No Bells, Just Whistles: Sports Field Registration by Leveraging Geometric Properties

Marc Gutiérrez-Pérez and Antonio Agudo
Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain

Abstract

Broadcast sports field registration is traditionally addressed as a homography estimation task, mapping the visible image area to a planar field model, predominantly focusing on the main camera shot. Addressing the shortcomings of previous approaches, we propose a novel calibration pipeline enabling camera calibration using a 3D soccer field model and extending the process to assess the multiple-view nature of broadcast videos. Our approach begins with a keypoint generation pipeline derived from SoccerNet dataset annotations, leveraging the geometric properties of the court. Subsequently, we execute classical camera calibration through DLT algorithm in a minimalist fashion, without further refinement. Through extensive experimentation on real-world soccer broadcast datasets such as SoccerNet-Calibration, WorldCup 2014 and TS-WorldCup, our method demonstrates superior performance in both multiple- and single-view 3D camera calibration while maintaining competitive results in homography estimation compared to state-of-the-art techniques. [1]

1. Introduction

Camera calibration is essential for a wide range of computer vision applications, such as structure from motion [2, 4, 12, 13], reconstruction [3, 24], and pose estimation [9, 36]. In the sports domain, accurately estimating pairwise correspondences between the sports field and broadcast video is crucial for sorting out some high-level tasks. This process not only streamlines manual labor but also enhances the visual appeal of broadcast matches through augmented reality and virtual advertisement insertion, while also facilitating the development of advanced tools for sports analytics.
Sports fields, with their well-defined dimensions [1], serve as calibration objects. However, achieving accurate camera calibration in the broadcast setting poses challenges due to multiple camera views and partial occlusion of the court, hindering the matching process between 2D and 3D correspondences. Additionally, the variability of the camera focal length further complicates calibration efforts. The recent surge in deep learning has led to several data-driven approaches for field-specific feature prediction [7, 9, 19, 25, 26] or direct homography matrix regression [21, 30]. Others investigate camera calibration as a search problem [6, 27, 28, 34, 35], generating camera pose databases and refining homography estimates to improve calibration accuracy. Moreover, some approaches [9, 10, 15, 16, 25] leverage temporal homography consistency between video frames, intending to better align with the nature of sports video broadcasts. Focusing on the soccer domain, despite the potential of estimating camera parameters for reconstructing non-planar points and enabling applications such as automatic camera control, offside detection or 3D ball tracking, previous studies [6, 7, 9, 21, 25–29, 34, 35] have predominantly treated the task as homography estimation rather than full calibration [31]. Inspired by the limitations of existing approaches, we propose a novel geometry-based keypoint retrieval pipeline (see Figs. 1-2) for 3D sports field registration, additionally capable of addressing the challenges posed by the multiple-view broadcast nature. This approach involves defining a hierarchical pipeline to extract a pre-defined keypoint grid from the court's geometric properties and leveraging an encoder-decoder neuronal to estimate keypoint positions. Specifically, we adopt HRNetv2 [32] as the backbone model for the keypoints prediction. Finally, the estimated keypoints are used to compute the projection matrix using RANSAC [14] and Direct Linear Transformation (DLT) [18] algorithms. The pipeline of our proposed method is outlined in Fig. 1. We extensively evaluate our approach on three real-world soccer broadcast datasets, (SoccerNet-Calibration [8], WorldCup 2014 [19] and TS-WorldCup [7]) and compare it with state-of-the-art methods in both 2D and 3D sports field registration. The experiments demonstrate that our model achieves superior performance on 3D camera calibration while maintaining comparable results on homography estimation with respect to competing approaches.

[1] https://github.com/mguti97/No-Bells-Just-Whistles

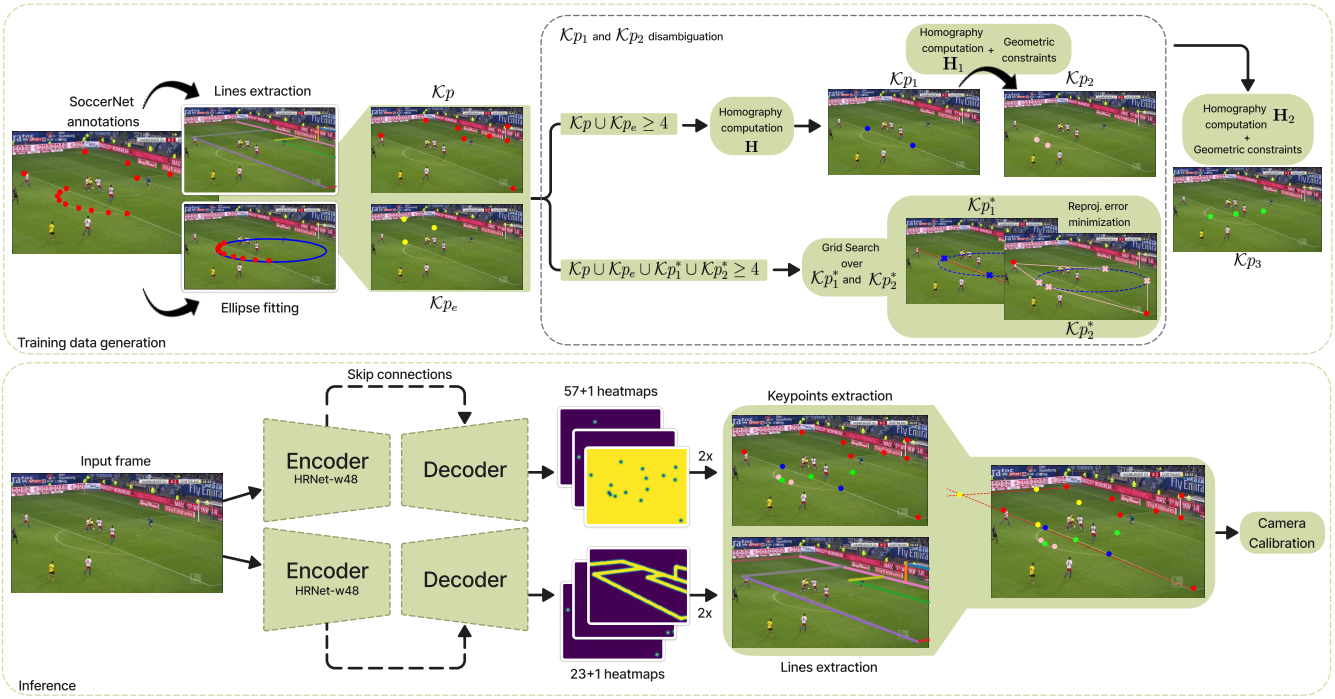

Figure 1. **Overview of our proposed framework. Top:** Training data generation pipeline. Beginning with SoccerNet [8] annotations, we utilize field line extraction and ellipse fitting to establish a hierarchical structure for computing each set of keypoints. **Bottom:** The encoder-decoder networks produce heatmaps for keypoints and extremities of soccer field lines to extract their positions in the image space. The obtained keypoint set is augmented with intersections of lines generated by the second model to ensure a sufficient number of points.

In a nutshell, this paper makes the following contributions:

- A novel geometry-based keypoints grid and a robust pipeline for their retrieval.
- A calibration pipeline capable of integrating non-planar points for 3D camera calibration and extending to multiple views from the broadcast.
- A pipeline structure focused solely on 2D-3D correspondences, without further refinement. In other words, a minimalist approach without bells and whistles—or, in soccer terms, no bells, just whistles.

2. Related Work

Sports field registration is a critical component of most sports applications in computer vision, whose common approaches intend to estimate homography matrices in team sports. Homography estimation has traditionally relied on the corresponding features, or keypoints, identification between images and the court field model. These keypoints, usually obtained by exploiting geometric primitives such as lines and/or circles, are subsequently used to estimate the mapping between the images by using RANSAC [14] algorithm with DLT [18] or non-linear optimization through the minimisation of a chosen loss function. More recent approaches either directly predict an initial homography matrix. Alternatively, they seek the optimal matching homography within a reference database that includes synthetic images with known homography matrices or camera parameters.

Homography Estimation. Various recent methods for sports field registration obtain an initial homography estimation through a grid of uniformly sampled and predicted keypoints [7, 9, 10, 20, 22, 25, 26], or line and circle pixels using semantic segmentation [19, 27, 30, 31, 34, 35]. These are obtained with Deep Neural Networks (DNN) and used as corresponding features between the field template and camera image. Following a prediction-based strategy, [21, 31] proposed an optimization-based framework to obtain homography estimation and camera parameters by minimizing image registration and segment reprojection error, respectively. Shi *et al.* [29] proposed an iterative estimation process to estimate any homography transformation, regardless of the degree of misalignment between the image and the template, in a self-supervised manner. Moreover, end-to-end approaches [30] have been proposed with promising results. Search-based approaches involve creating databases and searching for the best matching homography through the usage of edge maps [6, 28] or semantic segmentation [27, 34, 35]. However, while search-based methods are highly accurate and robust, they often incur a signif-

icant computational cost due to their time-consuming processing steps. Furthermore, recent approaches [9, 10, 25] leverage broadcast sequentiality to enforce temporal consistency between subsequent frames' homographies.

Homography Refinement. Homography refinement is a crucial step in camera calibration, aiming to achieve an even more accurate homography estimation, if necessary. Previous works use one or a combination of the following methods to refine the initial estimate. [21] and [29] directly regress the refined homography from the input image and field model with DNNs. [6, 28] obtained an estimate for the camera pose by matching with a feature-pose database, and used 100k templates to ensure the assumption of small transform between the input image and the matched template. Iterative optimisation of the camera pose or homography is also proposed in previous works [6, 21, 31], commonly based on re-projection or registration error. Moreover, other approaches [25, 27] exploit the temporal consistency between subsequent video frames.

Camera Parameters Retrieval. Approaches using feature-pose synthetic databases [6, 27, 28, 34, 35] allows for homography and/or direct camera parameters retrieval as templates are created through projective geometry. However, due to small transform assumption, smaller databases lead to larger reprojection errors, making homography refinement a crucial step. Carr *et al.* [5] estimate camera's extrinsic and intrinsic parameters leveraging gradient-based alignment to edge images. Moreover, homography decomposition [18] allows for access to individual camera parameters, as shown in [9, 31].

3. Methodology

Sports TV broadcasts consist of sequences of images featuring a fraction of the sports field from an uncalibrated moving camera perspective. Our approach focuses on both extrinsic and intrinsic camera parameters retrieval from each individual frame without any prior information about the camera position or orientation. The proposed method comprises four processing components: Soccer field modelling and keypoint generation, keypoint and line detection, DLT algorithm, and camera parameters retrieval. Next, these components are introduced.

3.1. Modelling the Soccer Field

A soccer field is composed of lines and circle segments, representing all field markings, goal posts, and crossbars. Our approach, like keypoint-based methods [7, 9, 10, 22, 25, 26], relies on the lines painted on the ground, their intersections, and the corners they define, due to its known position on the world coordinate system. We follow the segment definitions of Cioppa *et al.* [8], and set them as starting points to hierarchically compute our pre-defined keypoints from court geometric properties.

3.1.1 Keypoint Generation

The full set of sampled keypoints is organized into subgroups based on the specific geometric features they represent (Fig. 2). The hierarchical nature of the computation ensures that information from initially identified keypoints is exploited for computing the subsequent ones (some instances in Fig. 1-top). Next, we define the keypoints sets:

- **Line-Line intersections.** This set of points ($\mathcal{K}p$) includes the intersections of boundary lines or the penalty area markings. Considering the 23 lines depicted in [8], including goal posts and crossbars, up to 30 points can be included.
- **Extended Line-Line intersections.** This set ($\mathcal{K}e$) addresses the intersections of extended lines that represent non-adjacent segments of the soccer field. To obtain that, the lines are extended by exploiting their equations beyond their original boundaries.
- **Line-Ellipse intersections.** This set ($\mathcal{K}p_1$) is to consider the intersections between the field lines and the circles or semi-circles present on the court. Given the distortions introduced by the camera perspective, conics on the field are considered ellipses for equation computation. The parameters of these ellipses are fitted using the least squares method [17]. Intersection points were analytically derived using ellipse and line formulas.
- **Ellipse tangent points.** Augmentation of available points is achieved through the utilization of tangent points on tangent lines, extending from a specified point to the previously defined ellipses. These tangent points (denoted by $\mathcal{K}p_2$) were analytically determined by employing an ellipse equation, incorporating the known coordinates of an external point.
- **Additional points.** Once the previous points are inferred, and the corresponding homography, an additional set ($\mathcal{K}p_3$) of 9 points is integrated along the central pitch axis, encompassing the pitch center and penalty points. Additionally, 4 points are strategically placed to designate quarter turns along the central circle. Furthermore, the homography facilitates the inclusion of other points that are initially missing, addressing situations such as unannotated lines.

3.1.2 Keypoint Disambiguation

Due to the multi-view nature of the SoccerNet dataset [11] and, considering, for instance, one of the soccer field's semi-circles, ambiguity appears in its respective $\mathcal{K}p_1$ and $\mathcal{K}p_2$ keypoints candidates, as shown in Fig. 2-bottom. To avoid that, we define two different strategies to handle that disambiguation depending on the total number of keypoints generated in the previous sets: when there are sufficient points in the $\mathcal{K}p \cup \mathcal{K}e$ set to infer a homography, i.e., four points, $\mathcal{K}p_1$ is computed first by choosing the can-

didates combination that minimizes a reprojection error. Then, we include $\mathcal{K}p_1$ in the homography estimation and apply the same strategy to $\mathcal{K}p_2$. Otherwise, we perform a grid-search involving both $\mathcal{K}p_1$ and $\mathcal{K}p_2$ candidates when $\mathcal{K}p \cup \mathcal{K}p_e \cup \mathcal{K}p_1^* \cup \mathcal{K}p_2^* \geq 4$, where $*$ denotes a possible candidate combination. The grid-search iterates over all keypoints candidates in a set-wise manner to avoid unfeasible combinations and keeps the one with minimum reprojection error. It is worth noting that when none of the strategies can be applied, the current frame is deemed invalid for calibration purposes. Moreover, once we define the new sets, two additional constraints are applied in the case homography estimation or ellipse fitting is not accurate enough. Initially, we manually establish a reprojection error threshold to validate points. Subsequently, through iteration over combinations of keypoints, we construct vectors and ensure that the cross-products maintain consistent signs in both world and image coordinates. This final step is essential in cases where two distinct combinations yield valid top- and bottom-view perspectives of the field while exhibiting identical reprojection errors. Utilizing cross-products enables us to differentiate and retain the keypoint combination corresponding to the field's top-view. The full keypoint generation process is depicted in Fig. 1-top.

3.1.3 Left-Right Disambiguation

In sequences where the camera angle aligns with the longitudinal axis of the court, an ambiguity arises regarding the distinction between the right and left halves of the field. Hence, a critical step to ensure consistency and robustness across keypoints and lines detection processes involves differentiating between the two sides. This is accomplished by implementing a remap to the ground-truth (GT) values, ensuring that the goal area closest to the camera consistently represents the left side. The process to check whether or not the mapping should be applied is defined in a heuristic fashion. We compute angles of horizontal and vertical soccer field lines, respectively, and then set a threshold taking into account angle distribution and visual inspection.

3.2. Keypoints and Lines Detection

Our approach makes use of two encoder-decoder convolutional neural networks to estimate the position of the pre-defined keypoints and the soccer field lines depicted in [8] excluding conics, giving the last one an auxiliary role to enhance keypoint completeness. During inference, the former produces heatmaps for each pre-defined keypoint with a single Gaussian peak with $2px$ sigma positioned in the keypoint location, accompanied by an additional target channel. This additional channel reflects the inverse of the maximum value among the other target feature maps, ensuring that the resultant target tensor behaves as a probability distribution

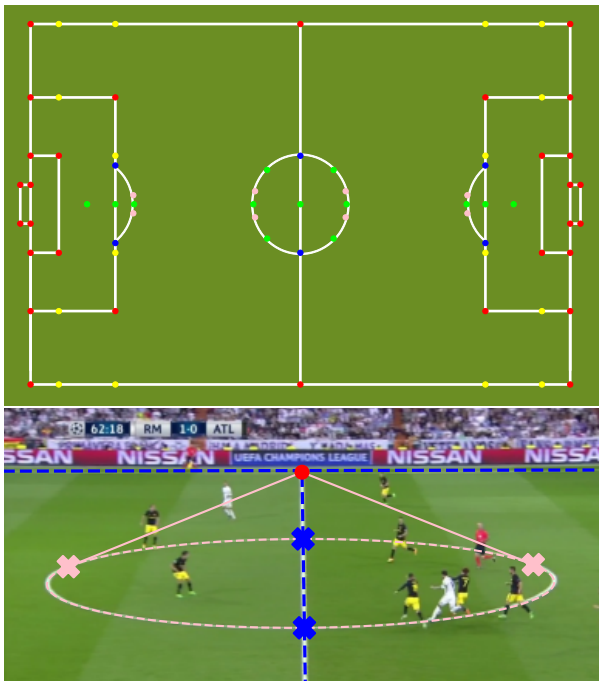

Figure 2. **Definition of keypoint positions on a soccer field. Top:** Distribution of points on a zenithal view, including all the relevant locations as a result of intersecting lines or curves in the field. $\mathcal{K}p$, $\mathcal{K}e$, $\mathcal{K}p_1$, $\mathcal{K}p_2$ and $\mathcal{K}p_3$ point sets are displayed in red, yellow, blue, pink and green points, respectively. **Bottom:** Given an external point, both $\mathcal{K}p_1$ and $\mathcal{K}p_2$ candidates are analytically derived, marked as blue and pink crosses, respectively.

function at every spatial point. Meanwhile, the latter network produces heatmaps for each visible soccer field line within the frame, assigning two Gaussian peaks at the line extremities' locations. Additionally, we introduce an extra channel, known as the boundary channel, to our heatmap following the approach outlined in [33]. This augmentation aims to enhance the efficient capture of global information regarding the soccer field and improve extremities detection, particularly near image borders. We effectively extract the positions of keypoints and line extremities from the generated heatmaps by employing a max pooling operation, drawing inspiration from the methodology proposed in [38]. This process is summarized on Fig. 1-bottom.

3.2.1 Architecture

The keypoint and line extremities detection utilized a modified HRNetV2-w48 [32] as the encoder's backbone network. HRNetv2 [32] is a new family of convolutional networks that maintains high-resolution representations through the whole process resulting in semantically richer and spatially more precise representations. To improve the spatial resolution of the predicted heatmaps, we incorpo-

rated $2\times$ upsampling and concatenated skip-connection features from the corresponding resolution of the convolution stem to fuse the features at different scales. The final predictions exhibit half the resolution of the original image, with softmax employed as the final activation function.

3.2.2 Keypoints Mask

When homography was unavailable due to a limited number of points, the heatmaps associated with points belonging to $\mathcal{K}p_1$, $\mathcal{K}p_2$, and $\mathcal{K}p_3$—which would have been indicated by homography—were masked out from the loss function as long as the line to which they belong is included in the GT annotation. Additionally, when the external point required to compute ellipse tangent points in $\mathcal{K}p_2$ is not present in $\mathcal{K}p$, the pair of tangent points candidates is also excluded.

3.3. Camera Projection Model

We employ a standard full perspective camera model:

$$\mathbf{P} = \mathbf{K}\mathbf{R}[\mathbf{I} \,|\, -\mathbf{c}] \in \mathbb{R}^{3\times 4}, \tag{1}$$

where $\mathbf{K} \in \mathbb{R}^{3\times 3}$ denotes the intrinsics to transform from camera coordinates to image ones, and $\mathbf{R} \in \mathbb{R}^{3\times 3}$ and $\mathbf{c} \in \mathbb{R}^3$ determine the extrinsics (rotation and translation) to map from scene coordinates to camera ones. Following [18], we assume zero skew and a known pixel aspect ratio. Additionally, for simplicity, we assume the principal point coincides with the center of the image, and we neglect astigmatism or distortions.

3.3.1 Camera Parameters Estimation

Extrinsic and intrinsic parameters are inferred by leveraging the coordinates of 3D object points and their corresponding 2D projections using the soccer field model as a calibration rig, following [37], which consists of a closed-form solution followed by a non-linear refinement based on the maximum likelihood criterion. To calibrate the 3D soccer field model rig, we also consider two additional vertical planes containing the goal polygons, including non-planar points such as keypoints belonging to the goal posts and crossbars. This strategy enhances completeness by providing estimations when insufficient points are on the ground plane.

To account for keypoint misdetections and other complexities in camera parameter retrieval, such as frames with only one non-planar keypoint visible, the calibration process was repeated on several subsets of keypoints. These subsets were selected based on various heuristics: *full-keypoints*, including all keypoints sets $\mathcal{K}p$, $\mathcal{K}p_e$, $\mathcal{K}p_1$, $\mathcal{K}p_2$ and $\mathcal{K}p_3$; *main-keypoints*, comprising only line-line intersections from the original SoccerNet annotations [8]; and *ground-plane-keypoints*, which excludes non-planar keypoints. Furthermore, we applied a grid of RANSAC [14] reprojection error thresholds to each subset. The final camera calibration values were determined through a heuristic voting process, prioritizing the method yielding a lower reprojection error, with emphasis on the *full-keypoints* subset.

3.3.2 Homography Estimation

Assuming the world coordinate system such that $z = 0$ corresponds to the ground plane, the ground-to-image homography $\mathbf{H}$ can be obtained from the first, second, and fourth columns of the camera projection matrix $\mathbf{P}$. Nevertheless, inaccurate estimations for keypoints associated with the non-planar rig, such as those belonging to the goal posts and crossbars, may result in a flawed homography estimation. To address this issue, we employ classical homography estimation with DLT [18] and RANSAC on the *ground-plane-keypoints* subset. We define a maximum allowable reprojection error to consider a point pair as an inlier and subsequently refine the computed homography matrix using the Levenberg-Marquardt method [18] on the point correspondences and initial homography estimation.

4. Experiments

We implement our proposed method for sports field registration. This section provides an overview of the datasets utilized, the evaluation metrics employed to assess our approach, and implementation specifics. Subsequently, we present both qualitative and quantitative results, comparing our method with state-of-the-art approaches.

4.1. Datasets

To evaluate our method, we follow state-of-the-art methods by using the SoccerNet Calibration [8], the WorldCup [19] and the TS-WorldCup [7] soccer datasets.

SN-Calib Dataset: The SoccerNetV3-Calibration (SN23) dataset [8] comprises 22,816 images extracted from SoccerNet [11] videos and encompasses a broadcast-based multi-view nature, offering a broader range of camera perspectives beyond the main broadcast camera. Cioppa *et al.* [8] provide annotations for all segments of the soccer field, encompassing lines, conics and goal posts. For each visible segment on the court, at least two annotated positions are provided, optimally representing the segment in a polyline format. For the conics drawn on the pitch, the annotations consist of a list of points that roughly give the circle shape when connected. Additionally, Theiner *et al.* [31] provided manual camera view annotations for the SoccerNetv3-Calibrationim 2022 (SN22) dataset version, allowing the creation of data subsets taking into account the camera view distribution.

WC14 Dataset: The WorldCup 2014 dataset (WC14) [19] stands as the reference benchmark for sports field registration and consists of 209 images from ten

games for training and 186 images from other ten games for testing and the corresponding manually annotated homography matrices from the FIFA WorldCup 2014. Additionally, Theiner *et al.* [31] provides segment annotations in SN-Calib [8] format.

TS-WC Dataset: The TS-WorldCup dataset (TSWC) [7] contains detailed field markings on 3,812 field images from 43 videos of Soccer WorldCup 2014 and 2018 in a time-sequence fashion, which is ten times larger than the WorldCup 2014 dataset.

4.2. Evaluation Metrics

The quality of estimated camera parameters or homography matrices can be evaluated both in 2D image and world spaces.

Accuracy@threshold (Acc@t): The evaluation relies on calculating the reprojection error between each annotated point and the line to which it belongs. Adopting a binary classification approach, each pitch segment is treated as a single entity. To be considered correctly detected, all points within the segment must have a reprojection error smaller than a threshold. The projection of pitch elements from densely sampled points of the soccer field 3D model yields a polyline for each segment. Therefore, a polyline representing a soccer field segment s is classified as a true positive (TP) if $\forall p \in s : \min(d(p, \hat{s})) < t$, being $\hat{s}$ the corresponding annotated segment and t the distance threshold in pixels. Otherwise, this segment is counted as a false positive (FP). Segments only present in the annotations are counted as false negatives (FN). Hence, the Accuracy for a threshold of t pixels is given by $Acc@t = TP/(TP+FN+FP)$. We also measure the completeness rate (CR) as the number of camera parameters provided divided by the number of images with more than four semantic line annotations in the dataset. The final score (FS) as an evaluation criterion is calculated as the product of CR and Acc@5.

Intersection over Union (IoU): The intersection over union (IoU) includes two components: IoU_{part} quantifies the visible area of the video frame by warping the video frame using both the refined homography and the GT homography, projecting them onto the template, and then calculating the IoU. IoU_{whole} evaluates the entire sports field by warping the template with the refined homography, projecting it onto the original template, and calculating the IoU.

Projection Error: The projection error was quantified as the average distance, in meters, between the projected points using the predicted homography and the GT homography. To achieve this, we uniformly sampled $2,500$ pixels from the visible field area of the camera image and projected them onto the field to compute the distance. The standard dimensions of a soccer field are 105×68 meters.

Reprojection Error: The reprojection error was calculated by averaging the distance between the reprojected points in the video frame, utilizing both the predicted and the GT homography.

4.3. Implementation Details

Due to the absence of publicly available results on the multiple-view SN23 distribution, we trained two models from scratch: Multi-view (MV) and Single-view (SV). The latter is composed almost entirely of non-replay frames, ensuring a high percentage of central camera shots. We train separate networks for the keypoints and line extremities detection tasks on the SN23-train dataset [8]. For the MV model, we train for 200 epochs, using an initial learning rate of $1e^{-2}$ and a batch size of 2. For the SV model, we train for 200 epochs, using an initial learning rate of $1e^{-5}$ and a batch size of 2. We utilize the Adam optimizer with default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. l_2-loss is used for heatmap regression in both neural networks. Data augmentation such as random horizontal flip, color jitter, and Gaussian noise are applied to enhance model robustness and generalization. Furthermore, we fine-tune both SV networks on the WC14 and TSWC datasets. GT homographies are transformed into our proposed keypoint sets and line extremities by projecting their respective world coordinates to the field's ground plane. The experiments are conducted on a single NVIDIA GeForce RTX 2080 Ti GPU with 12 GB of memory, and the implementation is carried out in the PyTorch framework.

4.4. Results and comparisons

In this section, we present the results of an extensive evaluation, divided into camera calibration and homography estimation. The former assesses the accuracy of individual camera parameters using Acc@t metric, while the latter evaluates the quality of homography estimation using IoU metrics, projection error, and reprojection error.

4.4.1 Camera Calibration

In team sports such as soccer, the action takes place on a nearly planar field. Consequently, most methods utilize homography estimation to map all elements positioned on this plane but cannot project non-planar points such as points belonging to goal posts or crossbars. Conversely, Theiner *et al.* [31] employs a 3D model of the soccer field to extract camera pose and intrinsic parameters directly. In homography-based approaches, parameter retrieval is accomplished through homography decomposition (HDecomp). We conduct a quantitative comparison of our proposed method with respect to state-of-the-art approaches [6, 21, 31] on the SN22-test-center dataset, comprising only images where the main camera center is vis-

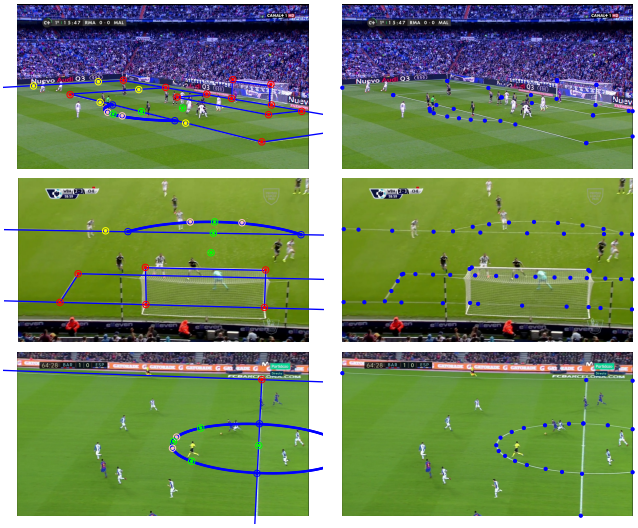

Figure 3. **Qualitative results of our MV model on SN23-test. Left:** Projection of soccer field lines and goal posts from world to image coordinates using predicted camera parameters. Blue lines correspond to segment projections, and colored points represent predicted keypoints (along with auxiliary points retrieved from line extremities detection). **Right:** SN23-test dataset annotations, where each soccer field line is delineated by a point set.

ible (1,454 images). Furthermore, utilizing the SoccerNet annotation format for the WC14-test dataset provided by Theiner *et al.* [31], we conduct a comparison of our proposed method's performance in camera parameter estimation on the WorldCup 2014 dataset distribution.

We report the statistics from [31] paper for the results of state-of-the-art approaches [6, 21, 31]. As shown in Tables 1-2 for SN22-test and WC14 datasets, respectively, our SV method outperforms state-of-the-art approaches on several metrics for both datasets. Minor variations in CR arise due to our approach requiring a minimum number of visible keypoints for calibration.

Finally, we evaluate our method on the entire SN23-test dataset, which defines the first public camera calibration benchmark on the SoccerNetV3-Calibration [8] dataset, and the first one extending calibration assessments to encompass multiple-view scenarios.

4.4.2 Homography Estimation

The proposed method is compared to state-of-the-art approaches [6, 7, 9, 21, 23, 25, 26, 29, 31, 34, 35] using the WC14-test dataset. Additionally, our method is also compared to state-of-the-art approaches [6, 7, 23, 25, 26] using the TSWC-test dataset. For computing IoU, projection error, and reprojection error, we adopt the approach outlined in [7]. The dimensions of the soccer field template are set at 115×74 yards for fair comparison.

Dataset	Approach	Acc@t [%] 5	10	20	CR	FS
SN23-test	Ours_{MV}	73.7	86.7	90.4	77.5	57.1
SN22-test -center	[6] + HDecomp	34.4	64.6	81.3	66.6	22.9
	TVCalib (τ) [31]	57.6	81.7	93.2	93.7	53.9
	TVCalib [31]	54.8	78.5	90.4	**100.0**	54.8
	Ours_{SV}	**75.3**	**89.4**	**91.1**	97.8	**73.7**

Table 1. **Evaluating camera calibration on SoccerNet distributions.** We assess our Multi-view model on the entire SN23-test dataset and quantitatively compare our Single-view model with other approaches on the SN22-test-center dataset.

Approach	Acc@t [%] 5	10	20	CR	FS
[6] + HDecomp	32.7	67.3	87.3	81.7	26.7
[21] + HDecomp	36.9	66.4	83.9	84.9	31.3
TVCalib (τ) [31]	41.3	73.6	91.4	95.7	39.5
TVCalib [31]	39.9	71.9	90.5	**100.0**	39.9
Ours_{SV}	**80.4**	**91.1**	**94.2**	99.5	**80.0**

Table 2. **Quantitative comparison** of our Single-view model on camera calibration conducted on the WC14-test dataset.

Regarding the evaluation on the WC14-test dataset, we present the performance metrics based on the findings reported in the respective papers of the state-of-the-art approaches, as detailed in Table 3. To ensure a fair comparison with [9], we use the results of the approach "ours-w/o-players" reported in their paper. Our fine-tuned model achieves competitive results in comparison with other methods, which are topped by [26, 29], on most of the metrics. Notably, we attain state-of-the-art performance in the median value of the reprojection error without requiring further homography refinement, unlike [6, 7, 26].

For the evaluation on the TSWC-test dataset, we report the statistics from the papers [7, 23, 26] in Table 3, observing how our fine-tuned model outperforms those methods in most of the metrics, once again demonstrating the effectiveness of our approach without the need for further homography refinement.

4.4.3 Ablation Study on Keypoint Sets Contribution

The contribution of each keypoint set, namely $\mathcal{K}p_e$, $\mathcal{K}p_1$, $\mathcal{K}p_2$, and $\mathcal{K}p_3$, is analyzed in Table 4. The integration of $\mathcal{K}p_1$ notably enhances accuracy, particularly along the midfield line, as no points were available in the previous sets. The inclusion of $\mathcal{K}p_2$ augments the completeness rate by

Dataset	Approach	$\text{IoU}_{\text{part}} \uparrow$ (%)		$\text{IoU}_{\text{whole}} \uparrow$ (%)		Proj. $\downarrow$ (m)		Reproj. $\downarrow$	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
WC14-test	Chen *et al.* [6]	94.5	96.1	89.4	93.8	-	-	-	-
	Jiang *et al.* [21]	95.1	96.7	89.8	92.9	-	-	-	-
	Citraro *et al.* [9]	-	-	90.5	91.8	-	-	0.018	0.012
	Zhang *et al.* [34]	95.9	97.5	91.4	94.2	-	-	-	-
	Nie *et al.* [25]	95.9	97.1	91.6	93.4	0.84	0.65	0.019	0.014
	Shi *et al.* [29]	96.6	97.8	**93.1**	**94.8**	-	-	-	-
	Chu *et al.* [7]	96.0	97.0	91.2	93.1	0.81	0.63	0.019	0.014
	Zhang et al [35]	95.9	97.3	91.4	94.1	-	-	-	-
	Maglo el at [23]	96.3	97.4	92.0	94.1	0.74	0.55	0.018	0.014
	Oo *et al.* [26]	**96.9**	**97.9**	92.9	94.6	**0.65**	**0.46**	**0.016**	0.012
	Theiner *et al.* [31]*	95.3	96.6	-	-	-	-	-	-
	$\text{Ours}^{*}_{\text{SV}}$	94.4	96.9	89.4	93.0	1.23	0.58	0.026	0.014
	$\text{Ours}^{*\dagger}_{\text{SV}}$	96.2	97.8	92.2	94.3	0.68	**0.46**	**0.016**	**0.011**
TSWC-test	Chen *et al.* [6]‡	96.8	97.4	90.7	94.1	0.54	0.38	0.016	0.013
	Nie *et al.* [25]‡	97.4	97.8	92.5	94.2	0.43	0.38	0.011	0.010
	Chu *et al.* [7]‡	98.1	98.2	94.8	95.4	0.36	0.33	0.009	0.008
	Maglo *et al.* [23]‡	98.3	98.5	95.7	96.2	0.26	0.23	0.008	0.006
	Oo *et al.* [26]‡	98.5	98.7	95.8	96.7	0.26	0.21	0.007	0.006
	$\text{Ours}^{*}_{\text{SV}}$	97.9	98.5	94.4	95.7	0.30	0.24	0.008	0.006
	$\text{Ours}^{*\ddagger}_{\text{SV}}$	**98.6**	**98.8**	**96.3**	**96.8**	**0.23**	**0.20**	**0.005**	**0.005**

Table 3. **Evaluating the homography estimation on WC14-test and TSWC-test.** $*$ denotes the methods trained on SoccerNet distribution, † denotes the methods fine-tuned on the WC14 dataset and ‡ denotes the methods fine-tuned on the TSWC one.

$\mathcal{K}p_e$	$\mathcal{K}p_1$	$\mathcal{K}p_2$	$\mathcal{K}p_3$	Acc@t [%] 5	10	20	CR	FS
✗	✗	✗	✗	66.8	73.9	91.9	85.9	57.4
✓	✗	✗	✗	66.9	86.0	91.9	85.9	57.5
✓	✓	✗	✗	74.6	89.1	92.1	91.8	68.6
✓	✓	✓	✗	73.8	87.8	91.1	96.5	71.3
✓	✓	✓	✓	75.3	89.4	91.1	97.8	73.7

Table 4. **Ablation study of our keypoint sets.** The table shows the effect of every keypoint set on the SN22-test-center dataset.

increasing the keypoint density within the field circles but shows a small decrease in accuracy metrics. This is due to the still low number of keypoints across these field circles; we are adding new low-quality detections, hence decreasing the overall accuracy. Furthermore, the integration of $\mathcal{K}p_3$ further elevates accuracy and completeness rate, ultimately defining our robust keypoints grid. These conclusions can be straightforwardly seen in the field model grid presented in Fig. 2. Although $\mathcal{K}p_e$ does not contribute to a score boost, its significance lies in increasing the number of points to allow the computation of the subsequent sets during the training data generation.

5. Conclusion

In this paper, we introduce a novel framework for 3D sports field registration. Our proposed pipeline adopts a minimalist approach by solely utilizing the geometric properties of the soccer field. We demonstrate superior performance in 3D camera calibration on SoccerNet and WorldCup 2014 datasets compared to state-of-the-art methods, while also achieving comparable results in homography estimation on WorldCup 2014 and TS-WorldCup datasets. Additionally, we extend our pipeline to multiple-view camera calibration, thereby establishing the first public multiple-view broadcast-based camera calibration benchmark in soccer. Our method exhibits promising results, highlighting the effectiveness of utilizing a robust field template without the need for further refinements. As long as the video frame distortions are not too harsh (i.e. fisheye shots) and a minimum of four keypoints are visible, sports field registration is shown to be effective. In future work, we plan to enhance our approach by incorporating temporal consistency between subsequent video frames, aligning better with the nature of sports video broadcasts.

Acknowledgment. This work has been supported by the project MoHuCo PID2020-120049RB-I00 funded by MCIN/ AEI /10.13039/501100011033.

References

[1] Fifa football stadiums guidelines. https://publications.fifa.com/en/football-stadiums-guidelines/. Accessed: 2024-02-26. 1

[2] Antonio Agudo. Total estimation from RGB video: Online camera self-calibration, non-rigid shape and motion. In *ICPR*, 2020. 1

[3] Antonio Agudo, Begoña Calvo, and Jose Montiel. FEM models to code non-rigid EKF monocular SLAM. *ICCVW*, 2011. 1

[4] Antonio Agudo, Francesc Moreno-Noguer, Begoña Calvo, and Jose Montiel. Sequential non-rigid structure from motion using physical priors. *TPAMI*, 38(5):979–994, 2016. 1

[5] Peter Carr, Yaser Sheikh, and Iain Matthews. Point-less calibration: Camera parameters from gradient-based alignment to edge images. In *WACV*, 2012. 3

[6] Jianhui Chen and James J Little. Sports camera calibration via synthetic data. In *CVPRW*, 2019. 1, 2, 3, 6, 7, 8

[7] Yen-Jui Chu, Jheng-Wei Su, Kai-Wen Hsiao, Chi-Yu Lien, Shu-Ho Fan, Min-Chun Hu, Ruen-Rone Lee, Chih-Yuan Yao, and Hung-Kuo Chu. Sports field registration via keypoints-aware label condition. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 8

[8] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up soccernet with multi-view spatial localization and re-identification. *Scientific data*, 9(1):355, 2022. 1, 2, 3, 4, 5, 6, 7

[9] Leonardo Citraro, Pablo Márquez-Neila, Stefano Savare, Vivek Jayaram, Charles Dubout, Félix Renaut, Andres Hasfura, Horesh Ben Shitrit, and Pascal Fua. Real-time camera pose estimation for sports fields. *MVA*, 31:1–13, 2020. 1, 2, 3, 7, 8

[10] Paul J Claasen and JP de Villiers. Video-based sequential bayesian homography estimation for soccer field registration. *arXiv preprint arXiv:2311.10361*, 2023. 1, 2, 3

[11] 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 understanding of broadcast soccer videos. In *CVPR*, 2021. 3, 5

[12] Felix Endres, Jürgen Hess, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. 3D mapping with an RGB-D camera. *TRO*, 30(1):177–187, 2013. 1

[13] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*, 2014. 1

[14] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 2, 5

[15] Bernard Ghanem, Tianzhu Zhang, and Narendra Ahuja. Robust video registration applied to field-sports video analysis. In *ICASSP*, 2012. 1

[16] Ankur Gupta, James J Little, and Robert J Woodham. Using line and ellipse features for rectification of broadcast hockey video. In *CRV*, 2011. 1

[17] Radim Halır and Jan Flusser. Numerically stable direct least squares fitting of ellipses. In *WSCG*, pages 125–132, 1998. 3

[18] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 2, 3, 5

[19] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun. Sports field localization via deep structured models. In *CVPR*, 2017. 1, 2, 5

[20] Nicolas Jacquelin, Romain Vuillemot, and Stefan Duffner. Efficient one-shot sports field image registration with arbitrary keypoint segmentation. In *ICIP*, 2022. 2

[21] Wei Jiang, Juan Camilo Gamboa Higuera, Baptiste Angles, Weiwei Sun, Mehrsan Javan, and Kwang Moo Yi. Optimizing through learned errors for accurate sports field registration. In *WACV*, 2020. 1, 2, 3, 6, 7, 8

[22] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. Kalicalib: A framework for basketball court registration. In *MMW*, 2022. 2, 3

[23] 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. 7, 8

[24] Etienne Mouragnon, Maxime Lhuillier, Michel Dhome, Fabien Dekeyser, and Patrick Sayd. Real time localization and 3D reconstruction. In *CVPR*, 2006. 1

[25] Xiaohan Nie, Shixing Chen, and Raffay Hamid. A robust and efficient framework for sports-field registration. In *WACV*, 2021. 1, 2, 3, 7, 8

[26] Yin May Oo, Ankhzaya Jamsrandorj, Vanyi Chao, Kyung-Ryoul Mun, and Jinwook Kim. A residual attention-based efficientnet homography estimation model for sports field registration. In *IECON*, 2023. 1, 2, 3, 7, 8

[27] Long Sha, Jennifer Hobbs, Panna Felsen, Xinyu Wei, Patrick Lucey, and Sujoy Ganguly. End-to-end camera calibration for broadcast videos. In *CVPR*, 2020. 1, 2, 3

[28] Rahul Anand Sharma, Bharath Bhat, Vineet Gandhi, and CV Jawahar. Automated top view registration of broadcast football videos. In *WACV*, 2018. 1, 2, 3

[29] Feng Shi, Paul Marchwica, Juan Camilo Gamboa Higuera, Michael Jamieson, Mehrsan Javan, and Parthipan Siva. Self-supervised shape alignment for sports field registration. In *WACV*, 2022. 1, 2, 3, 7, 8

[30] Shuhei Tarashima. SFLNet: direct sports field localization via cnn-based regression. In *ACPR*, 2020. 1, 2

[31] Jonas Theiner and Ralph Ewerth. Tvcalib: Camera calibration for sports field registration in soccer. In *WACV*, 2023. 1, 2, 3, 5, 6, 7, 8

[32] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 43(10):3349–3364, 2020. 1, 4

[33] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, 2019. 4

[34] Neng Zhang and Ebroul Izquierdo. A high accuracy camera calibration method for sport videos. In *VCIP*, 2021. 1, 2, 3, 7, 8

[35] Neng Zhang and Ebroul Izquierdo. A four-point camera calibration method for sport videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 2, 3, 7, 8

[36] Zhengyou Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *ICCV*, 1999. 1

[37] Zhengyou Zhang. A flexible new technique for camera calibration. *TPAMI*, 22(11):1330–1334, 2000. 5

[38] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 4