AutoSoccerPose: Automated 3D posture Analysis of Soccer Shot Movements

Calvin Yeung   Kenjiro Ide
Nagoya University, Japan

yeung.chikwong@g.sp.m.is.nagoya-u.ac.jp
ide.kenjiro@g.sp.m.is.nagoya-u.ac.jp

Keisuke Fujii
Nagoya University, Japan
RIKEN/JST PRESTO, Japan

fujii@i.nagoya-u.ac.jp

Abstract

Image understanding is a foundational task in computer vision, with recent applications emerging in soccer posture analysis. However, existing publicly available datasets lack comprehensive information, notably in the form of posture sequences and 2D pose annotations. Moreover, current analysis models often rely on interpretable linear models (e.g., PCA and regression), limiting their capacity to capture non-linear spatiotemporal relationships in complex and diverse scenarios. To address these gaps, we introduce the 3D Shot Posture (3DSP) dataset in soccer broadcast videos, which represents the most extensive sports image dataset with 2D pose annotations to our knowledge. Additionally, we present the 3DSP-GRAE (Graph Recurrent AutoEncoder) model, a non-linear approach for embedding pose sequences. Furthermore, we propose AutoSoccerPose, a pipeline aimed at semi-automating 2D and 3D pose estimation and posture analysis. While achieving full automation proved challenging, we provide a foundational baseline, extending its utility beyond the scope of annotated data. We validate AutoSoccerPose on Soccer-Net and 3DSP datasets, and present posture analysis results based on 3DSP. The dataset, code, and models are available at: https://github.com/calvinyeungck/3D-Shot-Posture-Dataset.

1. Introduction

Refining the interpretation of images is a fundamental aim in computer vision [26]. It encompasses tasks such as image classification [26], object detection [38], and human pose estimation [23]. In sports, image interpretation has significantly contributed to foul detection [43], player identification [42], and tracking [39]. Recently, such computer vision techniques have been applied to analyze soccer posture effectively [49].

In soccer posture analysis, a high-quality dataset and the analysis method hold profound effects on the analysis result. Concerning the pose dataset, for soccer pose and other
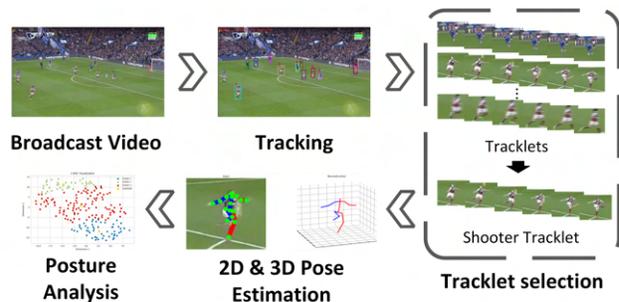
Figure 1. Overview of AutoSoccerPose.

human movements, the pose sequence is often captured in 2D given the use of a standard RGB camera [30]. Currently, publicly available 2D pose datasets include MPII [2], OCHuman [55], and COCO-WholeBody [23]. For soccer and sports image datasets, options include LSP [24], Sports-102 [4], LearningFromThePros [49], and Sport Image [48]. These datasets have been effectively utilized for their designated tasks. Regarding the analysis method, existing studies have analyzed the goalkeeper [36, 49] and shooter [19, 33] postures as they directly affect the match result. With a simplified scenario and linear model, studies have provided valuable insights into posture dynamics.

However, concerning the dataset, the large 2D pose datasets often lack a sufficient number of soccer images, and the soccer-specific datasets provide single images of postures instead of sequences, failing to represent entire movements. Regarding the analysis method, the simplified scenarios may not fully capture the complexities of real-world situations, limiting the applicability of results, especially at the highest competitive levels. Additionally, the utilization of linear models may overlook more intricate relationships in the data.

Therefore, in this study, we propose the 3D Shot Posture (3DSP) dataset, consisting of annotated 2D pose sequences of shooting instances from professional soccer matches. We introduce the 3DSP-GRAE (Graph Recurrent AutoEncoder) model, a non-linear model designed to en-

code pose sequences into vectors, and AutoSoccerPose, a semi-automated pipeline for extracting 2D and 3D pose sequences directly from professional soccer broadcast videos for analysis, as illustrated in Fig. 1. The contributions of this paper are as follows:

- 3DSP (3D Shot Posture) dataset, which comprises 2D pose annotations, 3D pose estimations, and tracklet information extracted from professional broadcast videos.
- 3DSP-GRAE model, a novel nonlinear spatiotemporal model designed to encode pose sequences into vector representations effectively.
- AutoSoccerPose, a pipeline tailored for semi-automated 3D posture analysis in soccer, specifically focusing on player shot movements.
- Extensive evaluations at each stage of AutoSoccerPose, accompanied by an analysis of shooting postures utilizing the 3DSP dataset.

2. Related work

2.1. 2D pose image datasets

For images consisting of humans, 2D human pose estimation has emerged as a well-established research topic. This task aims to accurately define human body part segments or joint keypoints in the image. Commonly used benchmarks for this task include OCHuman [55], COCO-WholeBody [23], both derived from COCO [26], and MPII [2] dataset, covering various scenarios such as daily activities, human interactions, and sports events.

In the realm of sports, image frequently revolves around the posture of athletes. While traditional computer vision tasks such as image classification [4], 2D/3D pose estimation [20, 24], and its extension, human parsing [48], can be undertaken, the intricate anatomy involved, coupled with challenging image conditions like motion blur [4, 24], amplify the difficulty of these tasks. In addition, advancements in computer vision techniques and soccer analytics, particularly those concerning player action value [11, 53, 54], have opened avenues for leveraging soccer player posture in computer vision-based player evaluation [49].

Nonetheless, while multiple datasets consist of pose images of professional soccer athletes, there are a few limitations when applying them for posture analysis:

1. **General-purpose dataset**: Despite the potential abundance of images in conventional datasets, the proportion of professional soccer images is low due to the multitude of classes they encompass. Furthermore, these images do not specifically cater to distinct soccer actions, posing challenges for comparison.
2. **Partial information**: Often, a complete soccer action, such as shooting, is captured through multiple sequential images (poses) [19, 33]. However, current datasets only offer a single pose image for each discrete action.

Moreover, 2D pose data relies on computer vision-based model estimation, introducing errors given the unique characteristics of sports images.

To mitigate these issues, we have collected the 3DSP dataset (see Sec. 4), currently the largest repository of professional soccer human pose data with 2D posture annotations. This dataset focuses on the shooting action, providing sequential images of the shooting movement. In Tab. 1, we summarize the existing datasets containing professional soccer pose images along with the 3DPS dataset.

2.2. Posture analysis in soccer

Posture analysis holds profound importance in soccer, given its direct impact on player performance and strategic optimization. Previous studies have predominantly focused on shooting and goalkeeping postures, recognizing their pivotal roles in influencing match outcomes.

In shooting analysis, researchers have delved into unraveling the intricate differences between various kicking techniques. For instance, [33] examined disparities between non-rotational shots and instep kicks, while [19] investigated distinctions among straight shots, knuckling shots, and curve shots. Both studies utilized advanced camera setups and reflective markers to capture 2D poses of college soccer players' shooting postures. The differentiation was based on 2D pose data, kinematic parameters, and principal component analysis techniques.

On the other hand, studies on goalkeeping posture, such as [49], have focused on analyzing goalkeeper saving techniques in penalty and 1-vs-1 situations. Utilizing techniques like estimated 3D pose extracted with PoseHG3D [58], engineered features, and K-means clustering, researchers examined the various postures used by goalkeepers. Similarly, [35, 36] examined the strategies employed by goalkeepers in penalty kicks. They employed logistic regression to infer the relationship between orientation and strategies, where the orientation was engineered from the 2D pose of goalkeepers and penalty takers, utilizing OpenPose [6] for pose estimation.

While previous studies provide valuable insight into the posture of soccer players, several drawbacks could limit the development of posture analysis:

1. **Simplified scenario**: Previous datasets are mainly collected under simplified scenarios, such as competitive settings with few players or controlled experimental environments. However, soccer actions often occur in occluded and crowded situations captured in broadcast videos.
2. **Labor-intensive**: Collecting data for posture analysis involves labor-intensive processes including image preprocessing, pose estimation or annotation, and quality control. Even for the datasets used in this study, the collection process had taken months to complete.

Dataset	2D Pose Annotation	Sequential Images	Images in Soccer	Number of Images	Soccer Action	Purpose
LSP [24]	✓	✗	——	2k	——	2D Pose Estimation
Sport Image [48]	✓	✗	——	1.3k	——	Human Parsing
MPII [2]	✓	✗	0.2k	41k	——	2D Pose Estimation
OCHuman [55]	✓	✗	——	4.7K	——	2D pose estimation
COCO-WholeBody [23]	✓	✗	——	200K	——	2D pose estimation
LearningFromThePros [49]	✗	✗	1k	1k	GoalKeeping	Posture Analysis
Sports-102 [4]	✗	✗	0.2k	14k	——	Image Classification
3DSP (Ours)	✓	✓	4k	4k	Shooting	Posture Analysis

Table 1. Overview of human pose in sports (professional soccer included) dataset. The datasets are arranged in order of their publication years. Due to the challenges associated with collecting 3D pose annotations during professional football matches, there currently exists no publicly available dataset providing 3D pose annotations for professional football.

3. **Linear model**: Previous methods like principal component analysis and logistic regression offer valuable insights but may fail to capture non-linear spatiotemporal relationships present in posture data.

Given these limitations, this study proposes the AutoSoccerPose for semi-automated 3D posture estimation from professional soccer broadcast video (see Sec. 3) and a non-linear graph-based sequential model 3DSP-GRAE for posture analysis (see Sec. 3.5).

3. Proposed method

AutoSoccerPose[1] aims to extract 3D poses from broadcast videos and analyze the extracted posture. Each stage of AutoSoccerPose corresponds to an existing computer vision task in soccer. Fig. 1 provides an overview of the stages involved in AutoSoccerPose, and each stage is elaborated upon in the subsequent sections.

3.1. Broadcast video

Similar to all posture analyses, Clips that capture the player's posture are fundamental to AutoSoccerPose. For broadcast video, there exist publicly available professional matches videos that could be collected, as in [36], or existing publicly available datasets like SoccerNet [12, 17] that provide a large number of videos (nearly 800 hours of professional soccer matches). Nonetheless, while the posture of professional players in broadcast videos is the most informative and commonly available data, other less complex video footage like those in previous studies (see Sec. 2.2) could be utilized in AutoSoccerPose.

Besides, accurate timestamp annotation of the action, shooting in this case, is critical to cutting clips from the broadcast video and the performance of AutoSoccerPose

[1]For detailed specifications of the AutoSoccerPose, please refer to https://github.com/calvinyeungck/3D-Shot-Posture-Dataset.

(see Sec. 5.2). The timestamp of important events like shooting could easily be retrieved in the post-match report, manually annotated, or publicly available datasets [12, 17, 34] as it was a fundamental type of soccer data annotation. Furthermore, accurate timestamp annotation has been a growing soccer computer vision task in recent times, in SoccerNet Challenge 2023 [10] to facilitate broadcast video understanding, the task action spotting (locate when and what type of event happened in the broadcast video) has been proposed. We reserve the integration of action spotting for future research.

To be more specific, the AutoSoccerPose was currently developed with the videos and timestamp annotation in SoccerNet [12, 17], with frames resolution of 1280×720 and 25 fps. The retrieved clips used for AutoSoccerPose cover 0.5 seconds before and after the annotated timestamps. However via observation, the first 20 frames cover the entire shot movement for most of the shots, therefore only the first 20 were utilized for each shot.

3.2. Tracking

Tracking in AutoSoccerPose aims to identify soccer players in video clips and create tracklets: pixel coordinates of each respective player within each frame of the clips. Conventionally, two main tracking methods are utilized: tracking-by-detection [5, 13, 50], where objects are initially identified with bounding boxes in each frame and then associated (tracked) based on various conditions; and end-to-end tracking [32, 51, 57], where tracklets (bounding boxes of an object) are generated directly from the model.

State-of-the-art soccer player tracking models, achieving the best performance in the SoccerNet Challenge 2023, predominantly follow the tracking-by-detection approach. However, these tracking models are often not open-sourced. Here, we adopt the tracking by detection method using

widely used models: fintuned YOLOv8[2] as the detector and BoTSort [1] as the tracking algorithm. For performance evaluation and discussion of YOLOv8-BoTSort, refer to Sec. 5.1.

3.3. Tracklet selection

Tracklet selection aims to pinpoint the tracklet associated with the shooter. This task can be viewed as a facet of image comprehension, where the objective is to classify whether a sequence of images pertains to the shooter. Analogous image comprehension tasks in soccer encompass sport type classification [4], foul detection [14, 18], and action spotting [10]. Nonetheless, many existing studies propose intricate deep-learning architectures that necessitate substantial and diverse datasets for training or finetuning.

To avoid the need for an extensive dataset, we opt for a simpler approach employing a Convolutional Neural Network (CNN) model[1]. Our model processes the sequence of images through three standard convolutional blocks, each comprising a convolutional layer, ReLU activation function, max pooling layer, batch normalization, and dropout layer. Subsequently, the feature maps of the images within the sequence are concatenated. Finally, the concatenated feature maps are passed through a three-layer Multi-Layer Perceptron (MLP) with ReLU activation to predict whether the tracklet contains the shooter.

For each tracklet, we extract a sequence of images spanning frames 10 to 15. From each frame, we crop a tracklet image of size 96×96 pixels, with the center of the bounding box of the detected player (as described in Sec. 3.2) serving as the center of the tracklet image. The shooter tracklet for each clip is determined by selecting the tracklet with the highest estimated probability from the CNN model. Definitions of clips and frames can be found in Sec. 3.1, while details regarding the selection of necessary frames are outlined in Sec. 5.2.

3.4. 2D & 3D pose estimation

The objective of 2D and 3D pose estimation is to derive posture details from video clips, with a specific emphasis on capturing the movements of the shooter. 3D pose estimation methods can be categorized into direct 3D pose estimation [8, 21, 37] and 2D-3D lifting [30, 41, 52], which involves converting 2D pose coordinates into 3D space. The MPI-INF-3DHP dataset [31] is a widely used benchmark for in-the-wild 3D human pose estimation, resembling scenarios found in soccer broadcasts. Among the best-performing models, those employing the 2D-3D lifting approach, such as MotionAGFormer [30], are prevalent and thus adopted for 3D pose estimation in AutoSoccerPose.

For 2D pose estimation, methodologies are typically classified into bottom-up approaches [6, 7, 16], which de-

tect points in the image and then associate keypoints to form 2D poses, and top-down approaches [22, 41, 52], which first detect individuals using object detection models and then perform pose estimation within cropped bounding boxes. Top-down approaches, often more efficient, are favored, with RTMPose [22] selected for AutoSoccerPose due to its superior performance on datasets like COCO-whole body [23] and our dataset, 3DSP.

In implementation, for shooter tracklets, we extract sequences of 20 images, each sized 100×100 pixels, covering the detected players' bounding boxes and entire bodies, akin to Sec. 3.3. However, cropping images larger than the bounding box might include other players besides the shooter. To mitigate this, as the shooter is always centered in the cropped image, we select the shooter's 2D pose as the one with the center of body (torso) joint closest to the image center. The 2D pose from RTMPose [22] is then lifted to 3D using MotionAGFormer [30]. The performance of both 2D and 3D pose estimation is further discussed in Sec. 5.3.

3.5. Posture analysis

Posture analysis seeks to distinguish various shooting styles among soccer players. Given that shot movements are typically unlabeled, unsupervised learning represents the most suitable approach for this task. Furthermore, as outlined in Sec. 2.2, prior research [19, 36] predominantly employs linear models such as Principal Component Analysis (PCA) and logistic regression to infer 3D posture, potentially limiting their ability to capture non-linear spatiotemporal relationships. Hence, to conduct unsupervised learning on posture data with a model capable of capturing the non-linear spatiotemporal properties, we propose the 3DSP-GRAE (Graph Recurrent AutoEncoder) model[1] to encode the latent (representation) vector, illustrated in Fig. 2. This model draws inspiration from the Graph Convolutional Network (GCN) [25] and LSTM AutoEncoder (LSTM-AE) [40]. While several graph and sequential-based models exist [9, 30, 47], none are explicitly designed for 3D pose encoding. Thus, leveraging the relatively simple structure of 3DSP-GRAE offers the advantage of requiring less data for training compared to fine-tuning existing models.

In general, the 3DSP-GRAE model utilizes an encoder-decoder architecture. For the encoder, 3D poses pass through a GCN to extract spatial relationships. Subsequently, the processed 3D poses are sequentially fed into an LSTM to extract nonlinear and temporal relationships. The output of the LSTM at the last layer and the last hidden state serve as the latent variables (vector representation) of the respective shot movement. In the decoder, another LSTM takes the latent variables and the processed 3D poses with GCN as input, but in reverse order. The LSTM output is then converted back to coordinates using another GCN. Consequently, the decoder output comprises the pre-
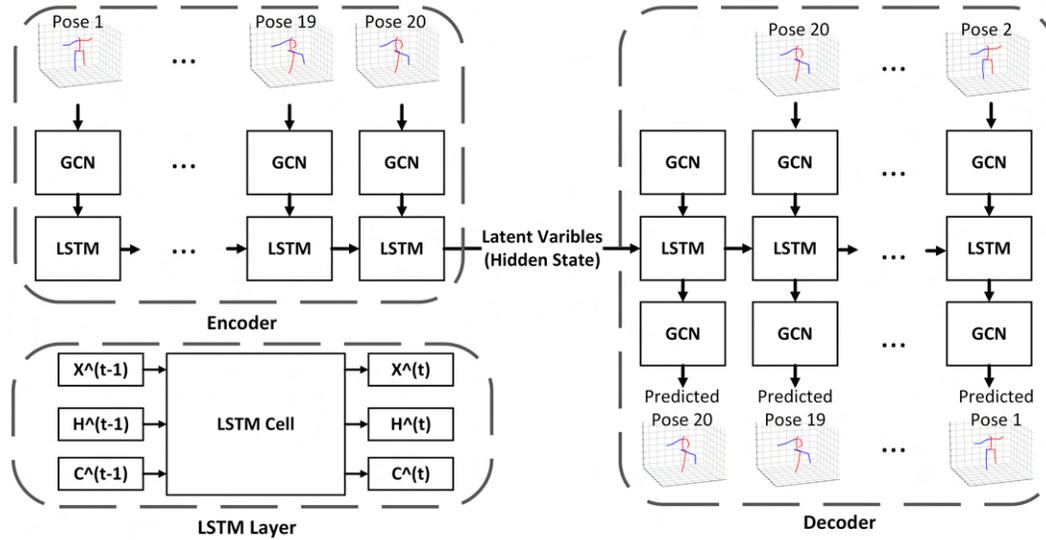
Figure 2. Overview of the 3DSP-GRAE model. The variables X, H, C, and t in the LSTM layer denote the input features, hidden state, cell state, and time step, respectively.

dicted 3D poses. Lastly, by employing the mean square error between the predicted 3D pose and the original input, the 3DSP-GRAE model can be trained.

In detail, each pose comprises 17 joints, as outlined in [30], with each joint represented by XYZ coordinates. The GCN consists of 2 simplified graph convolutional layers with ReLU activation, and the adjacency matrix represents connections between the joints with self-connections. Compared to the original graph convolutional layer [25], the parameter for the adjacency matrix is ignored, considering it to be the identity matrix. Furthermore, the encoder LSTM takes the pose input from pose 1 to 20 in temporal order. Meanwhile, in the decoder LSTM, the first input is zero-padded, followed by poses 20 to 2.

4. 3DSP dataset

In this section, we introduce the 3D Shot Posture (3DSP) dataset[3], a crucial resource for developing AutoSoccerPose, serving as a benchmark for 2D pose estimation in professional soccer match broadcast videos. Furthermore, as evidenced by the comparative study outlined in Tab. 1, 3DSP comprises sequential images capturing soccer movements previously unavailable, with a specific focus on distinct soccer actions such as shooting. 3DSP also represents the largest collection of soccer images annotated with 2D pose information. Additionally, the 3DSP dataset facilitates advancements in soccer posture analysis and offers direct applicability for future studies in this domain. Moreover, the

3DSP dataset could be extended easily with AutoSoccer-Pose.

Data collection process: 3DSP mirrors the process of AutoSoccerPose (with more details and code provided on GitHub [1]), albeit with manual refinement and annotation to ensure the highest quality of 2D poses. Initially, broadcast videos were sourced from SoccerNet [12, 17], and utilizing SoccerNet annotations on actions, the videos were trimmed (0.5 seconds before and after the annotated timestamp, totaling 25 frames). Subsequently, tracklets for these clips were generated employing a fine-tuned YOLO v8[2] in conjunction with BoT-Sort [1]. Furthermore, the tracklet corresponding to the shooter was manually selected and refined, and a cropped image was produced using the bounding box. The 2D poses of the first 20 frames (determined empirically) were then manually annotated and lifted to 3D utilizing the MotionAGFormer [30].

To summarize the dataset structure, we collected data on shot movements from 22 unique matches in the 2015-2016 English Premier League. The training set comprises 20 cropped images (100×100 pixels) per shot movement, with a total of 200 shot movements, resulting in 4000 images. Annotations encompass tracklets, 2D poses, and estimated 3D poses. In the test set, there are also 20 cropped images per shot movement, with a total of 10 shot movements; however, only tracklet information is available. For both sets of data, references to SoccerNet [12, 17] are provided, along with the code utilized for extending 3DSP with SoccerNet [12, 17].

Model	Precision	Recall	$AP_{0.5}$
YOLOv8[2]	95.24%	46.02%	80.33%
RT_DETR [28]	92.01%	50.83%	80.41%
RT_DETR [28] (finetuned)	85.00%	**94.30%**	85.10%
YOLOv8[2] (finetuned)	**95.70%**	93.80%	**97.30%**

Table 2. Soccer player detection models performance. Ranked by AP, with the top-performing result highlighted in bold. Precision quantifies the proportion of correctly detected bounding boxes out of all detected ones, recall measures the proportion of correctly detected bounding boxes out of all ground truth boxes, and AP calculates the area under the precision-recall curve.

Method	DetA	AssA	HOTA
SoccerNet Baseline [10]	38.38%	49.81%	43.67%
YOLOv8[2]-ByteTrack [56]	63.75%	42.31%	51.85%
YOLOv8[2]-BoTSort [1]	68.80%	53.40%	60.54%
YOLOX [15]-Kalisteo [29]	**73.64%**	**73.76%**	**73.65%**

Table 3. Tracking performance on SoccerNet player tracking challenge 2023 [10] test set. The evluation metrics include DetA (Detection Accuracy), AssA (Association Accuracy), and HOTA (Higher Order Tracking Accuracy) [27]. Ranked by HOTA, with the top-performing result highlighted in bold.

Required Frames	Precision	Recall	ACC	CLIP ACC
Frame 12-13	20.81%	**77.50%**	76.52%	52.50%
Frame 11-14	38.33%	57.50%	90.02%	65.00%
Frame 10-15	**57.78%**	65.00%	**93.90%**	**75.00%**
Frame 9-16	41.82%	57.50%	90.94%	65.00%

Table 4. Tracklet selection model performance utilizing different numbers of frames. Ranked by the number of required frames, with the top-performing result highlighted in bold.

5. Experiments

In this section, we verify each step of the AutoSoccerPose (see Sec. 3). More details is available at GitHub[1].

5.1. Tracking performance on SoccerNet [12]

In this subsection, we assess the performance of AutoSoccerPose tracking using the test set provided in the SoccerNet Player Tracking Challenge 2023 [10]. This dataset comprises ground truth tracking annotations derived from broadcast videos in SoccerNet [12, 17]. We conduct evaluations for both player detection and tracking.

For player detection, we employ evaluation metrics including precision, recall, and average precision (AP) (see Tab. 2 caption). The evaluated models encompass state-of-the-art object detection models introduced subsequent to the SoccerNet Player Tracking Challenge 2023 [10], including YOLOv8 [2] and RT_DETR [28], alongside their finetuned variants trained on the SoccerNet Player Tracking Challenge 2023 data [10], utilizing the Python package Ultralytics. The outcomes are delineated in Tab. 2. The result signified that YOLOv8 with finetuning yielded the most favorable player detection results. The relatively diminished recall rate before finetuning was attributed to player occlusion. However, with finetuning, YOLOv8 achieved a 97% AP, in contrast to RT_DETR [28], which exhibited an increase in AP but a decrease in precision.

Concerning tracking evaluation, the compared methodologies encompass the baseline and the top-performing approach (Kalisteo [29]) from the SoccerNet Player Tracking Challenge 2023 [10], in addition to two prevalent tracking techniques integrated with the finetuned YOLOv8. The outcomes are depicted in Tab. 3. It was evident that the Kalisteo method [29] continued to outperform popular methods. However, the absence of publicly available code for the Kalisteo method [29] posed a limitation. Furthermore, the test set of the SoccerNet Player Tracking Challenge 2023 [10] comprised numerous clips featuring substantial player occlusion, whereas in most shooting scenarios, player occlusion was less pronounced. Hence, we ascertain that BoT-Sort suffices for AutoSoccerPose.

5.2. Tracklet selection required features

After verifying the tracking, the next crucial step is tracklet selection, performed using a CNN model in AutoSoccerPose (see Sec. 3.3). Given the absence of existing models tailored for this specific task, we focus on validating essential features, namely the frames containing pivotal information. Our evaluation criteria encompass precision and recall metrics, alongside accuracy (ACC) and clip accuracy (CLIP ACC). The ACC metric assesses the CNN model's accuracy in classifying a tracklet as the shooter's tracklet. Additionally, CLIP ACC evaluates the accuracy of tracklet selection, specifically in identifying the shooter's tracklet from all tracklets within a clip (see Sec. 3.3). 200 clips are sourced from SoccerNet [12, 17], complemented by tracklet information from the 3DSP train set (see Sec. 4).

The feature set is based on tracklet images spanning frames 12 and 13, bounding the timestep of the shot movement (see Sec. 3). We iteratively enrich the feature set by progressively incorporating the frames preceding and succeeding the aforementioned frames. For each feature set, the CNN model is trained using PyTorch with an 80/20 train-validation split. Grid searching is employed to optimize the number of convolutional layers, MLP layers, and the hidden size of the MLP layer. The results are summarized in Table Tab. 4. Notably, employing tracklet images from frames 10 to 15 yields the best performance. While the tracklet selection model demonstrates reasonable per-

Model	Head	Sho	El	Wri	Body	Hip	Knee	Ank	PDJ	AUC
HRNet [41]	94.56%	63.38%	21.14%	17.29%	96.35%	78.68%	47.46%	34.94%	56.08%	47.38%
DWPose [52]	86.69%	85.80%	72.75%	58.88%	89.88%	71.41%	45.41%	32.98%	67.94%	55.80%
RTMPose [22]	**96.81%**	**95.08%**	**86.22%**	**76.01%**	**96.98%**	**95.62%**	**88.30%**	**79.00%**	**89.51%**	**73.56%**

Table 5. 2D pose estimation performance on 3DSP. Ranked by PDJ, with the top-performing result highlighted in bold. Sho, El, Wri, and Ank represent Shoulder, Elbow, Wrist, and Ankle respectively. Each column for body parts indicates the average PDJ for the specified body part, while the PDJ column denotes the mean PDJ across all 17 keypoints.
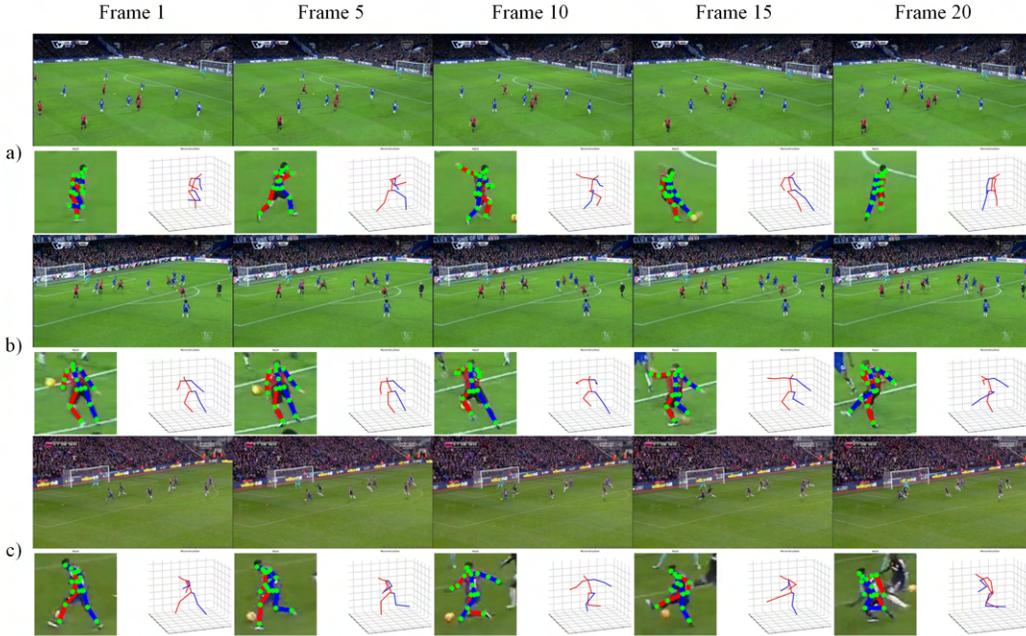
Figure 3. AutoSoccerPose qualitative results. Each row depicts the AutoSoccerPose 2D and 3D pose estimation from a broadcast video, and the columns denote which frame of the video. The top photo represents the broadcast video frame for each cell, while the bottom-left and bottom-right images correspond to the 2D and 3D pose estimations, respectively.

formance, there remains potential for further enhancement.

5.3. 2D pose estimation performance on 3DSP

This subsection aims to evaluate the performance of 2D pose estimation models under zero-shot conditions. We utilized the image and 2D annotations from the 3DSP dataset train set (see Sec. 4). The dataset provides annotations for 17 keypoints representing various body joints. To assess the performance, we employed the Percent of Detected Joints (PDJ) evaluation metric [45]. In this metric, a keypoint is considered detected if the normalized distance between the predicted and ground truth keypoints is under 0.5. Additionally, the area under the PDJ curve (AUC) was calculated to report the performance across different thresholds.

Given the dynamic nature of the shooting movements, especially significant rotation of the shoulders and hips during shots, the distance between the contralateral shoulder and hip keypoints may not accurately represent the torso's largest distance consistently. Therefore, we normalized the distance between detected keypoints and ground truth keypoints by the distance between the center of the shoulder and the center of the hip, as suggested in [44].

The evaluated models include DWPose [52] and RTM-Pose [22], which are the top-performing models in the COCO-WholeBody dataset [23] for 2D pose estimation tasks. Additionally, we included HRNet [41], which is the 2D pose estimation model integrated into MotionAGFormer [30]. The results were summarized in Tab. 5, indicating that RTMPose [22] exhibited the best performance across all joint detections. The detection errors predominantly occurred in limb detection. Through observation, two primary reasons were identified: when the shooter faced backward, the left and right limbs could be misidentified, and when the shooter faced sideways, one side of the limbs was covered by the torso.
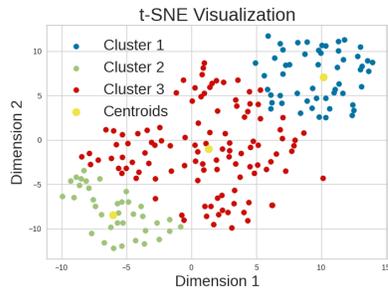
Figure 4. K-means clustering visualization.

Figure 5. Shot movement closest to the cluster centroid. The row and column denote the cluster and frame, respectively.

5.4. 3D pose estimation on non-annotated data

Here, we validate the results obtained by AutoSoccerPose for estimating 3D poses from broadcast videos, instead of relying on 2D annotations from 3DSP. This approach directly showcases what AutoSoccerPose users can retrieve without annotated data. Qualitative results are depicted in Fig. 3, comprising the outcomes of three clips: a) a shot towards the goal on the right from outside the box, b) a shot towards the goal on the left from outside the box, and c) a shot towards the goal on the left from inside the box. While the majority of 2D and 3D pose estimations were satisfactory, two limitations of AutoSoccerPose were identified: motion blur, as in Clip (a) Frame 15, which impeded the accurate detection of limb keypoints, and occlusion between the shooter and other players, as in Clip (b) Frame 20, hindered the detection of limb positions.

5.5. Soccer shots posture analysis

This subsection is dedicated to analyzing the 3D posture of professional soccer players. We leverage the 3D pose data available in the 3DSP dataset (see Sec. 4) to ensure high data fidelity for insightful analysis. Our goal is to identify various shooting techniques. To achieve this, we employ the 3DSP-GRAE model (see Sec. 3.5) to extract representative vectors from sequential shot movements (comprising 3D postures from 20 frames). Subsequently, we apply K-means clustering [3] to categorize these vectors. The resulting clusters are visualized in a 2-dimensional space using t-distributed stochastic neighbor embedding (t-SNE) [46] and illustrated in Fig. 4. Furthermore, we showcase the shot movements closest to each cluster centroid by displaying frames 11 to 15 in Fig. 5. Based on observation, Cluster 1 and Cluster 2 represent inside shots, whereas Cluster 3 denotes instep shots. The definitions are as follows:

- **Inside shot**: Precise technique where the ball is struck with the inside of the foot, allowing for accurate placement into goal corners. Involves swinging the foot inward toward the opposite foot for finesse shots.
- **Instep shot**: Powerful technique involving striking the ball with the area of the foot containing the laces. Com-

monly used for long-range or forceful shots. Players approach the ball straight on and use the laces to generate significant power.

Comparing Cluster 1 and 2 with statistics, Cluster 1 exhibited a 16% greater shooting feet ankle average travel distance. Moreover, the maximum vertical coordinate in Cluster 1 surpassed that of Cluster 2 by 33%, while the minimum knee angle was 17% lower. Statistics revealed that shots in Cluster 1 generated a larger swing motion compared to those in Cluster 2, likely attributable to deeper knee bending and a more pronounced vertical trajectory. This enabled enhanced momentum transfer and force generation during the shot. The identified shooting style holds promise for further utilization in player evaluation for optimizing posture and style, as demonstrated in [49]. Meanwhile, when the aforementioned analysis was performed with linear model PCA and K-means [3], the clustering result could only reflect shots to the left, center, and right, demonstrating the need for a non-linear spatiotemporal model (see GitHub[1]).

6. Conclusion

In this paper, we introduce AutoSoccerPose, a framework designed for estimating both 2D and 3D shooting postures of soccer players and conducting comprehensive posture analysis. Additionally, we present the 3DSP dataset, which stands as the largest annotated 2D pose dataset for professional soccer images. Through experimental validation, we showcase the performance of AutoSoccerPose across each stage of the framework, culminating in an assessment of its overall efficacy. Given that each component of AutoSoccerPose corresponds to a well-established task in computer vision, the seamless integration of future state-of-the-art models into AutoSoccerPose would readily enhance its capabilities. We envision that our research will inspire automated posture analysis in soccer and various other sports.

7. Acknowledgement

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Botsort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 4, 5, 6

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1, 2, 3

[3] David Arthur, Sergei Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Soda*, pages 1027–1035, 2007. 8

[4] Asish Bera, Mita Nasipuri, Ondrej Krejcar, and Debotosh Bhattacharjee. Fine-grained sports, yoga, and dance postures recognition: A benchmark analysis. *IEEE Transactions on Instrumentation and Measurement*, 2023. 1, 2, 3, 4

[5] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 3

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2, 4

[7] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 4

[8] Sungho Chun, Sungbum Park, and Ju Yong Chang. Learnable human mesh triangulation for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2850–2859, 2023. 4

[9] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2262–2271, 2019. 4

[10] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Deliège, Jan Held, Carlos Hinojosa, Amir M Mansourian, et al. Soccernet 2023 challenges results. *arXiv preprint arXiv:2309.06006*, 2023. 3, 4, 6

[11] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Vaep: an objective approach to valuing on-the-ball actions in soccer. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20*, pages 4696–4700. International Joint Conferences on Artificial Intelligence Organization, 2020. 2

[12] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic

[13] understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4508–4519, 2021. 3, 5, 6

[13] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 3

[14] Jiale Fang, Calvin Yeung, and Keisuke Fujii. Foul prediction with estimated poses from soccer broadcast video. *arXiv preprint arXiv:2402.09650*, 2024. 4

[15] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 6

[16] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14676–14686, 2021. 4

[17] 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. 3, 5, 6

[18] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. Vars: Video assistant referee system for automated soccer decision making from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5085–5096, 2023. 4

[19] Sungchan Hong, Chulsoo Chung, Keiko Sakamoto, and Takeshi Asai. Analysis of the swing motion on knuckling shot in soccer. *Procedia Engineering*, 13:176–181, 2011. 1, 2, 4

[20] Christian Keilstrup Ingwersen, Christian Møller Mikkelstrup, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. Sportspose-a dynamic 3d sports pose dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5218–5227, 2023. 2

[21] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7718–7727, 2019. 4

[22] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Realtime multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 4, 7

[23] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 196–214. Springer, 2020. 1, 2, 3, 4, 7

[24] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, page 5. Aberystwyth, UK, 2010. 1, 2, 3

[25] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4, 5

[26] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2

[27] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021. 6

[28] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu. Detrs beat yolos on real-time object detection. *arXiv preprint arXiv:2304.08069*, 2023. 6

[29] Adrien Maglo, Astrid Orcesi, Julien Denize, and Quoc Cuong Pham. Individual locating of soccer players from a single moving view. *Sensors*, 23(18):7938, 2023. 6

[30] Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6920–6930, 2024. 1, 4, 5, 7

[31] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 4

[32] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 3

[33] Taishu Nakamura, Tasuku Miyoshi, Shota Sato, Motoki Takagi, Yasuhisa Kamada, and Yoshiyuki Kobayashi. Differences in soccer kicking type identified using principal component analysis. *Sports Engineering*, 21:149–159, 2018. 1, 2

[34] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):236, 2019. 3

[35] Guilherme de Sousa Pinheiro, Vitor Bertoli Nascimento, Matt Dicks, Varley Teoldo Costa, and Martin Lames. Design and validation of an observational system for penalty kick analysis in football (ospaf). *Frontiers in Psychology*, 12:661179, 2021. 2

[36] Guilherme de Sousa Pinheiro, Xing Jin, Varley Teoldo Da Costa, and Martin Lames. Body pose estimation integrated with notational analysis: A new approach to analyze penalty kicks strategy in elite football. *Frontiers in sports and active living*, 4:818556, 2022. 1, 2, 3, 4

[37] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G Narasimhan. Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15190–15200, 2021. 4

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[39] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. Soccertrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3569–3579, 2022. 1

[40] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015. 4

[41] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 4, 7

[42] Tomohiro Suzuki, Kazushi Tsutsui, Kazuya Takeda, and Keisuke Fujii. Runner re-identification from single-view video in the open-world setting. *arXiv preprint arXiv:2310.11700*, 2023. 1

[43] Ryota Tanaka, Tomohiro Suzuki, Kazuya Takeda, and Keisuke Fujii. Automatic edge error judgment in figure skating using 3d pose estimation from a monocular camera and imus. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 41–48, 2023. 1

[44] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27, 2014. 7

[45] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 7

[46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[47] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*, pages 764–780. Springer, 2020. 4

[48] Yang Wang, Duan Tran, and Zicheng Liao. Learning hierarchical poselets for human parsing. In *CVPR 2011*, pages 1705–1712. IEEE, 2011. 1, 2, 3

[49] Matthew Wear, Ryan Beal, Tim Matthews, Tim Norman, and Sarvapali Ramchurn. Learning from the pros: Extracting professional goalkeeper technique from broadcast footage. *arXiv preprint arXiv:2202.12259*, 2022. 1, 2, 3, 8

[50] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017. 3

[51] Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixé, and Xavier Alameda-Pineda. How to train your

deep multi-object tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6787–6796, 2020. 3

[52] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 4, 7

[53] Calvin Yeung and Keisuke Fujii. A strategic framework for optimal decisions in football 1-vs-1 shot-taking situations: An integrated approach of machine learning, theory-based modeling, and game theory. *arXiv preprint arXiv:2307.14732*, 2023. 2

[54] Calvin Yeung, Tony Sit, and Keisuke Fujii. Transformer-based neural marked spatio temporal point process model for football match events analysis. *arXiv preprint arXiv:2302.09276*, 2023. 2

[55] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019. 1, 2, 3

[56] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 6

[57] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pre-trained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22056–22065, 2023. 3

[58] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE international conference on computer vision*, pages 398–407, 2017. 2