By combining object detection and classification algorithms, our approach provides a comprehensive solution to the challenges of traditional soccer analysis. Each area has been tested with different methods to determine the best algorithm for this niche problem.


To bridge the gap between on-field events and data-driven analysis, our research lays the ground for future work, where computer vision becomes an integral tool for extracting complex data from non-enriched video data that can be captured in any location in the stadium. This paper invites further exploration and refinement, ultimately contributing to the advancement of both computer vision and soccer analytics. The findings presented here are part of a larger research project in collaboration with KKS Lech Poznań, which includes an analysis of motor preparation, injury prevention, and player evaluation (see (Piłka et al., 2023; Sadurska et al., 2023)).


^a <https://orcid.org/0009-0001-6572-6000>

^b <https://orcid.org/0009-0001-9764-4798>

^c <https://orcid.org/0000-0002-6132-651X>

^d <https://orcid.org/0000-0002-9969-5257>

^e <https://orcid.org/0000-0003-1206-2076>

^f <https://orcid.org/0000-0002-2897-3176>

2 PROBLEM DESCRIPTION

The proposed pipeline addresses three core tasks: object detection, object tracking, and object classification within soccer match videos. Each of these tasks faces unique challenges, which we describe below:

2.1 Object Detection in Soccer

Object detection is key to computer vision-based soccer analysis, as it allows you to identify players, the ball, and the referee in video frames.

Region-based convolutional neural networks (Girshick, 2015) (R-CNNs), such as the Faster R-CNN, are among the more popular approaches and have shown excellent results. However, these models can have trouble with occlusions, player variations and real-time performance limits. Recent research has also explored single-stage detectors, such as the YOLO architecture, offering real-time inference capabilities (Reis et al., 2023).

Achieving high accuracy while maintaining real-time performance remains challenging, especially for small objects. Despite progress in object detection in soccer, certain challenges remain.

2.2 Object Classification in Soccer Matches

Detecting and tracking players, and classifying them by team is vital to soccer analysis. This helps us understand team dynamics, tactics, and player interactions during a match.

Traditionally, player classification relied on colour-based methods that segmented jersey colours to differentiate between teams. But variations in jersey design, lighting, and occlusion often limit these techniques.

Researchers have explored using deep learning techniques with colour-based features to overcome limitations. For example, Liu et al. (Liu et al., 2023) proposed a two-stage classification framework combining colour and deep learning features, achieving improved accuracy compared to colour-based methods.

Transformer-based architectures show promise. Wang et al. (Wang et al., 2022) demonstrated the potential of vision transformers for robust feature extraction in cluttered scenes, leading to improved player classification performance.

Research into multi-class classification tasks, e.g., player role identification and team formation identification, can provide coaches and analysts with deeper insights into team strategies and individual player

contributions (Asali et al., 2016). The use of contextual information has shown potential to improve classification accuracy (Kim et al., 2022).

By incorporating these cues, models can better disambiguate players.

3 METHODS

Our work focuses on the detection, tracking and recognition of players, referees (people) and balls. Many neural networks can be used for this task. Choosing the right methods allows us to prepare a test environment adapted to our case.

We used the Detectron2 (Wu et al., 2019) platform, which supports different neural network architectures through a single standard tool. We also used one of the new architectures, YOLOv8, which offers improved performance and better detection of small and occluded objects. The Deep SORT architecture was used for player and ball tracking. We tested two methods for player detection: the *K*-means algorithm and the Vision Transformer.

Our method consists of four main steps: Object Detection, Object Tracking and Team Classification. The flow diagram is shown in Figure 1. Balls, players, goalkeepers and referees were detected. Using Narya homography estimation (Garnier and Gregoir, 2021), all objects not on the pitch were excluded. Then each player was grouped with his team. Each belonging object has a unique ID and is tracked.

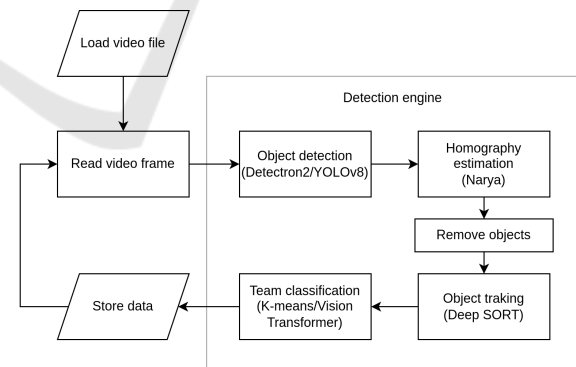


Figure 1: Detection flow diagram represents the sequence of algorithms used in our tool.

We used different algorithms for each of these stages. For object detection, we used three architectures: Faster R-CNN (Ren et al., 2015), Mask R-CNN (He et al., 2017) and YOLOv8 (Redmon et al., 2016). In the player grouping phase, *K*-means or Vision Transformer was used. Object tracking was done using the Deep SORT algorithm.

By using different algorithms, better results can be achieved for a given strategy. Dividing the tool into individual segments allows each algorithm to be focused on separately. This facilitates team collaboration and allows for modular swapping of algorithms. All the algorithms mentioned are described in more detail in the following subsections. Due to the limited space of the article, we have omitted the description of the homography estimation algorithm. More information about it can be found in (Garnier and Gregoir, 2021).

This section discusses the strategy for solving the problem of player and ball detection, tracking and recognition using the above platforms. It describes the structure, how each platform works and how each architecture was used.

3.1 Detection Strategy

For the Detectron2 platform, the Object Detection Algorithms Test Tool (ODATT) (Nowak and Dyczkowski, 2022) was used to train all the algorithms. This tool evaluates the accuracy of player and ball detection in soccer match videos or frames using different object detection algorithms (e.g. Faster R-CNN, Mask R-CNN and YOLOv8). The generated data can be used in evaluation programs (Padilla et al., 2021), and it supports fine-tuning of pre-trained Detectron2 baseline model zoo models (Wu et al., 2019). We have fine-tuned the You Only Look Once v8 (YOLOv8) architecture (Jocher et al., 2023). YOLOv8 integrates Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) modules to improve feature representation at different levels of abstraction.

3.1.1 Detectron2

Detectron2, developed by Facebook AI Research, is a library providing state-of-the-art detection and segmentation algorithms. It supports neural network architectures such as R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, and Panoptic FPN. We focused on Faster R-CNN and Mask R-CNN.

Detectron2 simplifies switching between architectures by adjusting two parameters: the model configuration file and model weights. Pre-trained baseline models are available for all supported architectures. Detectron2 downloads weights automatically if not locally available. To analyze soccer environments, we employed the FV2D tool (Nowak et al., 2022), which implements algorithms for pitch analysis, as shown in Figure 1. Its modular structure allows replacing components by overriding abstract classes. Details about the tool are in (Nowak et al., 2022).

We tested the Faster R-CNN X101 32x8d FPN 3x, which combines Faster R-CNN and ResNeXt object detection networks with FPN. It has 101 hidden layers, a cardinality of 32 (Xie et al., 2017) and a bottleneck width of 8 units. The architecture begins by passing the image through a backbone CNN for feature maps, which are sent to a region proposal network (RPN) to generate regions of interest (RoIs). These RoIs are scaled using RoI pooling and combined with the feature maps for classification, producing a confidence vector and an envelope vector (Ren et al., 2016).

Mask R-CNN X101 32x8d FPN 3x extends Faster R-CNN by adding a branch for instance segmentation masks and using RoIAlign for more accurate feature mapping and scaling. This improves the accuracy of small object detection, essential for tasks such as ball detection (He et al., 2018). Mask R-CNN's generated masks allow the elimination of non-player colours (e.g. turf).

These architectures, supported by pyramid neural networks, enhance the detection of small objects, improving the accuracy of ball detection in video files. The choice of algorithms was guided by the results of (Nowak, 2022).

3.1.2 YOLOv8

YOLOv8 (Redmon et al., 2016), introduced architectural improvements for detecting small and occluded objects in real time. By treating object detection as a regression problem, YOLOv8 simultaneously predicts bounding boxes and class probabilities.

YOLOv8 integrates Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) modules. FPN creates feature maps at multiple scales to recognise objects of different sizes, while PAN aggregates features across network levels using skip connections for contextual information. This combination improves feature representation and increases accuracy.

YOLOv8 uses anchor-free detection, directly predicting object centres and eliminating anchor box offset calculations. This simplifies post-processing and speeds up real-time detection. Soft-NMS refines overlapping boxes, improving accuracy and reducing redundancy.

With real-time speed and improved accuracy for small/obscured objects, YOLOv8 is well suited for challenging tasks such as ball detection in sports analysis where accuracy and performance are critical.

3.2 Deep SORT

Deep SORT (Wojke et al., 2017) was used to track players and balls. It uses a Kalman filter to track,

smooth and fill in missing data (Pei et al., 2019). Convolutional Neural Networks are used to extract features of the tracked objects, such as motion and appearance, while a Hungarian algorithm associates these features and assigns them to the corresponding objects tracked by the Kalman filter. Object detection algorithms (Detectron2, YOLO or others) pass the detected objects frame by frame to the Deep SORT architecture.

Deep SORT analyses the characteristics of each object in the frame and assigns it a corresponding identifier. In subsequent frames, it compares the features and Kalman filter data with previously tracked objects. If similarities and dependencies are detected, the identifier is retained; otherwise, a new identifier is assigned (Wojke et al., 2017). Having unique identifiers for each object, such as players or balls, enables detailed analysis of movements, distances covered and other metrics.

3.3 Classification Strategies

Two methods were used to classify players into teams: dominant colour detection using a K -means algorithm and a vision transformer technique. Details of each approach are given below.

3.3.1 K -means Algorithm

The K -means algorithm was used to detect the colours of the players' outfits and assign them to the appropriate team. For Faster R-CNN and YOLOv8 architectures, the entire bounding box area was considered, while Mask R-CNN provided object masks representing player silhouettes. The masks excluded the background of the soccer field, which improved the accuracy of the colour distribution of the outfits.

Colours extracted from bounding boxes or masks were converted from RGB to HSV to reduce the influence of lighting on colour recognition. These colours were then grouped using the K -means algorithm. The three dominant colours identified correspond to the teams and the category of goalkeepers, who typically wear different uniforms.

3.3.2 Vision Transformer

The Vision Transformer approach used the CogVLM multimodal architecture to classify players into their respective teams. This process consists of two main steps.

In the first stage, CogVLM processes a single video frame to infer the dominant team colours, allowing the model to learn the visual representation of team uniforms without explicit supervision.

In the second stage, the learned team colour representations are used to classify each detected player instance. The model analyses bounding boxes, segmentation masks, and input frames, using its multimodal capabilities to associate player appearances with inferred team colours. This adaptive approach achieves accurate player classification without relying on predefined colour models or heuristics, learning directly from the data.

4 EXPERIMENT

4.1 Experimental Setup

We evaluated the approach for soccer player detection and team affiliation on a diverse dataset with images from our local team (KKS Lech Poznań), as well as the SoccerDB and SoccerNet datasets.

This allowed for a more comprehensive evaluation. The KKS Lech Poznań dataset ensured accurate ground truth annotations, while the SoccerDB and SoccerNet datasets provided a broader representation. This combined dataset allowed a robust evaluation of the models used for player detection and team affiliation tasks.



Figure 2: Example frame from tactical camera.

4.2 Datasets

4.2.1 Tactical Camera Recordings

The hand-labeled dataset from KKS Lech Poznań's tactical camera consisted of 15 soccer match recordings, annotated to provide ground truth labels for player bounding boxes. This dataset encompassed various scenarios, including different stadiums, lighting conditions, camera angles, and player configurations, ensuring a comprehensive evaluation of the proposed approach. The dataset has been developed based on multiple videos from different stadiums and stages of the game. Videos have been created using

the special tactical camera that provides an uninterrupted shot of the entire match, as shown in Figure 2, with constant movements and no cuts in the footage. Those clips have been recorded in various locations and positions of the camera. Each recording has a standardized resolution of 1280×720 pixels and 60 frames per second rate. Part of the training set has been resized to improve the models' performance.

As shown in Figure 3, the labeled images contained a total of 16 280 players, 1 100 balls, 2 134 referees, and 693 goalkeepers instances, annotated with bounding boxes across 1 041 images, shown by the gray line in Figure 3, providing a diverse set of objects to evaluate the performance of the detection and classification models.

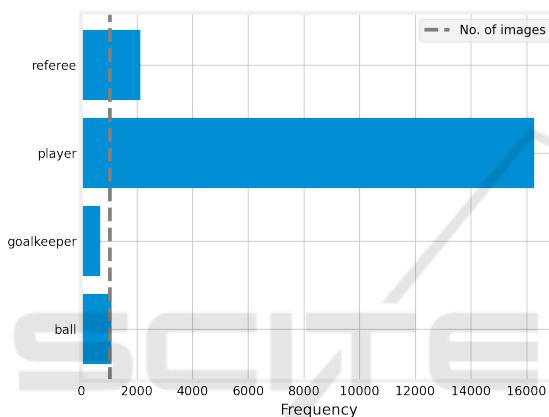


Figure 3: Class frequency across tactical camera dataset.

4.2.2 Data Augmentation

Data augmentation can significantly improve training effectiveness (Zoph et al., 2020). We applied several strategies:

Rotation. Images rotated between -15° and $+15^\circ$ to detect objects in varying orientations.

Blur. Up to 1-pixel blur applied to simulate defocusing and motion blur.

Noise. 0.26% of pixels affected by noise to introduce realistic distortions.

Bounding Box Blur. Boxes blurred up to 2 pixels to account for annotation imprecision.

Rotation and blur values were based on similar works, while noise percentage was selected to add variance without occluding features.

These augmentations, shown in Figure 4, expanded the dataset with realistic variations, increasing the total to 1 452 images (1 259 train, 193 validation). This improved model generalization for soccer game scenarios.



Figure 4: Examples of augmented frames from the dataset. Top left: noise, top right: rotation and blur, bottom left: rotation, bottom right: bounding box blur.

4.3 Evaluation Metrics

We evaluate multiple architectures using these metrics:

- **Recall** measures the ratio of true positive detections to ground truth objects, indicating detection completeness.
- **Precision** measures the ratio of true positives to all positive detections, indicating false positive avoidance.
- **mean Precision (mP)** averages precision values across all detected classes.
- **Average Precision (AP)** measures average precision per class, providing comprehensive performance assessment.
- **mean Average Precision (mAP)** averages AP across all classes for overall performance evaluation.
- **Average Precision at IoU=0.50 (AP50)** and **Average Precision at IoU range from 0.5 to 0.95 (AP50:95)** measure detection precision at specific IoU thresholds.
- **mean Average Precision at IoU=0.50 (mAP50)** and **mean Average Precision at IoU=0.75 (mAP75)** average AP50 and AP75 across classes, indicating performance at different localization strictness levels.

4.4 Object Detection Results

We evaluated the object detection performance using a variety of standard metrics for object detection models. Additionally, we analyzed the localization accuracy using different IoU thresholds – higher thresholds indicating stricter evaluation rules.

4.4.1 Detectron2 – Faster R-CNN and Mask R-CNN

The `faster_rcnn_X_101_32x8d_FPN_3x` and `mask_rcnn_X_101_32x8d_FPN_3x` architectures were pre-trained on `train2017` and evaluated on `val2017` datasets with the 3x schedule (≈ 37 COCO¹ epochs).

Next, models were fine-tuned using our dataset and ODATT tool. Networks were trained during nine epochs, with a batch size of 3 and a learning rate of 0.0125. The number of epochs and the learning rate were selected based on the research in (Nowak, 2022). As the dataset grows, we intend to fine-tune the hyperparameters further. We decided on batch size three due to hardware limitations.

4.4.2 YOLOv8

The model was trained using `yolov8l` and `NAdam` with a 0.06 dropout rate. All other parameters came from the Ultralytics (Jocher et al., 2023). library.

The model excelled at detecting players, followed by referees and goalkeepers. Detecting the ball was the hardest part, since it's often hidden, moves a lot, and remains a small, colourful object on the pitch.

Table 1: YOLOv8 model evaluation.

Class	Precision	Recall	AP50	AP50-95
ball	80.245	40.625	50.918	20.051
goalkeeper	82.421	87.302	88.153	53.668
player	89.717	91.237	90.697	54.222
referee	85.559	90.676	89.517	53.624
all	84.485	77.459	79.821	45.391

Table 1 shows the YOLOv8 model's high performance in detecting players.

It achieves high recall and precision over a wide range of intersection over union thresholds. This indicates its effectiveness in detecting and localising players on the soccer pitch. The model's superior performance in detecting players can be attributed to several factors. Players are larger than the ball, making them

¹COCO is a dataset containing labeled images of complex everyday scenes containing common objects in their natural context (Lin et al., 2015).

easier to detect. They have distinct visual features, such as jerseys and body shapes, which the model can learn to recognise.

The dataset contains enough player instances for the model to learn robust representations. The model performs better at detecting players than balls, with recall values ranging from 0.50 to 0.35 and precision values ranging from 0.80 to 0.82 for IoU thresholds up to 0.90. This is because balls are smaller and harder to detect. The ball is often obscured or partially visible due to player proximity, especially during close interactions or tackles. Appearance varies due to factors such as motion blur, lighting and distance from the camera. The dataset may contain fewer ball instances than player instances, potentially limiting the model's ability to learn robust representations.

4.5 Object Detection Performance Comparison

The models were compared across two datasets. Firstly, on hand labeled, prepared for this study real-world KKS Lech Poznań's tactical camera dataset and a more general comparison using a combination of SoccerDB (Jiang et al., 2020), SoccerNet (Deliege et al., 2021) and additional Polish Ekstraklasa League match recordings, which accumulate to over 20 000 annotated images.

Table 2: Object detection model comparison.

Model	mAP	mAP50	mAP75	mP
Faster RCNN	38.797	75.690	38.480	75.780
Mask RCNN	40.797	75.441	40.903	76.873
YOLOv8	45.389	79.821	47.896	84.485

In summary, the results shown in Table 2 demonstrate the superior performance of YOLOv8 over Faster R-CNN and Mask R-CNN for object detection in soccer game scenarios, particularly in terms of overall accuracy, moderate localization accuracy, and highly accurate object localization. YOLOv8's consistently higher scores across all evaluation metrics suggest its suitability for real-world applications in sports analytics and computer vision tasks.

Table 3: Object types detection model comparison (AP50).

Model	Ball	Referee	Player
Faster RCNN	20.064	44.889	55.509
Mask RCNN	22.260	43.990	54.202
YOLOv8	15.970	48.409	53.583

Table 3 compares the performance of Faster R-

CNN, Mask R-CNN, and YOLOv8 in detecting specific object classes in the soccer game scenario using the tactical camera dataset. While the models' AP values are similar to player detection, the YOLOv8 demonstrates significantly better goalkeeper detection, alongside a slight advantage in referee detection. The Mask R-CNN model better detects the ball, achieving the highest AP among the evaluated architectures.

4.6 Team Affiliation Classification Results

Random frames were selected from each video recording to create the dataset. In total, 309 close-up images of individual players were obtained. Each player image was tagged with a game identifier, which later allowed for the integration of these data into a broader game context for individual classification. Due to the zero-shot nature of the task and the small dataset, we did not split the images into training and validation sets.

Table 4: Object classification approach comparison (TP – True Positive, FP – False Positive, FN – False Negative).

Model	TP	FP	FN	Precision
<i>K</i> -means BBox	202	61	46	76.806
<i>K</i> -means Mask	210	52	47	80.152
CogVLM	202	101	6	66.667

Table 4 compares the results of *K*-means and CogVLM Multimodal Vision Transformer in the classification task.

K-means has two versions, BBox and Mask. The player team was assigned using the area of the bounding box detected by Faster R-CNN. Mask's player team is determined by Mask R-CNN. True Positive (TP) is the number of player appearances correctly assigned to their team. False Positive is the number misclassified as a player. False Negative is the number of players not assigned to any team. False negatives are due to misclassification or misalignment of the bounding box, so we evaluated methods mainly on precision and true positives. The *K*-means method is superior to the multimodal approach, especially when using a mask (80.152 vs. 76.806 vs. 66.667). Both models classified 202-210 players across all teams, showing significant potential. The advantage may be due to a masked algorithm.

5 CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

We have shown that it is possible to detect key objects in soccer match footage, despite challenges such as imperfect video resolution, complex image conditions and dynamic weather.

The best models accurately detected players, referees and goalkeepers. These results, along with significant advances in the classification of detected instances, provide a solid foundation for broader video analysis of soccer matches.

However, the proposed models face limitations, particularly in the ball detection task. It's hard to detect such a small, fast-moving object at 1280×720 resolution, where the ball appears as just a few pixels. This problem is made worse by the ball's similarity to other objects in the scene. Future work will focus on improving ball tracking by integrating trajectory estimation techniques such as (Liu, 2009), which use temporal data to predict the ball's position even when it is occluded or undetected. The use of super-resolution models and contextual data, such as player movement and game dynamics, could also improve detection accuracy.

K-means clustering performs well when players wear multicoloured uniforms or when the class distributions are very similar. The CogVLM model maintains high confidence when multiple players appear within the same bounding box. Future work will explore integrating these approaches to improve the classification.

We're developing a game-changing homograph model to determine players' positions on the pitch. By combining this with external data from providers like StatsBomb, we aim to compute a dynamic perspective matrix. This could enable more accurate spatial analysis of player positioning and movement.

The solutions presented in this article can be adapted to other team sports, such as basketball or hockey, where player tracking and classification face similar challenges. The flexibility of the proposed models allows them to be fine-tuned for different data types, player representations and team affiliations.

Finally, by associating detected player instances with their respective teams and integrating temporal information, it is possible to track individual players throughout a match. This opens the door to comprehensive player and team analysis. Combining player detection with ball tracking could provide a complete understanding of a match, enabling better tactics and decisions. Future work will focus on developing tracking algorithms for soccer, paving the way for a data-driven approach to soccer analysis.

ACKNOWLEDGEMENTS

The publication and the underlying research owe their existence to the invaluable support extended by the KKS Lech Poznań club, which provided access to datasets.

REFERENCES

- Asali, E., Valipour, M., Zare, N., Afshar, A., Katebzadeh, M., and Dastghaibifard, G. (2016). Using machine learning approaches to detect opponent formation. In *2016 Artificial Intelligence and Robotics (IRANOPEN)*, pages 140–144. IEEE.
- Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M. J., Dueholm, J. V., Nasrollahi, K., Ghanem, B., Moeslund, T. B., and Van Droogenbroeck, M. (2021). 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*, pages 4508–4519.
- Garnier, P. and Gregoir, T. (2021). Evaluating soccer player: from live camera to deep reinforcement learning.
- Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2018). Mask R-CNN.
- Jiang, Y., Cui, K., Chen, L., Wang, C., and Xu, C. (2020). SoccerDB: A large-scale database for comprehensive video understanding. In *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports, MM '20*. ACM.
- Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics YOLOv8.
- Kim, H., Kim, B., Chung, D., Yoon, J., and Ko, S.-K. (2022). SoccerCPD3: Formation and role change-point detection in soccer matches using spatiotemporal tracking data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3146–3156.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft COCO: Common Objects in Context.
- Liu, H., Adreon, C., Wagnon, N., Bamba, A. L., Li, X., Liu, H., MacCall, S., and Gan, Y. (2023). Automated player identification and indexing using two-stage deep learning network. *Scientific Reports*, 13(1):10036.
- Liu, S. (2009). *Object trajectory estimation using optical flow*. Utah State University.
- Nowak, J. (2022). Methods for detecting objects in video image and their application in the analysis of sports recordings. Master's thesis, Adam Mickiewicz University in Poznan.
- Nowak, J. and Dyczkowski, K. (2022). ODATT - Object Detection Algorithms Test Tool.
- Nowak, J., Galla, Z., and Dyczkowski, K. (2022). FV2D - Football video to 2 dimensional pitch.
- Padilla, R., Passos, W. L., Dias, T. L. B., Netto, S. L., and da Silva, E. A. B. (2021). A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, 10(3).
- Pei, Y., Biswas, S., Fussell, D. S., and Pingali, K. (2019). An elementary introduction to kalman filtering.
- Piłka, T., Grzelak, B., Sadurska, A., Górecki, T., and Dyczkowski, K. (2023). Predicting injuries in football based on data collected from GPS-based wearable sensors. *Sensors*, 23(3).
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.
- Reis, D., Kupec, J., Hong, J., and Daoudi, A. (2023). Real-time flying object detection with YOLOv8. *arXiv preprint arXiv:2305.09972*.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks.
- Sadurska, A., Piłka, T., Grzelak, B., Górecki, T., Dyczkowski, K., and Zaręba, M. (2023). Fusion of a fuzzy rule-based method and other decision-making models in injury prediction problem in football. In *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, pages 1–6.
- Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F., and Wang, Y. (2022). Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12186–12195.
- Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric.
- Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. <https://github.com/facebookresearch/detectron2>.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks.
- Zoph, B., Cubuk, E. D., Ghiasi, G., Lin, T.-Y., Shlens, J., and Le, Q. V. (2020). Learning data augmentation strategies for object detection. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 566–583. Springer.