Spiideo SoccerNet SynLoc: Single Frame World Coordinate Athlete Detection and Localization with Synthetic Data

Håkan Ardö[1][a], Mikael Nilsson[2][b], Anthony Cioppa[4][c], Floriane Magera[4], Silvio Giancola[3][d], Haochen Liu[1], Bernard Ghanem[3][e] and Marc Van Droogenbroeck[4][f]

[1]*Spiideo, Malmö, Sweden*

[2]*Centre for Mathematical Sciences, Lund University, Sweden*

[3]*Center of Excellence for Generative AI, KAUST, Saudi Arabia*

[4]*Montefiore Institute, Open-SportsLab, University of Liège, Belgium*

ﬂ

Keywords: Synthetic, Dataset, Sports, 3D, Human, Detection, Localization.

Abstract: Currently, most research and public datasets for video sports analytics are base on detecting players as bounding boxes in broadcast videos. Going from there to precise locations on the pitch is however hard. Modern solutions are making dedicated static cameras covering the entire pitch more readily accessible, and they are now used more and more even in lower tiers. To promote research that can take benefits of such cameras and produce more precise pitch locations, we introduce the *Spiideo SoccerNet SynLoc* dataset. It consists of synthetic athletes rendered on top of images from real world installation of such cameras. We also introduce a new task of detecting the players in the world pitch coordinate system and a new metric based solely on real world physical properties where the representation in the image is irrelevant. The dataset and code are publicly available at https://github.com/Spiideo/sskit.

1 INTRODUCTION

The object detection research field has seen a lot of successes using bounding-boxes to represent the detected object and evaluate the performance of detectors c.f. (Wang et al., 2024). This particular representation has proven effective for a lot of downstream tasks and applications, but it is not sufficient for all of them. For instance, several applications require to infer information about the physical world by analysing captured images of the scene. However, the image itself is only an intermediate representation, and evaluations in image-space is therefore not representative of the physical detection of the object. In those cases, evaluations in physical-space are critical, *i.e.*, to measure the localisation errors in meters instead of pixels.

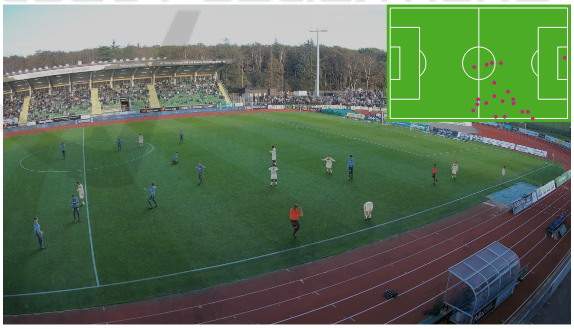

Figure 1: Example synthetic image form our proposed *Spiideo SoccerNet SynLoc* public dataset. The players are 3D generated on a real-captured image of a soccer pitch. The proposed task is to detect and locate the player on the pitch (purple) given the image and the camera calibration.

Sports analytics often requires player localization on the pitch to analyze their positions relative to each other, the ball, and the field. Applications include shot and goal locations, ball possession losses, heatmaps, and Game State Recognition (GSR).

Such analytics can be conducted using broadcast video streams with moving cameras or dedicated static cameras covering the entire pitch. Historically,

[a] https://orcid.org/0000-0001-6214-3662
[b] https://orcid.org/0000-0003-1712-8345
[c] https://orcid.org/0000-0002-5314-9015
[d] https://orcid.org/0000-0002-3937-9834
[e] https://orcid.org/0000-0002-5534-587X
[f] https://orcid.org/0000-0001-6260-6487

Ardö, H., Nilsson, M., Cioppa, A., Magera, F., Giancola, S., Liu, H., Ghanem, B. and Van Droogenbroeck, M.
Spiideo SoccerNet SynLoc: Single Frame World Coordinate Athlete Detection and Localization with Synthetic Data.
DOI: 10.5220/0013108200003912
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 278-285
ISBN: 978-989-758-728-3; ISSN: 2184-4321
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

static cameras were rare and costly, so most datasets and research relied on broadcast video. However, modern solutions have made static cameras more accessible, even in lower tiers, increasing their relevance for research.

A first step toward physical-space evaluation was made with SoccerNet-GSR (Somers et al., 2024), but the dataset is finite and was expensive to produce. It defines the player's physical location as the projection of the image bounding box's bottom edge center onto the ground. How this relate to any physical property of the players is unclear. In this paper, we define the physical location as the orthogonal projection of the player pelvis onto the ground plane, and shows in Section 7.1, that the SoccerNet-GSR definition is a poor approximation of this.

World-space evaluation datasets are more complex than standard image-bounding-box datasets, requiring precise camera calibration and physical ground truth data. While self-driving car datasets often use expensive sensors like radars and lidars, research shows synthetic data can effectively train systems that generalize to real data (Rematas et al., 2018; Black et al., 2023). Synthetic data is now also used in sports analytics training (Leduc et al., 2024; Zhu et al., 2020), with ground truth extracted from 3D rendering models.

Team clothing is crucial in sports analytics, making parametric clothed human models valuable for generating diverse subjects. Layered human representations like SynBody (Yang et al., 2023) and BEDLAM (Black et al., 2023), are particularly relevant.

In this work we propose a new public dataset, called *Spiideo SoccerNet SynLoc*, designed for soccer player analytics. The dataset proposed is based on real world installations of dedicated static cameras.

Contributions. The main contributions are as follows. (**i**) A Bird's-Eye View (BEV) detection and localization task for players using real world locations on a pitch. (**ii**) A publicly released synthetic dataset, called *Spiideo SoccerNet SynLoc*, of soccer scenes with static, calibrated cameras covering the entire pitch. (**iii**) A new metric, called *mAP-LocSim*, for evaluating player localization in real world pitch space. (**iv**) A baseline detector based on YOLOX-pose (Maji et al., 2022; Ge et al., 2021).

Table 1: Comparing the proposed dataset, *Spiideo SoccerNet SynLoc*, with other popular 3D human datasets. The numbers refers to the numer of training images (Imgs), image width in pixesl (Width), number of annotated humans (Hum) and maximum camera height (MaxH). Bold faced names are synthetic datasets.

	Imgs	Width	Hum	MaxH
Kitti	14K	1224	4K	2m
nuScenes	204K	1600	11K	2m
Human3.6M	5.8M	1000	1.4M	2m
MPI-INF3DHP	102K	2048	102K	3m
JTA	230K	1920	300M	N/A
SynBody	1.2M	1024	2.7M	5m
BEDLAM	285K	1280	750K	6m
Proposed	65K	3840	668K	29m

2 RELATED WORK

2.1 Datasets

Publicly available datasets for sports analytics, such as those released by SoccerNet (Giancola et al., 2018; Deliège et al., 2021; Cioppa et al., 2022a; Cioppa et al., 2022b; Mkhallati et al., 2023; Held et al., 2023; Somers et al., 2024)[1], are often based on broadcast or single-view video. This means that the cameras are in motion, which makes them hard to calibrate due to limited visual cues and motion blur. That makes it challenging to get accurate real-world ground truth. These datasets have driven research for a lot of different analytics tasks as shown by the success of the SoccerNet challenges (Giancola et al., 2022; Cioppa et al., a; Cioppa et al., b). However, for some sports analytics, more precise locations of all players in the real world are needed, and in those cases, dedicated static calibrated cameras covering the entire pitch are an interesting alternative. Hence, in this work, we capture data from single static cameras covering half a pitch each, that we release publicly.

Lots of efforts have been put into producing datasets with 3D information about humans, as illustrated in Table 1, leveraging both annotated real world footage and rendered synthetic data. The level of details of the 3D information in those datasets varies from full 3D meshes based on the SMPL (Loper et al., 2015) model to 3D pose keypoints to 3D bounding boxes. These datasets can aid sports analytics training but require bridging a domain gap. For example, in autonomous driving, Kitti (Geiger et al., 2012; Menze and Geiger, 2015) and nuScenes (Caesar et al., 2020) datasets use car-mounted cameras with low viewing angles, unlike typical soccer analytics setups.

[1] www.soccer-net.org

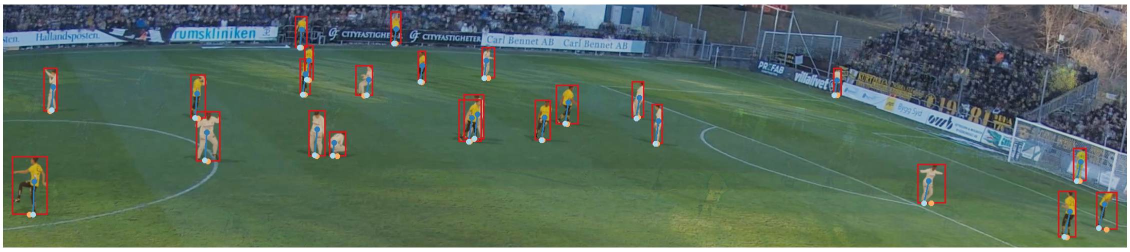

Figure 2: Example of image data annotations available in our proposed *Spiideo SoccerNet SynLoc* dataset. These are only provided for convenience as the evaluation is performed entirely in world pitch coordinates. Our annotations comprise i) bounding boxes (red), ii) ground position, defined as the orthogonal projection of the pelvis onto the ground (light blue) and iii) pelvis (blue). To show that the ground position does not always align with center of bottom edge of the bounding box it is also shown (orange).

Then there are the studio datasets such as Human3.6M (Ionescu et al., 2014) and MPI-INF3DHP (Mehta et al., 2017) that use actors in a studio recorded by multiple cameras and some motion capture system for generating ground truth. These datasets only contains small scenes with one or a few humans. Setting up such a capturing system for an entire soccer pitch would be almost impossible.

A more promising approach consists in rendering synthetic datasets. This can be achieved by intercepting the 3D data passing through the graphics hardware while playing computer games. Examples of this are the JTA (Fabbri et al., 2018), NBA2K (Zhu et al., 2020) and SoccerNet-Depth (Leduc et al., 2024) datasets. However, this limits the variability of the produced datasets to that of the game as it is hard to extend them beyond their original framework.

There are also approaches that build up the 3D models explicitly in layers such, as BEDLAM (Black et al., 2023) and SynBody (Yang et al., 2023). Here, each sample is constructed by randomly choosing an combination of 3D models from large pools of models representing different aspects of the scene, such as background, body shapes, cloths, hair styles, textures and motions. This can create large variations in the datasets as the different aspects can be combined in several different ways. The randomly created 3D model is then rendered using photorealistic rendering engines such as Blender (Blender Online Community, 2018) or Unreal. By choosing which models are available for each aspect it is also possible to control the kind of scenes produced, and it is also easy to extend by adding more 3D models. In this work, we leverage this last approach by superimposing 3D parametric athlete models over a real stadium.

2.2 Metrics

When it comes to metrics, the classical way to measure how well a detected object fits the ground truth annotations is to look at the Intersection over Union (IoU) between the detected image-bounding-box and the ground truth image-bounding-box. This IoU value is typically thresholded to get a criteria specifying if an object have been detected or not. For some use-cases, such as sports analytics, a more relevant detection criteria is to threshold the distance between the estimated location in the real world and the ground truth location.

In tasks where a more detailed representation of the objects is available, this IoU can be replaced with other similarity measures. In pose keypoint detectors for example, the COCO Object Keypoint Similarity, OKS, (Lin et al., 2015a) is typically used instead. It is defined as the mean similarity over all visible keypoints. It does however still measure the similarity in pixels in the image space and not in the real world.

The SoccerNet Game State dataset (Somers et al., 2024) introduces a tracking metric called GS-HOTA. It is based on the HOTA (Luiten et al., 2020) metric, but replaces the IoU similarity measure used there with another measure called LocSim. It is based on the distance, d, between a detected objects location and its ground truth location in the real world ground plane, and defined as

$$e^{\ln 0.05 \frac{d^2}{\tau^2}}, \tag{1}$$

where τ is a constant distance tolerance set to 5 in their work. This similarity measure allows the similarity to be measured in the real world, but HOTA is a tracking metric not suitable to evaluate a single frame detector. However the same approach of replacing the IoU with LocSim can be applied to the common detection metric mAP, which then becomes the proposed metric mAP-LocSim.

This gives a single metric that allows a whole parametric ensemble of detectors to be evaluated as a single entity. The different detectors in the ensemble are formed by varying the score threshold used to discard week detections. However, for a practical usage, this threshold has to be chosen and a single metric does not describe all aspects of a model. Also, it's hard to interpret what the exact values represent. This prompts for the use of additional metrics whose values are more intuitively interpretable. In this work, we propose to choose the final threshold that maximizes the F1-score on the validation-set and use the classical precision, recall and F1-score as additional metrics together with the *frame accuracy* metric, defined in Section 5.

3 TASK DEFINITION: ATHLETE DETECTION AND LOCALIZATION

In this paper a new athlete detection and localization task is proposed. It focuses on the real world problem of detecting players, referees, and bystanders and locating them on a soccer-pitch. More precisely, given an image and the camera calibration parameters relating the image to the real world coordinate system, the objective is to locate each athlete based on the projection of its pelvis onto the ground plane. How the player is represented in the image is irrelevant to the task. The entire evaluation is performed in the real world pitch coordinate system. This allows different representations in the image (bounding-box, pose-keypoints, pixel segmentations, etc.) to be utilized while solving the task.

4 PROPOSED DATASET: SPIIDEO SOCCERNET SYNLOC

To support this task we release a new dataset with ground-truth locations of players in the real world. The data consists of background images from real world installation with synthetic players rendered on top.

4.1 Data Generation

The 3D rendering technique used to render the *Spiideo SoccerNet SynLoc* dataset is based on a combination of the techniques used in BEDLAM (Black et al., 2023) and HeNIT (Ardö et al., 2022). Each scene to render is created from an random combination of different assets which leads to an exponential combination of possible scenes. The tools used (with the exception of the cloth 3D models) are available as open-source[2].

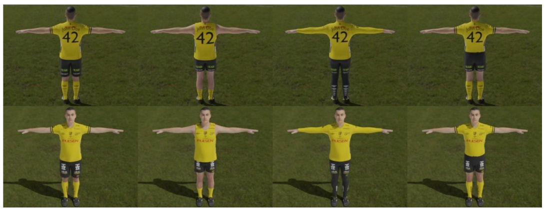
Figure 3: Example renderings of the different cloths 3D models used with one of the teams textures applied to all of them.

At the base real world installations are used to form the background and camera placements. Installations at 17 different arenas with two 4K cameras covering half the pitch each have been used, c.f. Fig. 1. From these scenes, background images were constructed from 134 different games by taking the temporal median over several minutes of video to get rid of any moving objects.

Lighting is applied to the scene by randomly choosing one of 683 different HDRI sky images as described in HeNIT (Ardö et al., 2022), and player bodies (shapes and poses) are sampled from the BEDLAM (Black et al., 2023) dataset. The clothing in BEDLAM is unsuitable for soccer players or referees, so a 3D artist designed custom models (Fig. 3). These included three upper-body, three lower-body, and two sock variations, along with a procedural texture for customizable player names, jersey numbers, colors, stripes, and badges. This allowed for the creation of home and away uniforms for 11 teams, including players and goalkeepers, totaling 44 uniforms. Additionally, three referee uniforms were designed.

For each scene two teams are then randomly chosen, including 10 players and one goalkeeper each. The goalkeeper location is chosen randomly using a Gaussian distribution centered at the center of the goal with a standard deviation $\sigma = 2$ meters. For the players a ball position is chosen randomly using a uniform distribution over the entire pitch, then the player positions are chosen from a Gaussian distribution, truncated to the pitch and centered around this point with $\sigma = 10$ meters.

In addition to the players, the referees are placed in the scene. The main referee position is chosen using the same distribution as the players while the two side line referees are placed along the long sides with a center position chosen uniformly along the line and then drawn from a Gaussian distribution centered at that point with $\sigma = 1$ meter. Outside the pitch 4 bystanders are placed at a distance from the pitch uniformly chosen between 0 and 2 meters and positioned uniformly along the edge of the pitch. The bystander model uses cloths from the BEDLAM dataset.

[2] https://github.com/AxisCommunications/blenderset-addon

4.2 Dataset Annotations

Annotations in the *Spiideo SoccerNet SynLoc* dataset are the players ground locations in world space and camera calibrations. All the annotations are presented in a format compatible with the COCO annotations format (Lin et al., 2015b). That is a list of images and a list of athletes.

For each image a camera calibration consisting of a camera matrix, P, a distortion polynomial, p_{dist} and an undistortion polynomial, p_{undist} is provided. The camera matrix,

$$P = [R \;\; t], \tag{2}$$

specifies the orientation, R, and translation t of the camera related to the ground plane with the origin at the pitch center and the x-axis along the center line and the y-axis perpendicular to it in the ground plane.

The distortion polynomial models the entire lens, including all the intrinsic parameters of the camera, using an industrial distortion model (Trioptics,). This means no intrinsic parameters, K, are present in the camera matrix. The polynomial relates pixels distance from the principal point to the angle of the world ray that is projected onto that pixel. It assumes that the principal point is centered in the image and is here defined on normalized image coordinates,

$$(u_{\text{n}}, v_{\text{n}}) = \frac{1}{w}\left(u - \frac{w}{2}, v - \frac{h}{2}\right), \tag{3}$$

where (u,v) are pixel coordinates in the camera image and (w,h) its size in pixels. It is a radial distortion model and the distortion polynomial, p_{dist} relates the magnitude of the normalized image coordinates, $r_{\text{n}} = \sqrt{u_{\text{n}}^2 + v_{\text{n}}^2}$ to the magnitude of the undistorted coordinates, $r_{\text{u}} = \sqrt{u_{\text{u}}^2 + v_{\text{u}}^2}$ as

$$r_{\text{n}} = p_{\text{dist}}\left(\arctan\left(r_{\text{u}}\right)\right). \tag{4}$$

For convenience there is also an undistortion polynomial fitted to the inverse of this function,

$$r_{\text{u}} = \tan\left(p_{\text{undist}}\left(r_{\text{n}}\right)\right). \tag{5}$$

For each athlete, the annotations consists of both image and world information. In the image there is a bounding-box, the area of a pixelwise segmentation and two 2D keypoints, the pelvis and the physical location on the pitch, projected into the image, see Fig. 2. The world data consists of two 3D keypoints, the pelvis and the physical location on the pitch, see Fig. 1.

Annotations are stored in JSON format, with polynomials as coefficient lists in decreasing monomial degree. Keypoints and the camera matrix are stored as lists of lists, and the image bounding box is represented as $(u, v, w_{\text{box}}, h_{\text{box}})$, denoting the top-left corner and box dimensions.

4.3 Dataset Statistics

The *Spiideo SoccerNet SynLoc* dataset has been split into 42 504 training images, 6 777 validation images, 9 309 test images and 11 352 challenge images. Among the 17 arenas used, two have been solely dedicated to the test-set and two to the challenge-set. About half the images in the test and the challenge sets are based on those dedicated arenas. The other half is based on the same arenas as is present in the training data, but from different games. In total, the entire dataset consists of 1 107 009 annotated humans.

5 EVALUATION METRICS

The main metric proposed for this task is called *mAP-LocSim* and it is based on the common detection metric mAP, but replaces the IoU similarity measure with the LocSim similarity measure defined in Equation 1.

The entire evaluation, using mAP-LocSim, is performed in world space and, since the image representation is irrelevant for the metric, the algorithms are freed from using a specific representation there (i.e. bounding boxes). This allows other representations to be explored.

In the benchmark presented, the LocSim parameter τ was chosen to 1 m based on empirical experiments. This is in contrast to SoccerNet-GSM that used 5. This allows the proposed task to focus on more precise localisation, which is made possible by using real 3D information.

This approach evaluates the model without requiring a final threshold, but one must be selected for practical use. Since false positives and negatives are equally problematic in many sports analytics cases, the final threshold is typically chosen by maximizing the F1-score on the validation set post-training. Using this threshold, we propose *frame accuracy*, a more interpretable metric measuring the percentage of images with perfect predictions—no false positives or negatives, with all players correctly detected. Detection is defined by a LocSim similarity below 0.5, corresponding to a 0.48-meter distance. Figure 5 shows this is achievable across the pitch without sub-pixel image localization precision.

To give an even more detailed picture of the capabilities of different algorithms it is also proposed to use other metrics that show different aspects of their performance. The additional metrics are F1-score, precision and recall.

6 BASELINE METHOD

As a baseline, an off-the-shelf 2D keypoint detector will be used to detect two points: the pelvis and its projection onto the ground plane (player location on the pitch). Each detection provides an estimated image location, projected back to world coordinates using the camera calibration. Image bounding-box bottom-line-centers will also be projected for comparison, highlighting the benefits of 3D information over bounding boxes. The architecture used is YOLOX-pose (Maji et al., 2022; Ge et al., 2021), which consists of a CNN backbone that extracts features directly from the images, followed by a head that for a set of anchor-boxes detects bonding box coordinates, pose point coordinates and scores. A non maximum suppression algorithm is used to suppress similar detections, and then a score threshold can be used to prune week and negative detections.

7 EXPERIMENTS

7.1 Baseline Experimental Setup

The 2D pose detector implementation is based on mmpose (Contributors, 2020) with code released on github[3]. It was trained on the training set of the *Spiideo SoccerNet SynLoc* dataset and evaluated on the test set. The original YOLOX-Pose uses a learning rate of $4 \cdot 10^{-3}$ and introduces an auxiliary loss with a second-stage preprocessing after 280 epochs. This training schedule did not converge when applied to the proposed dataset. Instead the learning rate had to be reduced to 10^{-4} and the auxiliary loss introduced already after 200 epochs. Experiments were performed with training detectors for different resolutions, 640 × 640 and 960 × 960. Otherwise the training process was not altered. Results are shown in Table 2 and Fig. 4.

The models regressing the location significantly outperforms the models that use the center of the bottom edge of the bounding box as the location in all the metrics investigated. Also, increasing the resolution gives a more significant performance boost as compared to increasing the model size. This is most likely due the fact that the athletes are small compared to the image size, and therefore they will consist of very few pixel when the image is scaled down. That means that there is not enough information to distinguish them from the background and thus increasing the model size will not help. Also, the performance of the model using the bounding box flattens out, which probably means that increasing the model size even more will not improve the performance for it.

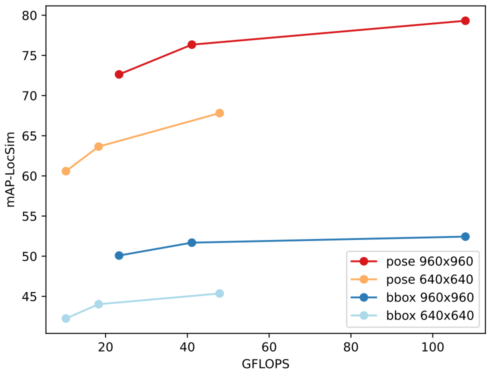

Figure 4: Results of different detectors with different input resolutions on the *Spiideo SoccerNet SynLoc* dataset. The bbox models uses the center of the bottom edge of the image bounding box as the player position while the pose models are regressing the position as a keypoint in the pose detector.

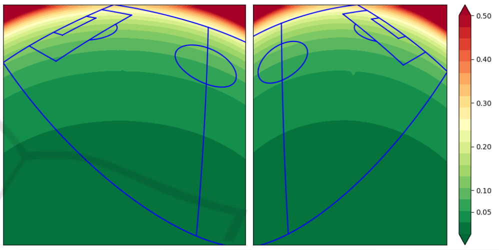

Figure 5: Localisation errors in meters in different parts of the pitch corresponding to a pixel error of one pixel for a some of the real world installations the *Spiideo SoccerNet SynLoc* dataset is based on.

7.2 Pitch Location Uncertainty

Since the camera rig is located on one side of the pitch, a one pixel detection error has a different metric impact depending on the distance of its real world location with respect to the camera.

In order to highlight this impact, for each pixel belonging to the pitch image, we simulate a confusion with one of its direct neighbouring pixels and compute the distance in meters between the deprojection of the pixel on the pitch and its neighbour. For a pixel that is not located on the edges of the image, this gives us 8 metric distances that are averaged and plotted in Fig. 5. The resulting errors in meters remain low and does not exceed 30 cm for pixels corresponding to the opposite side of the pitch. Note that the pixels here refer to the pixels in the original 4k images, which are scaled down 6 respective 4 times in the baseline experiments presented in Table 2.

[3] https://github.com/Spiideo/mmpose/tree/spiideo_scenes/configs/body_bev_position/spiideo_scenes

Table 2: Results of different detectors with different input resolutions on the *Spiideo SoccerNet SynLoc* dataset. Mean average precision, mAP, metrics are presented for two different cases: IoU - standard intersection over union based image bounding box similarity and LocSim - proposed world distance based similarity of predicted pitch location. The YOLOX-pose architecture is used with the bbox models using the center of the bottom edge of the image bounding box as the player location while the pose models are regressing the location as a keypoint in the pose detector. To give a more detailed picture, the classical Precision, Recall and F1 metrics are also reported, and to give a metric that is easier to interpret intuitively, Frame Accuracy is proposed. It presents the amount of images for with a perfect result in terms of false positives/negatives is predicted.

			mAP					Frame
Model	Input Res.	GFLOPS	IoU	LocSim	Precision	Recall	F1	Accuracy
yolox-tiny bbox	640 × 640	**10.3**	50.2	42.2	77.9	70.0	73.7	6.2
yolox-s bbox	640 × 640	18.3	54.5	44.0	79.7	72.0	75.7	6.8
yolox-m bbox	640 × 640	47.9	59.2	45.4	82.1	74.0	77.9	9.2
yolox-tiny bbox	960 × 960	23.3	61.3	50.1	84.8	80.0	82.3	13.6
yolox-s bbox	960 × 960	41.1	65.7	51.7	85.6	82.0	83.7	15.6
yolox-m bbox	960 × 960	108.0	**69.5**	52.4	87.8	83.0	85.3	17.1
yolox-tiny pose	640 × 640	**10.3**	50.2	60.6	81.7	75.0	78.2	10.0
yolox-s pose	640 × 640	18.3	54.5	63.6	84.9	77.0	80.8	11.3
yolox-m pose	640 × 640	47.9	59.2	67.8	87.5	80.0	83.6	15.4
yolox-tiny pose	960 × 960	23.3	61.3	72.6	90.4	84.0	87.1	20.9
yolox-s pose	960 × 960	41.1	65.7	76.3	88.0	88.0	88.0	28.0
yolox-m pose	960 × 960	108.0	**69.5**	**79.3**	**92.8**	**89.0**	**90.9**	**31.6**

8 CONCLUSIONS

In this work we introduce and publish a new synthetic dataset for soccer analytics. It is based on background images from real world installations on top of which synthetic players are rendered. This allows the ground truth annotations to consist of precise 3D camera calibrations and pitch locations of the athletes. Baseline experiments show that this kind of data can improve localisation of players on a pitch significantly compared to using the center of the bottom edge of the image bounding box as the players location projected into the camera image. We also present a new task for detecting and locating players on a pitch and propose a new metric, mAP-LocSim, for evaluation performed entirely in the world pitch coordinate system. We see this dataset as a first step towards opening up new research opportunities for the field without the limitations imposed by using broadcast video as source. Potential future steps could involve extracting more annotations from the rendering pipeline, such as keypoint poses, athlete classes (player, referee, bystander), Jersey numbers, names, pixel segmentations or pixel depths.

ACKNOWLEDGEMENTS

This research has been supported by VINNOVA project 2023-02689 ”DAIDESS”. The rendering of the synthetic data were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre. The training of the baseline models were enabled by resources provided by Chalmers e-Commons at Chalmers. Anthony Cioppa is funded by the F.R.S-FNRS (https://www.frs-fnrs.be/en/).

REFERENCES

Ardö, H., Ahrnbom, M., and Nilsson, M. (2022). Height normalizing image transform for efficient scene specific pedestrian detection. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–11.

Black, M. J., Patel, P., Tesch, J., and Yang, J. (2023). BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737.

Blender Online Community (2018). *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *CVPR*.

Cioppa, A., Deliège, A., Giancola, S., Ghanem, B., and Van Droogenbroeck, M. (2022a). Scaling up SoccerNet with multi-view spatial localization and re-identification. 9(1):1–9.

Cioppa, A., Giancola, S., Deliege, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., and Van Droogenbroeck, M. (2022b). SoccerNet-tracking: Multiple object tracking dataset and benchmark in soccer videos. pages 3490–3501.

Cioppa, A., Giancola, S., Somers, V., and et al. SoccerNet 2023 challenges results.

Cioppa, A., Giancola, S., Somers, V., and et al. Soccernet 2024 challenges results.

Contributors, M. (2020). Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose.

Deliège, 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. pages 4503–4514.

Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., and Cucchiara, R. (2018). Learning to detect and track visible and occluded body joints in a virtual world. In *European Conference on Computer Vision (ECCV)*.

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). Yolox: Exceeding yolo series in 2021.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Giancola, S., Amine, M., Dghaily, T., and Ghanem, B. (2018). SoccerNet: A scalable dataset for action spotting in soccer videos. pages 1792–179210.

Giancola, S., Cioppa, A., Deliège, A., and et al. (2022). SoccerNet 2022 challenges results. pages 75–86. ACM.

Held, J., Cioppa, A., Giancola, S., Hamdi, A., Ghanem, B., and Van Droogenbroeck, M. (2023). VARS: Video assistant referee system for automated soccer decision making from multiple views. pages 5086–5097.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339.

Leduc, A., Cioppa, A., Giancola, S., Ghanem, B., and Van Droogenbroeck, M. (2024). Soccernet-depth: a scalable dataset for monocular depth estimation in sports videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3280–3292.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015a). Microsoft coco: Common objects in context.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015b). Microsoft coco: Common objects in context.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16.

Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., and Leibe, B. (2020). HOTA: A higher order metric for evaluating multi-object tracking. 129(2):548–578.

Maji, D., Nagori, S., Mathew, M., and Poddar, D. (2022). Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. pages 2636–2645.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., and Theobalt, C. (2017). Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE.

Menze, M. and Geiger, A. (2015). Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mkhallati, H., Cioppa, A., Giancola, S., Ghanem, B., and Van Droogenbroeck, M. (2023). SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. pages 5074–5085.

Rematas, K., Kemelmacher-Shlizerman, I., Curless, B., and Seitz, S. (2018). Soccer on your tabletop. In *CVPR*.

Somers, V., Joos, V., Giancola, S., Cioppa, A., Ghasemzadeh, S. A., Magera, F., Standaert, B., Mansourian, A. M., Zhou, X., Kasaei, S., Ghanem, B., Alahi, A., Van Droogenbroeck, M., and De Vleeschouwer, C. (2024). SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap.

Trioptics. Imagemaster. https://www.trioptics.com/products/imagemaster-hr-tempcontrol-universal-image-quality-mtf-testing/.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., and Ding, G. (2024). Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*.

Yang, Z., Cai, Z., Mei, H., Liu, S., Chen, Z., Xiao, W., Wei, Y., Qing, Z., Wei, C., Dai, B., Wu, W., Qian, C., Lin, D., Liu, Z., and Yang, L. (2023). Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20282–20292.

Zhu, L., Rematas, K., Curless, B., Seitz, S., and Kemelmacher-Shlizerman, I. (2020). Reconstructing nba players. In *Proceedings of the European Conference on Computer Vision (ECCV)*.