This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# Semi-Supervised Training to Improve Player and Ball Detection in Soccer

Renaud Vandeghen*
University of Liège

Anthony Cioppa*
University of Liège

Marc Van Droogenbroeck
University of Liège

## Abstract

*Accurate player and ball detection has become increasingly important in recent years for sport analytics. As most state-of-the-art methods rely on training deep learning networks in a supervised fashion, they require huge amounts of annotated data, which are rarely available. In this paper, we present a novel generic semi-supervised method to train a network based on a labeled image dataset by leveraging a large unlabeled dataset of soccer broadcast videos. More precisely, we design a teacher-student approach in which the teacher produces surrogate annotations on the unlabeled data to be used later for training a student which has the same architecture as the teacher. Furthermore, we introduce three training loss parametrizations that allow the student to doubt the predictions of the teacher during training depending on the proposal confidence score. We show that including unlabeled data in the training process allows to substantially improve the performances of the detection network trained only on the labeled data. Finally, we provide a thorough performance study including different proportions of labeled and unlabeled data, and establish the first benchmark on the new SoccerNet-v3 detection task, with an mAP of* 52.3%*. Our code is available at [https://github.com/rvandeghen/SST].*

## 1. Introduction

Sports analytics has been steadily growing over the last decade [22], pushed by the development of advanced artificial intelligence and computer tools. Last year, the market was estimated at more than 1 billion dollars, with most indicators pointing out a growth by 500% within the next 5-10 years [14, 18]. Therefore, sports analytics will become even more central for the sports industry in the coming years. Some companies already offer analytics services to clubs with the purpose to improve their playing performances and ascend the championship ladder, thus generating more revenues from ticket sales, advertisements and merchandising.

(*) Denotes equal contributions. Contacts: r.vandeghen@uliege.be and anthony.cioppa@uliege.be.

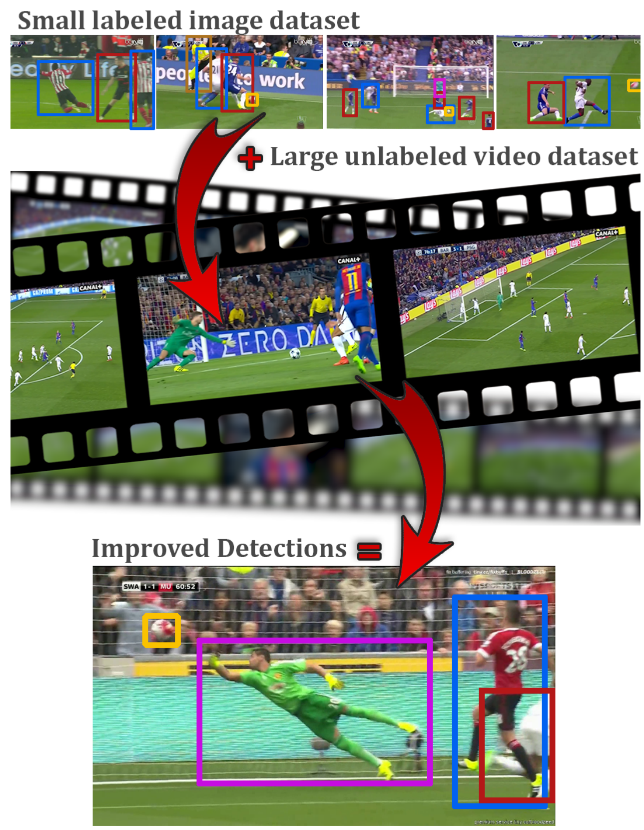


Figure 1. **Overview.** Given a small labeled image dataset for object detection in soccer such as the players, the ball, or the referees, we leverage a large unlabeled dataset of soccer broadcast videos for training an object detector in a semi-supervised fashion. Our training technique allows us to significantly improve the performance of the object detector for the targeted soccer application.

Nowadays, most sports analytics products either rely on manual inspection, which has a heavy cost in terms of manpower, or more recently on automated analysis systems based on artificial intelligence and computer vision techniques. The first step of automated systems often relies on accurately retrieving the players and the ball, which are the key elements to grasp the course of the game. From this information, deeper analyses may be performed such as tracking the players to extract individual speed performance, estimating the field coverage by a defending team to unveil potential weaknesses, or analyze critical pass decisions. All these are powerful indicators of an individual's performance, and game strategy analyses may reveal the strengths and weaknesses of opponent or one's own team. Accurately detecting the players and the ball is therefore crucial since analyses rely on these preliminary results.

Over the past few years, artificial intelligence techniques have surpassed their hand-crafted features algorithms counterparts in many areas including player and ball detection in sports. Even though many deep learning detection networks are publicly available for sports companies and researchers, they are often trained on generic data that are not specifically tailored for each sport. The domain gap between the training dataset and the targeted application often results in performances lower than expected, which is why training, or at least fine-tuning, on sport specific data is often required. However, this may require huge amounts of data, which can be costly to annotate and cannot be transferred from one sport to another. Furthermore, some recent works showed that training the network on sport, and even stadium or team, specific data allows to substantially improve the performance of those networks [8].

In this paper, we present a novel generic semi-supervised method for training an object detector with few annotated soccer images, by leveraging a large unlabeled dataset of soccer broadcast videos as illustrated in Figure 1. More specifically, we develop an iterative teacher-student training approach with three different training loss parametrizations for the student, which may doubt the detections performed by the teacher based on their confidence score. We show that including unlabeled data in the training process allows to substantially increase the performances of the detection network on unseen soccer games. Specifically, we provide a complete performance study for different proportions of labeled and unlabeled data, and establish the first benchmark for the detection task on the new SoccerNet-v3 [5] dataset. It is important to note that the presented ideas and achievements do not rely on any data knowledge about soccer, nor on the network architecture. Therefore, our method is applicable to any other sport or domain, characterized by a low amount of annotated data and a large dataset of unlabeled data, and for any detection network.

**Contributions.** We summarize our contributions as follows. **(i)** We propose a novel semi-supervised method for training a player and ball detection network in soccer games with a teacher-student approach. **(ii)** We introduce three loss parametrizations for training the student with the objective to doubt detections performed by the teacher based on their confidence scores. **(iii)** We establish the first detection benchmark on the new SoccerNet-v3 dataset.

## 2. Related Work

**Object detection in sport analytics.** Object detection has been massively studied in the context of sports analytics as it provides a strong basis for further analyses techniques [44]. Even though the first detection algorithms used background subtraction to detect players [2, 35], they have been quickly overthrown by deep learning networks such as convolutional neural networks (CNN). For instance, the authors of [39] use a shallow CNN to detect players on a hockey field with different image representations. Other methods rely on pre-trained networks such as Mask R-CNN [15, 34, 47]. Recently, Cioppa *et al.* [7] proposed a cross-modality online distillation method for player detection and counting on low budget stadium. Liu *et al.* [28] developed a method to detect players and automatically match them with object such as hockey players and their stick.

Some other works use detection as a first step for various downstream tasks such as improving action spotting using camera calibration and player localization [6], player and ball tracking [17, 21, 30], or to model pass feasibility [1].

In order to train deep learning networks, the AI for sports community can count on a large variety of datasets for sports analytics. SoccerNet [11] and SoccerNet-v2 [9] propose 500 complete broadcast soccer games with annotated action events, camera cuts and classes, and replay information. A complementary dataset with spatio-temporal event annotations focusing on player statistical analyses was released by Pappalardo *et al*. [32]. Yu *et al*. [48] and SoccerDB, published by Jiang *et al*. [20], provide annotations for more than 200 soccer games with player bounding boxes and shot transitions. Lately, SoccerNet-v3 [5] was released, providing manual bounding box annotations for player and other objects of interest such as the ball, the lines, and the goal, with extra annotations such as jersey numbers and re-identification of players across multiple views.

**Object detection in general.** Together with image classification, object detection is among the most studied task in computer vision. Many object detection architectures have been developed in the past few years thanks to the availability of large-scale datasets such as Pascal VOC [10] or MS COCO [26]. Usually, object detectors come into one of two main flavors: two-stage detectors [12, 13, 15, 24, 38], and one-stage detectors [25, 27, 36, 37, 42]. For two-stage detectors, a proposal module is used to propose regions of interest where potential object candidates are likely to be located, for example with a region proposal network such as in Faster R-CNN [38]. The proposals are later refined in a second module, where a class is associated with each predicted bounding box. One-stage detectors operate differently, and directly output the bounding boxes with their classes, leading to faster inference, but often at the price of a lower accuracy compared to their two-stage counterparts.

For these reasons, in this work, we will focus on the two-stage Faster-RCNN [38] architecture, which is widely used in semi-supervised object detection. Note however that our method is applicable regardless of the network architecture.

**Semi-supervised object detection.** Following the successes of semi-supervised methods achieved for image classification [3, 4, 33, 40, 45, 49], many semi-supervised learning methods for object detection have been developed over the past few years. In 2019, Jeong *et al*. [19] proposed a consistency method for the detections made for an image and its horizontally flipped version. More recently, Sohn *et al*. [41] designed a teacher-student approach [23, 29, 43, 46, 49], where the teacher model is trained with labeled data in a supervised manner, and used to produce pseudo-labels on the unlabeled data. These pseudo-labels, along with the labeled data, are then used to train the student model, leading to better performances. This teacher-student approach relies on a selection mechanism to include or reject pseudo-labels, which is often performed by comparing their confidence score to a threshold. However, determining the appropriate threshold value is an arduous process as it is prone to generate noise, resulting in false positives or false negatives. Therefore, authors have promoted different learning strategies for the student, including Unbiased Teacher [29], which addresses the bias issue regarding the dominant classes with a weighted focal loss [25] for the classification head, and Soft Teacher [46], which uses a confidence score for each pseudo-label to weight the background classification loss. In this paper, we present a weighting strategy on the foreground boxes rather than the background ones, with a doubt mechanism based on the confidence score of the pseudo-labels.

## 3. Method

**Problem statement.** We leverage the availability of unlabeled data to improve the detection performance as follows. Given a model tailored for a detection task on images, and trained with a dataset $\mathcal{D}_l$ comprising $N_l$ labeled images, we make use of a dataset $\mathcal{D}_u$ comprising $N_u$ unlabeled images to increase the detection performance of the model; annotations of a labeled image consist in the bounding boxes and classes for all objects contained in it.

This setup is very common in artificial intelligence as datasets are extremely time-consuming and expensive to annotate. Therefore, only a tiny portion of the available data is usually annotated and used for training a model. In this work, we show how to exploit unlabeled images in a semi-supervised fashion for sports analysis. In particular, we propose a method based on a teacher-student approach, where a teacher model $\mathcal{T}$ is trained only with the labeled data, and a student model $\mathcal{S}$ is trained with the labeled and unlabeled, for which pseudo-labels are produced by $\mathcal{T}$.

**Iterative semi-supervised training.** The first step of our method consists in training the teacher model $\mathcal{T}$ with a standard supervised learning technique on the labeled dataset $\mathcal{D}_l$. Once $\mathcal{T}$ is properly trained, we generate pseudo-labels for images of the unlabeled dataset $\mathcal{D}_u$. More precisely, $\mathcal{T}$ processes each image of $\mathcal{D}_u$ and outputs the box, class and confidence score for each detected object. To avoid multiple predictions of the same object, a classical non-maximum suppression is performed. Let us note that, at this point, the performance of $\mathcal{T}$ corresponds to the typical case of training a model in a supervised fashion on a labeled dataset. Hence, the performance of the first teacher $\mathcal{T}$ is the baseline for comparisons in Section 4.

The next step consists in training a student $\mathcal{S}$, which has the exact same architecture as $\mathcal{T}$, on both $\mathcal{D}_l$ and $\mathcal{D}_u$. The training is performed in a supervised fashion, identical to that of $\mathcal{T}$, but on a larger concatenated dataset (that could be seen as a dataset augmented by $\mathcal{D}_u$). The training loss of $\mathcal{S}$ is taken as the sum of two equal contributions, that is

$$\mathcal{L} = \mathcal{L}_l + \mathcal{L}_u \ , \tag{1}$$

with $\mathcal{L}_l$ and $\mathcal{L}_u$ corresponding to the loss on the labeled dataset and unlabeled dataset, which now contains pseudo-labels, respectively. Once the training is stopped, we fine-tune $\mathcal{S}$ with $\mathcal{D}_l$, to make sure to finalize the training on real ground-truth annotations. While being known in the machine learning community and to the best of our knowledge, the fine-tuning step has only been used once before by Li *et. al* [23] in a self-training method for object detection, despite being highly efficient, as shown in Section 4.

These two steps (generating the pseudo-labels with $\mathcal{T}$ and training $\mathcal{S}$) may be iterated, by considering the last student as the new teacher and re-generating the pseudo-labels on $\mathcal{D}_u$. Hopefully, since the prediction quality of $\mathcal{S}$ is expected to be higher than $\mathcal{T}$, the next pseudo-labels should be better as well and improve the training of the next student.

Since $\mathcal{T}$ is not perfect (otherwise we could stop the training process there), $\mathcal{D}_u$ will contain truly detected objects (true positives), but also some predictions that do not correspond to any real objects (false positives), as well as some missing objects (false negatives). These errors in $\mathcal{D}_u$ affect the training of $\mathcal{S}$, and require to find the best practical trade-off. In the following, we propose three training loss parametrizations for the student based on the confidence score of the proposals in order to reduce the impact of potential errors. The whole pipeline is drawn in Figure 2.

**Loss parametrization 1: single threshold.** A first way to alleviate false positives in the dataset consists in selecting a subset of the pseudo-labels in $\mathcal{D}_u$ to only retain the true positive predictions and remove the false positive ones. This is usually done by solely keeping predictions with a confidence score higher than a given threshold $\tau_h$. This reduces the number of positive proposals in the $\mathcal{D}_u$ dataset and in-

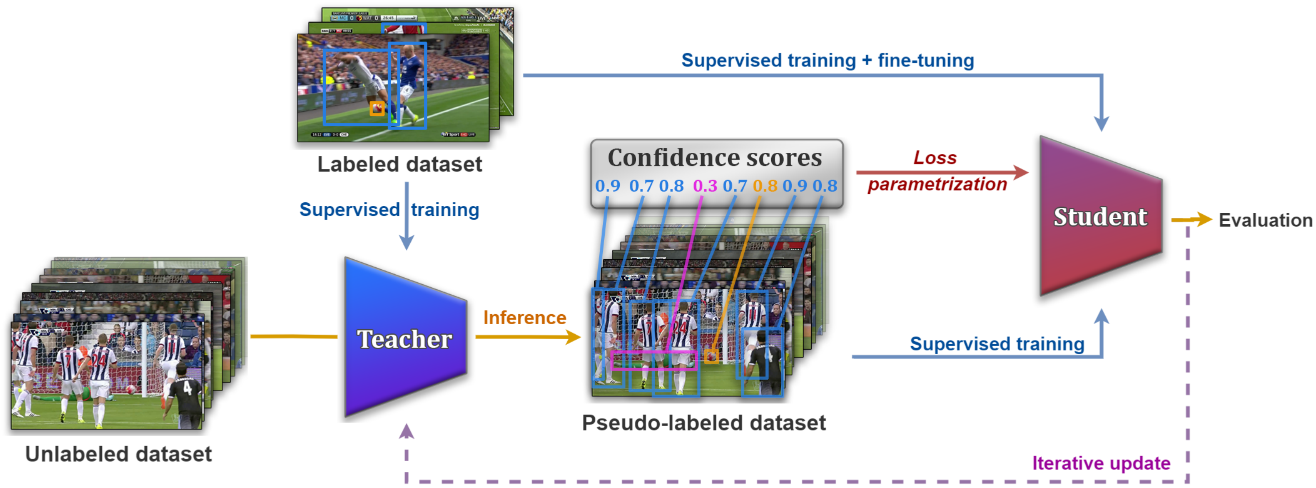


Figure 2. **Overview of our semi-supervised training method for player and ball detection.** We first train a teacher network on a labeled dataset in a fully supervised fashion. Then, we use the trained teacher to produce pseudo-labels on the unlabeled dataset. This creates a first pseudo-labeled dataset, with a confidence score for each prediction. The labeled and pseudo-labeled datasets are then used to train a student network, whose training loss is parameterized based on the confidence score with one of the three parametrization introduced in this paper. This allows the student to doubt unsure proposals by the teacher and achieve good performances on the test dataset. At the end of the training, a final fine-tuning phase is performed with the labeled data, and the student becomes the new teacher for the next iteration.

creases the number of background proposals. The training loss term $\mathcal{L}_l$ of Equation 1, corresponding to the labeled dataset during the training of the student, can be written as:

$$\mathcal{L}_l = \sum_{i=1}^{N_l} \sum_j \mathcal{L}_{cls} + \mathcal{L}_{reg} \ , \tag{2}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ denote the classification and box regression loss respectively, and the superscript $j$ stands for the $j$th proposal for image $i$. Likewise, the training loss on the unlabeled dataset, $\mathcal{L}_u$, can be written as:

$$\mathcal{L}_u = \sum_{i=1}^{N_u} \sum_j \mathcal{L}_{cls} + \mathcal{L}_{reg} \ . \tag{3}$$

Recent works [29, 41, 43, 46] have shown that using a relatively high threshold value ($\tau \geq 0.7$) ensures pseudo-labels of high quality. This parametrization has two effects: (1) it allows to keep predictions which are supposedly true positives, and (2) predictions boxes with low confidence score are associated to the background and therefore correctly removed. However, the downside is that true positive predictions may also have a confidence score lower than this threshold, leading to the introduction of incorrect false negatives in the dataset. In fact, the threshold value acts as a trade-off between precision and recall, given that lower values tend to increase the recall despite lowering the precision, whereas higher threshold values have the opposite effect. Thus, with the choice of a high threshold value, the trade-off tends towards a higher precision, at the price of introducing false negatives.

**Loss parametrization 2: double threshold and doubt.** In order to take into account the potential false negatives, we introduce a second threshold value $\tau_l$ separating true background predictions with a very low confidence score from the remaining predictions. The goal of this second threshold is to create a range of confidence scores, that is $[\tau_l; \tau_h]$, for which we ignore whether the predictions belong to an actual objects or not. For all predictions with a confidence score in this range, we set the loss to $0$ so that the proposals are neither used as positive nor negative examples. This allows to introduce doubt in the training process of the student for unsure predictions of the teacher. The training loss for $\mathcal{D}_l$ is the same as for the first parametrization, but now for $\mathcal{D}_u$, we modify Equation (3) to introduce the new doubt range:

$$\mathcal{L}_u = \sum_{i=1}^{N_u} \sum_j \alpha_j \left(\mathcal{L}_{cls} + \mathcal{L}_{reg}\right) \ , \tag{4}$$

where the term $\alpha_j$ is defined as follows:

$$\alpha_j = \begin{cases} 0 & \text{if} \quad \tau_l \leq s_j < \tau_h, \\ 1 & \text{otherwise}, \end{cases} \tag{5}$$

where $s_j$ is the confidence score associated to the $j$th proposal. Thus, pseudo-labels whose confidence score lies between $\tau_l$ and $\tau_h$ do not contribute anymore to the loss term $\mathcal{L}_u$. By doing so, we can increase the value of $\tau_h$, ensuring that the positives that we introduce actually correspond to true positives regardless of false negatives introduced in the previous parametrization. This provides more flexibility than for the first parametrization.

**Loss parametrization 3: double threshold and progressive doubt.** Finally, one could argue that predictions with a confidence score close to $\tau_h$ are more reliable than predictions with scores close to $\tau_l$. Therefore, we adapt the second parametrization by introducing a doubt that decreases between the two thresholds. This allows us to tune the uncertainty from high for predictions close to $\tau_l$, to low for predictions as their confidence score approaches $\tau_h$. Equations (2) and (4) stay the same, but Equation (5) becomes:

$$\alpha_j = \begin{cases} \frac{s_j - \tau_l}{\tau_h - \tau_l} & \text{if} \quad \tau_l \leq s_j < \tau_h, \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

The weighting term of our three parametrizations, for the loss associated with each positive proposal, is illustrated in Figure 3. Note that for the three parametrizations, the weight loss associated with negative proposals is unchanged, regardless of the confidence score, as the background follows a different dynamic than the foreground. Indeed, it is not possible to assign a confidence score to a region without proposals from the teacher. In other terms, this means that we cannot alleviate false negatives already present in the pseudo-labeled dataset. False negative region proposals based on video analysis as in [7], and a loss parametrization for the rejected proposal ($\leq \tau_l$) may be considered. However, they are out of the scope of this paper and could be studied in a further work.

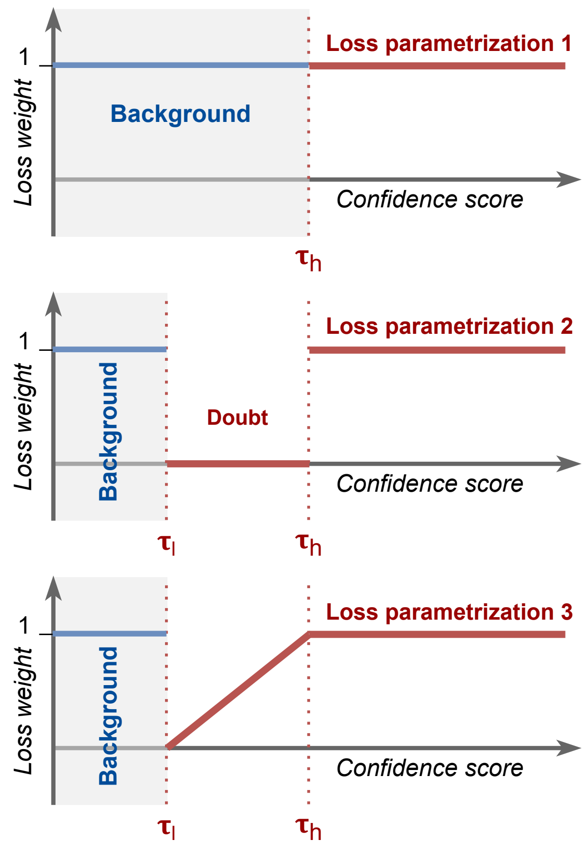


Figure 3. **Our three loss parametrizations for positive candidates.** Comparison of the evolution of the proposal loss weight (corresponding to $\alpha_j$) with respect to the prediction confidence score for our three parametrizations for positive candidates (in red). (1) Simple threshold value to discriminate between the positive proposals and the background by assigning the same loss weight to all positive samples. (2) Introduction of a second threshold to delimit a doubt zone where the loss is zeroed out. (3) Soft linear approximation for the loss weight in the doubt zone to give more importance to predictions close to $\tau_h$. Note that the loss weight is always 1 for background proposals (in blue), regardless of the parametrization for the positive proposals.

## 4. Experiments

**Dataset.** The SoccerNet [11] dataset provides the largest public soccer video collection, including 550 complete broadcast games from the six most influential soccer championships in Europe. Recently, new annotations were released as part of SoccerNet-v3 [5] including 344,660 human bounding boxes of players, referees, and staff, and 26,939 annotations of salient objects such as the ball. These annotations are spread across 33,986 images representing salient moments in soccer with actions such as goals, cards, corners, and their replays.

We choose the training set of SoccerNet-v3 as our labeled dataset, which contains 24,459 frames, its validation set to evaluate performance during training and compare the different loss parametrizations, with 4,797 frames, and its test set for evaluating our final performance, with 4,730 frames. For our unlabeled set, we first retrieve the broadcast videos of the training set games of SoccerNet, which accounts for about 435 hours of video, and extract images at 1 frame per second. This amounts to almost 1,6 million unlabeled frames across 290 different games, which is 64 times more images than the labeled training set!

For the detection task, we focus on the six most important classes for soccer analysis: player, goalkeeper, main referee, side referee, staff, and ball. This amounts to more than 250,000 ground-truth bounding boxes with a highly non-uniform class distribution. This dataset allows us to study our method in many cases ranging from few to many labeled and unlabeled data, with class imbalance and a wide range of object sizes, covering most practical use cases.

**Training setup.** Both the teacher and student models are based on the same Faster R-CNN [38] architecture with FPN [24] and a ResNet-50 [16] backbone pre-trained on ImageNet. Therefore, these networks are composed of a first-stage region proposal network (RPN) and a second-stage detection network, each having their own classification and regression losses for training. Regarding Equations (1), (2),

(3) and (4), we simply equivalently consider the RPN and detection losses as described in those equations, with the total loss becoming the sum of all four losses.

For the first training phase of the teacher on the labeled dataset, we use the SGD optimizer with an initial learning rate of $0.02$, momentum of $0.9$, and a weight decay of $10^{-4}$. We choose to evaluate our model on the validation set with the mAP ($AP_{50:95}$) metric after every epoch, which is a common metric for object detection. If no improvement is made regarding the mAP for $5$ consecutive epochs, we reduce the learning rate by a factor of $10$. The models are trained using $4$ GPUs with $8$ images per batch per GPU, with synchronized batch normalization layers across the different GPUs. For both the RPN and detection modules of Faster R-CNN, we use the standard smooth L1 loss for the regression part $\mathcal{L}_{reg}$ and the cross-entropy loss for the classification part $\mathcal{L}_{cls}$. Note that for the detection module, we also weight the classification loss for each proposal according to the class proportion in $\mathcal{D}_l$, which is a common procedure to counter the class imbalance problem. Specifically, this prevents the networks from focusing too much on the most represented class such as players compared to less represented ones like the balls. Furthermore, we use a simple data augmentation process in which we randomly apply horizontal flipping and color jittering for each training sample. Finally, as an early stopping strategy, we cut off the training of the model if no improvement is made with respect to the mAP on the validation set for $10$ consecutive epochs or if the training reaches $200$ epochs.

Next, during the inference phase of the teacher, we process all frames of the unlabeled dataset and gather all detection with their confidence scores, localization, and classes, creating the pseudo-labeled dataset. Afterwards, the student network is trained on both the labeled and pseudo-labeled dataset by randomly mixing the samples of both datasets. The exact same training procedure than the one for the first teacher is used except that for each sample of the pseudo-labeled dataset, we parameterize the training loss according to one of the three techniques introduced in Section 3. Once the student finishes training, either by early stopping or by reaching the maximal number of epochs, we fine-tune it on the labeled dataset only.

Finally, the student network is evaluated and becomes the new teacher network for the next iteration. The pseudo-labeled dataset is re-computed with this new teacher and a new student is trained following the above procedure.

**Quantitative results.** We evaluate our method on increasing labeled dataset sizes to study scenarios ranging from very few to lots of annotated data. In particular, we select the following sizes: $1\%, 5\%, 10\%$, and $100\%$ of $\mathcal{D}_l$, which corresponds to 3, 14, 29 and 290 games (193, 1,196, 2,475, and 24,459 frames respectively). The sampling is operated at the match level rather than at the frame level to stay close to a real-world application in which new data comes from a whole game. For the unlabeled dataset, it is unfortunately too slow to train the model on the whole unlabeled dataset for each setup. Therefore, for most of our experiments, we sample $10$ extra matches, not belonging to the labeled matches, which represents around $55{,}000$ frames. Nevertheless, we evaluate our method once on the entire labeled and unlabeled datasets (corresponding to $1{,}596{,}387$ frames) for the best set of parameters found on the restricted unlabeled dataset, which defines the first detection benchmark on the SoccerNet-v3 dataset. Those choices follow the recommendations of Oliver *et al*. [31] regarding the evaluation of semi-supervised learning methods.

Table 1. **Best performances of the teacher and the fine-tuned student after a single iteration.** Performance of our method are given for several labeled dataset sizes, trained with a fixed amount of 10 extra unlabeled games (that is $55{,}000$ frames). According to best practices, hyper parameters such as the threshold values of our parametric losses are optimized on the validation set only. In addition, the performances for the test set are calculated after training with the entire labeled and unlabeled datasets, and the optimal parameters obtained on the validation set. The mAP value of $\mathbf{52.3}\%$ is the first detection benchmark on the new SoccerNet-v3 dataset. ($^\dagger$corresponds to $\tau_h = 0.9$)

| Method | $\tau_l$ | $\tau_h$ | Validation set 1% | 5% | 10% | 100% | Test set 100% |
|---|---|---|---|---|---|---|---|
| Teacher | - | - | 18.1 | 31.9 | 39.5 | 52.7 | 51.0 |
| Param. 1 | - | 0.99 | 25.8$^\dagger$ | 38.6 | **44.3** | 53.7 | – |
| Param. 2 | 0.9 | 0.99 | 26.0 | 38.7 | **44.3** | **53.8** | – |
| Param. 3 | 0.9 | 1 | **26.2** | **38.9** | 43.7 | **53.8** | **52.3** |

For each labeled dataset size and each loss parametrization, we optimize the threshold values $\tau_l$ and $\tau_h$ using a grid search strategy on the validation set according to good practice in semi-supervised learning. A complete ablation study of these parameters is presented in the next subsection. The results for the fine-tuned student models after the first iteration may be found in Table 1. As can be seen, the optimal threshold values $\tau_l$ and $\tau_h$ are quite high for the three loss parametrizations, indicating that we select predictions for which the teacher is extremely confident. Furthermore, for all dataset sizes, each parametrization systematically outperforms the teacher, which is the baseline corresponding to a strictly supervised approach. We can also see that the second and third parametrizations have comparable results, but operate better than the first parametrization with a single threshold. This indicates that doubt introduced by those parametrizations is beneficial for training the student.

Then, we evaluate only once our method trained with the entire labeled and unlabeled datasets on the test set, choosing the best performing loss parametrization and thresholds based on the previous experiments with the restricted unlabeled dataset. As can be seen in Table 1, the best performing method on $100\%$ of the training data with $10$ extra games is

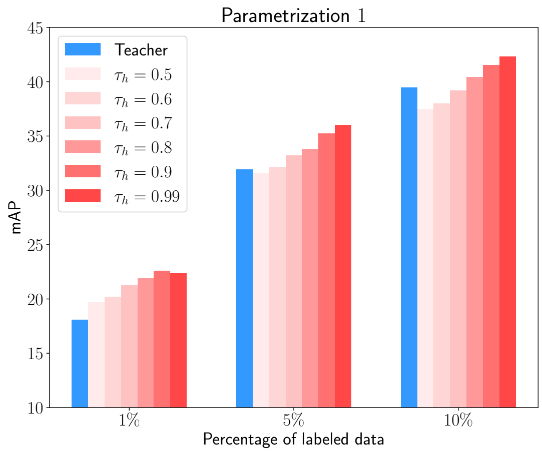


Figure 4. **Optimal threshold value for the first parametrization.** Comparison of the performance of the first parametrization for different threshold values $\tau_h$ on various labeled dataset sizes, with 10 extra unlabeled games. The performance of the student increases with the threshold value indicating that only predictions for which the teacher is certain should be considered. Also, the student manages to surpass the teacher for each dataset size.

obtained with the third parametrization and threshold values of $\tau_l = 0.9$ and $\tau_h = 1$. Therefore, we train a student model on the whole labeled and unlabeled dataset with those parameters as well. Since this experiment has a high training time, a single iteration is performed. We achieve an mAP of 52.3% with the fine-tuned student, improving the performance of the teacher by 1.3%, which is slightly better than with 10 extra unlabeled games (52.0% on the test set). This shows that our method improves the detection performance compared with fully supervised methods, especially when considering few annotated data and that more unlabeled data leads to greater improvements.

**Ablation study.** In this analysis, we start by reviewing the effect of fine-tuning the student, then we propose a thorough study of $\tau_l$ and $\tau_h$ for our three loss parametrizations, and finally, we explore the further gain one can expect when considering multiple iterations of our method.

First, we discuss the benefit of fine-tuning the student on $\mathcal{D}_l$ at the end of the training process. Table 2 shows the performance of the student before and after fine-tuning for each dataset sizes on the validation set (the results on the right of the arrow are the ones of Table 1). As can be seen, fine-tuning allows to significantly improve the performance no matter the parametrization or the labeled dataset size. For this reason, in this ablation study, we only consider the performance *before fine-tuning* as this step takes consequent computation time and that the important observations can be made on the differences between the performances rather

Table 2. **Fine-tuning comparison.** Performance improvement when fine-tuning the student network on the labeled dataset at the end of the training for different labeled dataset sizes, with 10 extra unlabeled games. After fine-tuning, the performance increase for all dataset sizes and all parametrizations, showing the importance of this last training step ($^\dagger$corresponds to $\tau_h = 0.9$).

| Method | 1% | 5% | 10% | 100% |
|---|---|---|---|---|
| Teacher | 18.1 | 31.9 | 39.5 | 52.7 |
| Param. 1 | $22.6^\dagger \rightarrow 25.8$ | $36.0 \rightarrow 38.6$ | $42.3 \rightarrow 44.3$ | $52.6 \rightarrow 53.7$ |
| Param. 2 | $23.1 \rightarrow 26.1$ | $36.6 \rightarrow 38.7$ | $43.0 \rightarrow 44.3$ | $52.6 \rightarrow 53.8$ |
| Param. 3 | $23.0 \rightarrow 26.2$ | $36.1 \rightarrow 38.9$ | $41.9 \rightarrow 43.7$ | $52.7 \rightarrow 53.8$ |

Table 3. **Optimal threshold values for the second parametrization.** Comparison of the performance of the second parametrization before fine-tuning for different threshold values $\tau_l$ and $\tau_h$ on 10% of the labeled dataset size with 10 extra games as unlabeled data. The performance of the student increases with both threshold values, indicating that predictions should be considered as background samples for high values of the confidence score as well.

| $\tau_l$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.99 |
|---|---|---|---|---|---|---|---|---|---|
| $\tau_h$ | 0.6 | 0.7 | 0.8 | 0.9 | 0.9 | 0.9 | 0.9 | 0.99 | 0.999 |
| mAP | 38.1 | 39.0 | 39.5 | 40.4 | 40.9 | 41.1 | 41.4 | **43.0** | 41.0 |

than their absolute values.

Second, we investigate the influence of the threshold values on our three loss parametrizations. For the *first loss parametrization*, we study the influence of $\tau_h$ which conditions the proportion of false positive and false negative proposals introduced in the pseudo-labeled dataset. The performance of the teacher and student models for the different sizes of labeled dataset and for values of $\tau_h$ ranging from 0.5 to 0.99 are shown in Figure 4. For all sizes, increasing the threshold value tends to increase the performance. Furthermore, all student models achieve better performance than the teacher for high threshold values, indicating that even with a simple strategy it is possible to improve on supervised methods using unlabeled data. For the student model trained with 5% and 10% of the labeled dataset, the optimal threshold value corresponds to $\tau_h = 0.99$, showing that it is better to be more selective at the expense of generating false negatives, rather than introducing false positives in the unlabeled dataset.

For the *second parametrization*, we analyze the influence of $\tau_l$ and $\tau_h$ independently, and provide the results only on 10% of the labeled dataset with 10 extra unlabeled games, since the other labeled dataset sizes lead to similar observations. Our setup is the following: (1) we vary $\tau_h$ from 0.6 to 0.9 with a fixed value of $\tau_l = 0.5$, and (2) we vary $\tau_l$ from 0.6 to 0.8 with a fixed value of $\tau_h = 0.9$. We also evaluate this parametrization with higher threshold values ($\tau_l = 0.99$ and $\tau_h = 0.999$). All results are presented in Table 3. Similarly to the first parametrization, we see that the performance increases with $\tau_h$. In addition, higher val-

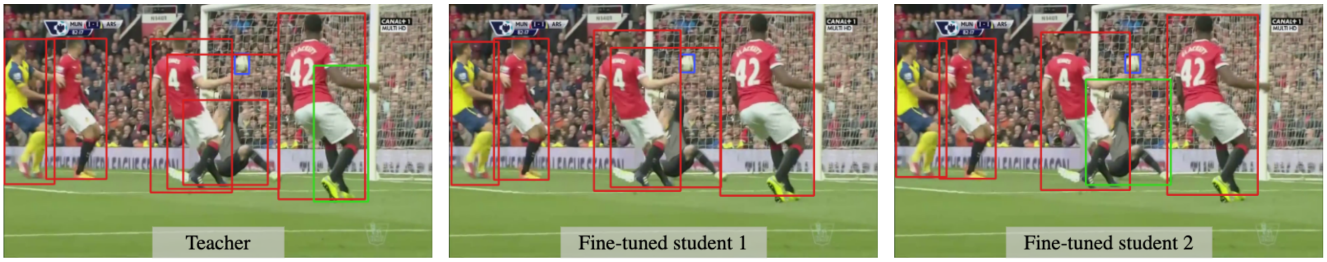


Figure 5. **Qualitative results.** Comparison of the detections on a test set image for the first teacher (left), fine-tuned student model after 1 iteration (middle), and fine-tuned student model after 2 iterations (right). The considered labeled dataset size is 10%, with 10 extra unlabeled games, using the third parametrization for both iterations, with the optimal threshold values presented in Table 1.

Table 4. **Optimal threshold values for the third parametrization.** Comparison of the performance of the third parametrization before fine-tuning for different threshold values $\tau_l$ when $\tau_h = 1$, on 10% of the labeled data and 10 extra games as unlabeled data. The performance of the student increases with $\tau_l$ showing that only high confidence samples should be considered.

| $\tau_l$ | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|
| mAP | 39.1 | 40.1 | 40.8 | 41.3 | **41.9** |

ues for $\tau_l$ also lead to better performance. This means that the transition zone between true negatives and positives is around high confidence scores. In other words, detected objects with confidence scores lower than 0.8 should be considered as negative samples rather than being ignored. By construction, this observation is dependent on the considered network architecture and dataset. However, it provides good insights on how we should consider the Faster R-CNN predictions based on their confidence scores. We can also observe that a very high value for $\tau_l$ and $\tau_h$ reduces the performance of the student.

For the *third parametrization*, we also study the influence of $\tau_l$ and $\tau_h$ on the performance. From our experiments, we noticed that the best performance is always obtained when choosing $\tau_h = 1$. This means that we should increasingly give credit to the predictions based on their confidence score with no upper limit, independently of the value of $\tau_l$. Therefore, we show the performance when varying $\tau_l$ only for this optimal threshold ($\tau_h = 1$). As can be seen in Table 4, the performance increases with the value of $\tau_l$, showing that we should consider predictions with a higher prediction score than before ($\tau_l = 0.9$). In fact, the predictions between the thresholds are not completely ignored compared to the second parametrization, but are simply less considered when approaching $\tau_l$.

Finally, since our method may also be used in an iterative fashion, we provide some insights on to what extend a second iteration of pseudo-labelling using the first student as the new teacher and training a second student further improve the performance. In particular, we study the iterative process with 10% of the labeled dataset and the third parametrization since it gives good performance for one iteration and that its training time is reasonable. As mentioned earlier in Table 1, for this setup, the first teacher and the first fine-tuned student have performances of 39.5% and 43.7%, respectively. After fine-tuning, the second student model reaches an mAP of 45.1%, which further increases the performance compared to the teacher and the first student. In further work, we will study more deeply our iterative process, especially when considering the whole labeled and unlabeled dataset, which is computationally intensive.

**Qualitative results.** Illustrations of our method's predictions for consecutive iterations are shown in Figure 5 for the first teacher, the first student, and the second student. As can be seen, the first student does not produce false positives, unlike the teacher, but fails at correctly localizing and classifying the goalkeeper. However, the second student manages to correctly detect the goalkeeper. This perfectly illustrates the detection improvements at each iteration.

## 5. Conclusion

In this work, we propose a new generic semi-supervised method based on a teacher-student approach for object detection. In particular, we show how unlabeled data improves the detection performance of a model trained solely on labeled data. Our method consists in using a teacher trained on labeled data to produce surrogate ground-truth annotations on the unlabeled dataset, later added to the labeled data to train a student model. To adapt the training process to our scenario, we propose three loss parametrizations based on the confidence score of the teacher's predictions to introduce doubt. By doing so, our method substantially improves the performance compared to supervised training. A side result is that we set the first detection benchmark on the new SoccerNet-v3 dataset. Since our method is data and network agnostic, we presume that it is always possible to use available unlabeled data, a common situation in sports analysis, to further improve a detection network.

## References


[1] Adrià Arbués Sangüesa, Adriàn Martín, Javier Fernández, Coloma Ballester, and Gloria Haro. Using player's body-orientation to model pass feasibility in soccer. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Workshops (CVPRW)*, pages 3875–3884, Seattle, WA, USA, June 2020. 2

[2] M. Archana and M. Geetha. An efficient ball and player detection in broadcast tennis video. In *Intelligent Systems Technologies and Applications*, volume 384 of *Adv. in Intell. Syst. and Comput.*, pages 427–436. Springer, 2015. 2

[3] David Berthelot, Nicholas Carlini, Ekin Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. ReMixMatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *Int. Conf. on Learn. Rep. (ICLR)*, Addis Abada, Ethiopia, Apr.-May 2020. 3

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. MixMatch: A holistic approach to semi-supervised learning. In *Adv. in Neural Inform. Process. Syst. (NeurIPS)*, volume 32, Vancouver, Canada, Dec. 2019. Curran Associates, Inc. 3

[5] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-v3: Scaling up soccernet with multi-view spatial localization and re-identification. *Submitted to Scientific Data*, 2022. 2, 5

[6] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Floriane Magera, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. Camera calibration and player localization in SoccerNet-v2 and investigation of their representations for action spotting. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Workshops (CVPRW), CVsports*, pages 4537–4546, Nashville, TN, USA, June 2021. 2

[7] Anthony Cioppa, Adrien Deliège, Noor Ul Huda, Rikke Gade, Marc Van Droogenbroeck, and Thomas B. Moeslund. Multimodal and multiview distillation for real-time player detection on a football field. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Workshops (CVPRW), CVsports*, pages 3846–3855, Seattle, WA, USA, June 2020. 2, 5

[8] Anthony Cioppa, Adrien Deliège, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. ARTHuS: Adaptive real-time human segmentation in sports through online distillation. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Workshops (CVPRW), CVsports*, pages 2505–2514, Long Beach, CA, USA, June 2019. 2

[9] Adrien Deliège, 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 understanding of broadcast soccer videos. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Workshops (CVPRW), CVsports*, pages 4508–4519, Nashville, TN, USA, June 2021. Best CVSports paper award. 2

[10] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *Int. J. Comp. Vis.*, 88(2):303–338, June 2010. 2

[11] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Workshops (CVPRW)*, pages 1711–1721, Salt Lake City, UT, USA, June 2018. 2, 5

[12] Ross Girshick. Fast R-CNN. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 1440–1448, Santiago, Chile, Dec. 2015. 2

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 580–587, Columbus, OH, USA, June 2014. 2

[14] Christina Gough. Market size of the sports analytics industry worldwide in 2020 and 2028, 2021. https://www.statista.com/statistics/1185536/sports-analytics-market-size/. 1

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 2980–2988, Venice, Italy, Oct. 2017. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. 5

[17] Samuel Hurault, Coloma Ballester, and Gloria Haro. Self-supervised small soccer player detection and tracking. In *Int. ACM Workshop Multimedia Content Anal. in Sports (MMSports)*, pages 9–18, Seattle, WA, USA, Oct. 2020. 2

[18] Mordor Intelligence. Sports analytics market – Growth, trends, COVID-19 impact, and forecasts (2022 - 2027), 2022. https://www.mordorintelligence.com/industry-reports/sports-analytics-market. 1

[19] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Adv. in Neural Inform. Process. Syst. (NeurIPS)*, volume 32, Vancouver, Canada, Dec. 2019. Curran Associates, Inc. 3

[20] Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. SoccerDB: A large-scale database for comprehensive video understanding. In *Int. ACM Workshop Multimedia Content Anal. in Sports (MMSports)*, pages 1–8, 2020. 2

[21] Paresh R. Kamble, Avinash G. Keskar, and Kishor M. Bhurchandi. A deep learning ball tracking system in soccer videos. *Opto-Electronics Review*, 27(1):58–69, Mar. 2019. 2

[22] DTAI Sports Analytics Lab. Why sports analytics, 2019. https://dtai.cs.kuleuven.be/sports/. 1

[23] Yandong Li, Di Huang, Danfeng Qin, Liqiang Wang, and Boqing Gong. Improving object detection with selective self-supervised self-training. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 12374 of *Lect. Notes Comp. Sci.*, pages 589–607. Springer, Oct. 2020. 3

[24] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 2117–2125, Honolulu, HI, USA, July 2017. 2, 5

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 2, 3

[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 *Eur. Conf. Comput. Vis. (ECCV)*, volume 8693 of *Lect. Notes Comp. Sci.*, pages 740–755. Springer, Sept. 2014. 2

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander Berg. SSD: Single shot multibox detector. *CoRR*, abs/1512.02325, 2016. 2

[28] Yang Liu, Luiz Hafemann, Michael Jamieson, and Mehrsan Javan. Detecting and matching related objects with one proposal multiple predictions. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. Workshops (CVPRW)*, pages 4515–4522, Nashville, TN, USA, June 2021. 2

[29] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Pzizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Int. Conf. on Learn. Rep. (ICLR)*, May 2021. 3, 4

[30] Mehrtash Manafifard, Hamid Ebadi, and Hamid Abrishami Moghaddam. A survey on player tracking in soccer videos. *Comp. Vis. and Image Underst.*, 159:19–46, June 2017. 2

[31] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Adv. in Neural Inform. Process. Syst. (NeurIPS)*, volume 31, Montréal, Canada, Dec. 2018. Curran Associates, Inc. 6

[32] 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*, 6:1–15, Oct. 2019. 2

[33] Hieu Pham, Zihang Dai, QIzhe Xie, Minh-Thang Luong, and Quoc Le. Meta pseudo labels. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 11557–11568, Nashville, TN, USA, June 2021. 3

[34] Miran Pobar and Marina Ivasic-Kos. Mask R-CNN and optical flow based method for detection and marking of handball actions. In *Int. Congress on Image and Signal Process., BioMedical Eng. and Inform. (CISP-BMEI)*, pages 1–6, Beijing, China, Oct. 2018. 2

[35] Upendra M. Rao and Umesh C. Pati. A novel algorithm for detection of soccer ball and player. In *Int. Conf. Commun. and Signal Process. (ICCSP)*, pages 344–348, Melmaruvathur, India, Apr. 2015. 2

[36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, June 2016. 2

[37] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *CoRR*, abs/1804.02767, Apr. 2018. 2

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, June 2017. 2, 3, 5

[39] Melike Sah and Cem Direkoglu. Evaluation of image representations for player detection in field sports using convolutional neural networks. In *International Conference on Theory and Application of Fuzzy Systems and Soft Computing (ICAFS)*, volume 896 of *Adv. in Intell. Syst. and Comput.*, pages 107–115. Springer, 2018. 2

[40] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Han Kurakin, Alexand Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Adv. in Neural Inform. Process. Syst. (NeurIPS)*, volume 33, pages 596–608. Curran Associates, Inc., Dec. 2020. 3

[41] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *CoRR*, abs/2005.04757, 2020. 3, 4

[42] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and efficient object detection. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 10778–10787, Seattle, WA, USA, June 2020. 2

[43] Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *IEEE Winter Conf. Applicat. Comp. Vis. (WACV)*, pages 2291–2301, Waikoloa, HI, USA, Jan. 2021. 3, 4

[44] Graham Thomas, Rikke Gade, Thomas B. Moeslund, Peter Carr, and Adrian Hilton. Computer vision for sports: current applications and research topics. *Comp. Vis. and Image Underst.*, 159:3–18, June 2017. 2

[45] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *IEEE Int. Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, pages 10684–10695, Seattle, WA, USA, June 2020. 3

[46] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 3060–3069, Montréal, Canada, Oct. 2021. 3, 4

[47] Yukun Yang, Min Xu, Wanneng Wu, Ruiheng Zhang, and Yu Peng. 3D multiview basketball players detection and localization based on probabilistic occupancy. In *Digit. Image Comp.: Tech. and Applicat.*, pages 1–8, Canberra, ACT, Australia, Dec. 2018. 2

[48] Junqing Yu, Aiping Lei, Zikai Song, Tingting Wang, Hengyou Cai, and Na Feng. Comprehensive dataset of broadcast soccer videos. In *IEEE Conf. on Multimedia Inform. Process. and Retrieval (MIPR)*, pages 418–423, Miami, FL, USA, June 2018. 2

[49] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D. Cubuk, and Quoc V. Le. Rethinking pre-training and self-training. In *Adv. in Neural Inform. Process. Syst. (NeurIPS)*, volume 33, pages 3833–3845. Curran Associates, Inc., Dec. 2020. 3