Camera Calibration in Sports Broadcasting

A PhD Dissertation by Floriane Magera

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (PhD) in Engineering Science

Marc Van Droogenbroeck – Supervisor
Olivier Barnich – Co-Supervisor

Doctoral College in Electricity, Electronics and Computer science
December 2026

A word about the front page...

Science and education should be fun and accessible. If you are fortunate enough to study a topic that brings joy to millions of people, then you might as well enjoy yourself along the way. I certainly did. As someone who has often been the only woman in the room throughout her studies and career, I wanted my front page to express both my personality and my experience.

Jury Members

Pr. Bernard Boigelot — President
Department of Electrical Engineering and Computer Science,
University of Liège (ULiège), Belgium

Dr. Olivier Barnich — Co-supervisor
EVS, Liège, Belgium

Pr. Anthony Cioppa
Department of Electrical Engineering and Computer Science,
University of Liège (ULiège), Belgium

Pr. Javier Civera
Engineering Research Institute of Aragon (I3A),
University of Zaragoza, Spain

Pr. Christophe De Vleeschouwer
Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM),
Catholic University of Louvain (UCLouvain), Belgium

Pr. Marc Van Droogenbroeck — Supervisor
Department of Electrical Engineering and Computer Science,
University of Liège (ULiège), Belgium

Abstract

Camera calibration, or the process by which we identify how an image was formed, is a key component in the analysis and enrichment of sports videos, as it bridges the images to the 3D geometry of the sports scene. This problem is extensively studied and is at the core of many applications. Yet in the context of this industrial thesis, we found a gap between the demanding environment of broadcasting and the existing research. Indeed, sports scenes present several challenges for classical camera calibration solutions: scenes are captured from distant viewpoints and contain both texture-less and repetitive patterns like grass, brand advertisements, and seats in stadiums. On the other hand, the advantage of sports is the presence of sports field markings, which offer a convenient calibration target. However, each sport needs a specific detector and dedicated datasets to learn the detection of these markings in challenging conditions. The problem is that we found the ground truth of available datasets incompatible with the level of precision required in an industrial environment. Through this thesis, we have addressed these gaps, namely by working on ways to annotate datasets in the absence of physical ground truth for camera parameters, and by proposing a metric that is compatible with industrial needs within the limits of available data. We also proposed different methods that advance the state of the art, and importantly, that are grounded in the physical realities of broadcast cameras. In a context where everything is easily learned, we investigated how geometry can be of use when our practical scenarios reach the limits of the existing datasets.

Finally, through collaborations with industry, and our strong community involvement in the SoccerNet initiative, we enforced our vision and requirements to the scientific community, and thus enabled new research directions in the realm of sports analytics that are aligned with sports broadcasters' needs.

Through its contributions and by laying the groundwork for future research, this thesis advances camera calibration for sports and opens new perspectives for robust geometric understanding of sports scenes.

Acknowledgements

We thank the *Service Public de Wallonie (SPW) Recherche* of Belgium for supporting this work with the Grant N° 8573.
These acknowledgements will be updated.

Contents

Abstract ii

Acknowledgements iii

List of Abbreviations vii

Notation viii

1 Introduction 1
- 1.1 Context . . . 1
- 1.2 Outline . . . 2

2 A Guide to Camera Calibration 5
- 2.1 Modeling a camera . . . 5
  - 2.1.1 Pinhole camera . . . 5
  - 2.1.2 Homographies . . . 7
  - 2.1.3 Lens distortion . . . 9
- 2.2 Estimating camera parameters . . . 9
  - 2.2.1 From correspondences . . . 10
  - 2.2.2 Retrieval-based camera calibration . . . 11
  - 2.2.3 Absolute pose regression . . . 12
  - 2.2.4 Perspective and field-of-view fields . . . 13
  - 2.2.5 Geometry foundation models . . . 13
  - 2.2.6 Conclusion . . . 14

3 Benchmarking Camera Calibration in Sports (1^{st} article) 15
- 3.1 Introduction . . . 17
- 3.2 Related work . . . 20
  - 3.2.1 Datasets to benchmark camera calibration . . . 20
  - 3.2.2 Evaluation of camera calibration in sports . . . 21

3.3 A model-agnostic camera calibration benchmark . . . 24
3.3.1 Annotations: beyond homographies . . . 24
3.3.2 Evaluation metric . . . 26
3.4 Results . . . 28
3.4.1 Description of the datasets . . . 28
3.4.2 A better camera model for broadcast cameras . . . 28
3.5 Discussion . . . 30
3.6 Conclusion . . . 31

4 Broadcast Camera Tracking (2^{nd} article) 33
4.1 Introduction . . . 35
4.2 Related work . . . 37
4.3 Method . . . 39
4.3.1 Broadcast camera model . . . 39
4.3.2 Complete description of our tracking system . . . 41
4.3.3 Reinitialization algorithm . . . 44
4.4 Results . . . 45
4.4.1 Tracking system on SoccerNet-gamestate . . . 45
4.4.2 Reinitialization on SoccerNet-calibration . . . 47
4.4.3 Ablation study . . . 47
4.4.4 Physical soundness . . . 48
4.4.5 Long-term tracking . . . 48
4.5 Conclusions . . . 49

5 Leveraging Circular Markings (3^{rd} article) 53
5.1 Introduction . . . 54
5.2 Related work . . . 56
5.3 Method . . . 59
5.3.1 Case 1: Line correspondence . . . 61
5.3.2 Case 2: Point correspondence . . . 61
5.3.3 Case 3: Unknown center . . . 62
5.4 Experiments . . . 63
5.4.1 PnLCalib keypoint detector . . . 63
5.4.2 Qualitative results . . . 64
5.4.3 Quantitative Results . . . 64
5.4.4 Unknown center . . . 68

5.5 Discussion . . . 71
5.6 Conclusion . . . 71

6 Sport-agnostic Metric Rectification of Sports Field (4^{th} article) 72
6.1 Introduction . . . 74
6.2 Related Work . . . 76
6.2.1 Sports camera calibration . . . 76
6.2.2 Single-image camera calibration . . . 76
6.2.3 Monocular depth estimation . . . 77
6.2.4 Line and vanishing point detection . . . 77
6.2.5 Rectification . . . 78
6.3 Method . . . 78
6.3.1 Problem statement . . . 78
6.3.2 Pipeline overview . . . 79
6.3.3 Plane segmentation from depth models . . . 79
6.3.4 Parallel lines detection . . . 80
6.3.5 Focal length from parallel lines . . . 80
6.3.6 Focal length aggregation . . . 81
6.4 Experiments . . . 82
6.4.1 Experimental setup . . . 82
6.4.2 Baselines . . . 84
6.4.3 Main results . . . 84
6.4.4 Qualitative results . . . 85
6.5 Discussion . . . 86
6.6 Conclusion . . . 87

7 Conclusion 89
7.1 Contributions . . . 89
7.2 Future Work . . . 90
7.3 Perspectives . . . 90

Bibliography 92

List of Abbreviations

APR Absolute Pose Regression. 12, 57, 58

AR augmented reality. 19, 40, 73, 78, 82, 83, 90

DLT Direct Linear Transform. 10, 11, 59

LM Levenberg-Marquardt. 11, 13

MDE Monocular Depth Estimation. 13, 77

OOD Out-of-Domain. 63, 65–67

PCA Principal Component Analysis. 79

PnP Perspective-n-Point. 10, 37, 55, 59

RANSAC RAndom SAmple Consensus. 10, 11, 38, 45, 69

VP Vanishing Point. 80

Notation

This section summarizes the main mathematical notation used throughout this thesis.

General mathematical notation

Symbol	Description
$\langle \mathbf{a}, \mathbf{b} \rangle$	Dot product of vectors $\mathbf{a}$ and $\mathbf{b}$.
$\mathbf{a} \times \mathbf{b}$	Cross product of vectors $\mathbf{a}$ and $\mathbf{b}$.
$\|\mathbf{x}\|_2$	Euclidean norm, or ℓ_2 norm, of vector $\mathbf{x}$.
$\mathbb{P}$	Projective space.
$\mathbf{x}$	Vector, represented by a bold symbol.
$\mathtt{M}$	Matrix, represented by a typewriter symbol.
$\mathtt{I}$	Identity matrix.
$\mathtt{M}^\mathsf{T}$	Transpose of matrix $\mathtt{M}$.
$\mathtt{M}^{-1}$	Inverse of matrix $\mathtt{M}$.
$\widehat{x}$	Estimate of a quantity x.

Camera geometry and calibration

Symbol	Description
κ	Camera or camera parameterization.
$\mathtt{P}$	Camera projection matrix.
$\mathtt{K}$	Camera calibration matrix.
$\mathtt{R}$	Camera rotation matrix.
$\mathbf{t}$	Camera translation vector.
$\mathbf{C}$	Camera optical center, or focal point.

Symbol	Description
$\mathbf{T}$	Rotation center of a camera mounted on a tripod.
C_x	Coordinate of the camera optical center along axis x.
$\mathbf{p} = (p_x, p_y)$	Principal point and its image coordinates.
f_x, f_y	Horizontal and vertical focal lengths.
s	Skew coefficient of the intrinsic calibration matrix.
ϕ, θ, γ	Camera pan, tilt, and roll angles, respectively.
t	Time or frame index.
$\mathbf{X}$	Three-dimensional point.
$\mathbf{x}$	Two-dimensional image point.
$\pi_\kappa(\mathbf{X})$	Projection of the 3D point $\mathbf{X}$ into the image by camera κ.
x, y	Coordinates in the image plane.
$\overline{x}, \overline{y}$	Coordinates in the normalized image plane.
k_1, k_2	First- and second-order radial-distortion coefficients.

Sports-field representation

Symbol	Description
$\mathcal{F}$	Geometric template of the sports field.

Classification notation

Symbol	Description
TP	Number of true positives.
TN	Number of true negatives.
FP	Number of false positives.
FN	Number of false negatives.

Evaluation metrics

Symbol	Description
IoU	Intersection over Union.
IoU_{part}	Sports field registration metric: intersection over union computed over the visible field region.
$\text{IoU}_{\text{whole}}$	Sports field registration metric: intersection over union computed over the entire field region.
JaC	Jaccard-based camera calibration metric.
JaC_{τ}	Jaccard-based camera calibration metric evaluated with threshold τ.
MRE	Mean reprojection error.

Chapter 1

Introduction

1.1 Context

This thesis is the result of an industrial doctorate in collaboration between the Faculty of Applied Sciences of the University of Liège and EVS, a Belgian company, and a world leader in broadcasting technology. EVS is renowned for its replay servers; they are used around the world to record the camera streams from *live* events like sports, and to transmit them *live* to end-watchers. This industrial thesis is an attempt at leveraging recent advances in computer vision to enable the creation of value for EVS, while contributing to the state of the art.

As a company operating in the broadcasting industry, EVS has access to a high volume of video data. These videos originate from a wide range of sports productions and include not only the main broadcast feed, *i.e.* the program seen by all on television, but also the complete set of camera streams available during an event. This access to the full set of cameras capturing an event is a key advantage for enriching the content of EVS clients. These enrichment possibilities include automatic tagging of events, summarization, cropping, reframing, and tracking of the ball and players in two dimensions and even in three dimensions. These features, in turn, enable advanced applications such as tactical analysis, virtual cameras, free-viewpoint videos, or 3D replays.

This access to data makes EVS particularly well positioned to recover three-dimensional information from the various camera feeds recorded on its servers. This is a differentiating factor and, at the same time, a demanding position, as EVS will work with any production's cameras without assuming much control or information about the setup, which are the constraints for real-world production considered in this work. In addition, cost considerations prevent dedicated acquisition efforts, meaning that we must rely on existing footage captured by unknown cameras, under

unknown settings, and for purposes unrelated to computer vision.

Under these constraints, extracting valuable three-dimensional information about the game nevertheless remains possible through geometric vision techniques such as camera calibration and triangulation. These methods transform raw images into quantitative representations of the scene, enabling the recovery of player positions, object trajectories, and spatial relationships. In practice, camera calibration constitutes the cornerstone of spatial game understanding, as it provides the geometric foundation upon which all subsequent reasoning can be extracted from images.

At the beginning of this thesis, we observed that this fundamental problem had received limited attention from the research community. Camera calibration is rarely an end goal in itself and is more often an intermediate step to achieve higher-level tasks such as player tracking, pose estimation, or tactical analysis. Consequently, small calibration errors are often tolerated, provided that the final application remains competitive on benchmark datasets.

In a broadcasting context, however, the requirements are considerably more demanding. While an average calibration result may have little impact on academic benchmark scores, even a small geometric misalignment can become immediately visible when graphics, virtual advertising, offside lines, or other augmented-reality elements are overlaid onto a live broadcast viewed by millions of spectators. In this setting, calibration accuracy is not merely a benchmarking metric but a prerequisite for maintaining the quality and credibility of the production. This difference in priorities creates a gap between research objectives and industrial needs.

In this thesis, we sought to address this gap. Beyond proposing novel methods and analyses, we believe that our involvement in the research community has contributed to renewed interest in the problem and helped raise awareness of the importance of accurate camera calibration for sports broadcasting applications.

Through this work, we asked ourselves how best to evaluate camera calibration given the available datasets and pseudo-ground truth; which camera model represents best the cameras used in live production; how to exploit different types of geometric correspondences; and how to enable content enrichment without sport-specific knowledge, relying solely on the planarity of the sports field and the presence of straight line markings.

1.2 Outline

This cumulative thesis presents four papers written as first author, either published or under review. Except where explicitly noted, all code and experiments were

developed exclusively by the author. First, as a prelude to our contributions, we give a concise overview of camera calibration in **Chapter 2**.

Chapter 3 challenges current benchmarking procedures in sports environments and introduces an alternative benchmarking protocol. The chapter presents the following publication: Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. "A Universal Protocol to Benchmark Camera Calibration for Sports". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, Washington, USA: IEEE, June 2024, pp. 3335–3346.

Chapter 4 is an attempt at modeling broadcast cameras and addresses camera calibration over time, following the camera modeling work initiated in the previous chapter. For the results presented in this chapter, a part of the code regarding the optical flow was written by Thomas Hoyoux. The chapter presents the publication: Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. "BroadTrack: Broadcast Camera Tracking for Soccer". In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Tucson, Arizona, USA: IEEE, Feb. 2025, pp. 6177–6187.

In **Chapter 5**, we address an edge case of camera calibration in sports, where we can leverage the presence of circular field markings. **Chapter 5** explores the use of these particular geometric primitives and reports the work published in the following paper: Floriane Magera, Thomas Hoyoux, Martin Castin, Olivier Barnich, Anthony Cioppa, and Marc Van Droogenbroeck. "Can Geometry Save Central Views for Sports Field Registration?" In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, Tennessee, USA: IEEE, June 2025, pp. 6060–6069.

Chapter 6 integrates geometric foundation models in camera calibration for sports, with the challenge of being agnostic to the sport at hand. This last chapter is about a paper that is currently under review: Floriane Magera, Olivier Barnich, Marc Van Droogenbroeck, and Daniel Barath. *Beyond Field Registration: Sports-Agnostic Metric Rectification from a Single Image*. Submitted to a conference. 2026.

Other collaborations not reported in this thesis are:

- Anthony Cioppa et al. "Camera Calibration and Player Localization in SoccerNet-v2 and Investigation of their Representations for Action Spotting". In: *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), CVsports*. Nashville, Tennessee, USA, June 2021, pp. 4532–4541.

  For this paper investigating the action spotting potential uncovered by players

localization on the sports field, the first step to obtain the player tracking data was to calibrate the cameras of the SoccerNet dataset. My contribution was to generate a dataset of camera calibrations with EVS' certified camera calibration product [6] and then to train an existing camera calibration method [7] on this pseudo-ground truth. The camera calibration code is available.

- Vladimir Somers et al. "SoccerNet Game State Reconstruction: End-to-End Athlete Tracking and Identification on a Minimap". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, Washington, USA: IEEE, June 2024, pp. 3293–3305
  For the Game State Reconstruction challenge of SoccerNet, I was in charge of providing the players positions in the tactical view of the football field, which required calibrating all images of the dataset and ensuring the smoothness of the camera parameters to retrieve valid player positions.

- Silvio Giancola et al. "SoccerNet 2022 Challenges Results". In: *International ACM Workshop on Multimedia Content Analysis in Sports (MMSports)*. Lisbon, Portugal: ACM, Oct. 2022, pp. 75–86
  I was the task lead for both field localization and camera calibration challenges. I provided baseline and evaluation code, and oversaw both challenges organization.

- Anthony Cioppa et al. "SoccerNet 2023 challenges results". In: *Sports Engineering* 27.2 (July 2024), pp. 1–18
  The same challenges were reconducted; I was supervising the annotation of more frames for both challenges and oversaw the organizational details of the challenges.

Finally, several of my works are available on GitHub. The list is the following:

- https://github.com/evs-broadcast/BroadTrack provides the code of BroadTrack [2].

- https://github.com/SoccerNet/sn-calibration provides the metric defined in our benchmarking protocol [1], and the baseline provided for SoccerNet camera calibration and field-marking detection challenges.

- https://github.com/evs-broadcast/CCBV provides the camera calibration algorithm used in Cioppa et al. [5], which is an implementation of [7].

Chapter 2

A Guide to Camera Calibration

Camera calibration is the process of estimating the parameters of a model that characterizes how an image was formed. These parameters depend on the camera model, which is the mathematical definition of how a three-dimensional scene is turned into a two-dimensional plane, the image. Among various standard camera models (*e.g.* pinhole, orthographic), the pinhole model is assumed throughout this thesis, as it applies to the vast majority of real-world images.

Understanding how images are formed is a way to interpret the three-dimensional world from this condensed information. This is the whole purpose of this thesis: enabling interpretation and image augmentation through implicit three-dimensional world understanding.

2.1 Modeling a camera

2.1.1 Pinhole camera

The pinhole camera model forms images by projecting light rays through a single point, known as the pinhole, onto an image plane. The physical phenomenon underlying this model, referred to as the camera obscura (see Fig. 2.1), has been known since antiquity and can easily be observed in nature through small openings, such as gaps between tree leaves or a hole in a dark enclosure.

In this model, light rays are assumed to travel along straight lines. When a scene is observed through a small aperture in an otherwise dark chamber, an inverted image of the scene is formed on the opposite side of the pinhole. The size of the projected image depends on the geometric configuration, including the distance between the pinhole and the image plane.

This geometric construction can be formalized using similar triangles. As illus-

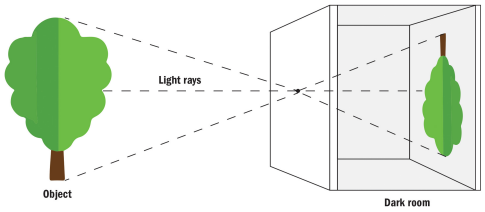


Figure 2.1: Camera obscura: a dark chamber projecting an inverted image through a small hole. *Source: camera-obscura.*

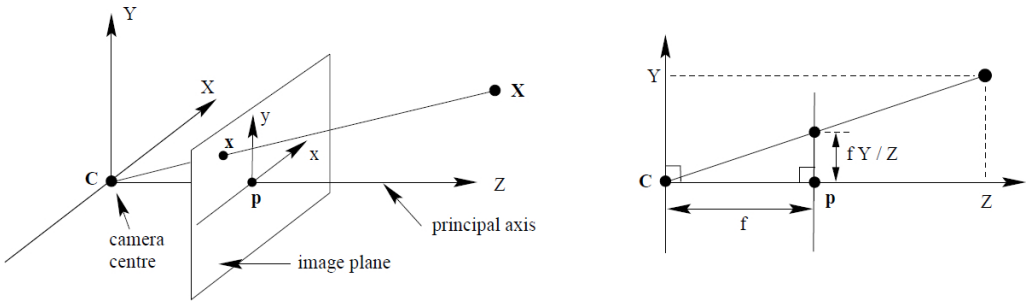


Figure 2.2: Pinhole camera. A point $\mathbf{X}$ is imaged on the image plane at the point $\mathbf{x}$, its coordinates are obtained using similar triangles. Image taken from Hartley and Zisserman [11].

trated in Fig. 2.2, a three-dimensional point $\mathbf{X} = (X, Y, Z)$ is projected onto the image plane at position $\mathbf{x} = (x, y)$ according to

$$\begin{aligned} x &= f\frac{X}{Z}\,, \\ y &= f\frac{Y}{Z}\,, \end{aligned} \tag{2.1}$$

where f denotes the distance between the pinhole and the image plane, commonly referred to as the focal length. These equations establish the fundamental relationship between a three-dimensional scene and its two-dimensional image, forming the basis of camera calibration. These projection equations are not linear because of the division by Z. To obtain linear equations, it is convenient to introduce homogeneous coordinates. A two-dimensional point (x, y) is represented by any triplet (u, v, w) such that $x = u/w$ and $y = v/w$, while a three-dimensional point (X, Y, Z) can be represented as $(X, Y, Z, 1)$, in its normalized form. In this representation, image formation can be expressed as a matrix multiplication defined up to an arbitrary scale factor.

Accounting for the fact that the image origin is commonly located at its upper left corner, we introduce a new quantity, the principal point $\mathbf{p} = (p_x, p_y)$, which is

the point of intersection of the principal axis, the Z-axis, or look-at orientation of the camera and the image plane. Equation (2.1) can then be rewritten as:

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \overbrace{\begin{bmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix}}^{\mathtt{K}} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}. \tag{2.2}$$

This transformation is carried out by the calibration matrix $\mathtt{K}$, whose parameters are called the *intrinsic* camera parameters.

Up to this point, the three-dimensional point coordinates were expressed in the system of axes relative to the camera. In sports and multi-view setups, it is convenient to express 3D coordinates in a common 3D world reference system. This requires applying a rotation and a translation to a 3D point to convert it to the camera reference system before projecting it to the image plane. With the camera center, or focal point, expressed in the world reference system by $\mathbf{C}$, the rotation $\mathtt{R}$ and translation $\mathbf{t} = -\mathtt{R}\mathbf{C}$ form the *extrinsic* camera parameters. Together, the intrinsic and extrinsic parameters can be combined and form the projection matrix $\mathtt{P}$:

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \mathtt{KR}(\mathbf{X} - \mathbf{C}) = \mathtt{K}(\mathtt{R}\mathbf{X} \overbrace{-\mathtt{R}\mathbf{C}}^{\mathbf{t}}) = \overbrace{\mathtt{K}[\mathtt{R}|\mathbf{t}]}^{\mathtt{P}} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}. \tag{2.3}$$

In the end, the coordinates in the image can be obtained from the homogeneous coordinates $(x, y) = (u/w, v/w)$.

2.1.2 Homographies

Another essential mathematical element to introduce is the homography: *a homography is an invertible mapping h from the projective plane $\mathbb{P}^2$ (homogeneous 3-vectors) to itself such that three points $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$ lie on the same line if and only if $h(\boldsymbol{x}_1)$, $h(\boldsymbol{x}_2)$ and $h(\boldsymbol{x}_3)$ do* [11, p. 32]. This property is particularly useful in sports broadcasts, where most semantic information (field markings, player positions, referee locations) lies on the playing surface. As sports fields are mostly planar, it is convenient to use homographies to represent the mapping between the sports field plane and the image plane. Furthermore, this mapping can conveniently be

expressed in terms of camera parameters. If we set the sports field plane as $Z = 0$, then the projection equation becomes:

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \overbrace{\mathtt{K}[\mathbf{r}_1\mathbf{r}_2\mathbf{t}]}^{\mathtt{H}} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} = \mathtt{K}[\mathtt{R}|\mathbf{t}] \begin{pmatrix} X \\ Y \\ 0 \\ 1 \end{pmatrix}, \tag{2.4}$$

with $\mathbf{r}_i$ the i^{th} column vector of the rotation matrix $\mathtt{R}$. The homography matrix is a homogeneous 3×3 matrix that maps 2D points of the sports field plane to their imaged point. As sports fields have fixed dimensions, this homography is a way to transfer 2D points lying on the pitch plane to a common reference system; this homography performs sports field registration.

Through Eq. (2.4), a homography can be constructed from the parameters of a pinhole camera observing a planar scene. While the reverse conversion is also possible, it is generally not lossless in practice. Indeed, an estimated homography does not necessarily satisfy the geometric constraints imposed by the pinhole model, such as the chosen intrinsic parameterization and the orthonormality of the rotation matrix. By using the construction of the homography from Eq. (2.4), we can estimate $\mathtt{K}$ using the algorithm explained by Hartley and Zisserman [11, p. 225], which exploits constraints on the calibration matrix and planar homographies. The rotation matrix $\mathtt{R}$ can be obtained from $\mathtt{K}^{-1}\mathtt{H} \sim [\mathbf{r}_1\mathbf{r}_2\mathbf{t}]$ followed by orthonormalization of the recovered rotation vectors.

Homographies also enable playing with other geometric primitives easily, which is of prime interest in the case of sports broadcasting, with sports field markings being a collection of straight lines, circles, and points of regulated dimensions. Lines in the projective plane $\mathbb{P}^2$ are defined as vectors $\mathbf{l} = (a, b, c)^\mathsf{T}$, and if a point $\mathbf{x} = (x, y, 1)^\mathsf{T}$ belongs to the line $\mathbf{l}$, then $\mathbf{x}^\mathsf{T}\mathbf{l} = ax + by + c = 0$. Through the transformation of a homography, the line $\mathbf{l}$ becomes $\mathbf{l'}$ and points belonging to $\mathbf{l}$ become: $\mathbf{x'} = \mathtt{H}\mathbf{x}$ and by definition, $\mathbf{x'}^\mathsf{T}\mathbf{l'} = \mathbf{x}^\mathsf{T}\mathtt{H}^\mathsf{T}\mathbf{l'} = 0$ hence $\mathbf{l} = \mathtt{H}^\mathsf{T}\mathbf{l'}$, or $\mathbf{l'} = \mathtt{H}^{-\mathsf{T}}\mathbf{l}$.

Finally, for circular field markings, a circle is a particular type of conic. A conic $\mathtt{C}$ is defined by the quadratic equation $ax^2 + dy^2 + 2bxy + 2cx + 2ey + f = 0$ that can be rewritten using homogeneous coordinates as:

$$\mathbf{x}^\mathsf{T}\mathtt{C}\mathbf{x} = \begin{pmatrix} x & y & 1 \end{pmatrix} \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = 0\,. \tag{2.5}$$

Under a homography transformation, the conic equation becomes: $\mathbf{x}'^{\mathsf{T}}\mathtt{C}'\mathbf{x}' = \mathbf{x}^{\mathsf{T}}\mathtt{H}^{\mathsf{T}}\mathtt{C}'\mathtt{H}\mathbf{x} = 0$ and thus $\mathtt{C} = \mathtt{H}^{\mathsf{T}}\mathtt{C}'\mathtt{H}$ or

$$\mathtt{C}' = \mathtt{H}^{-\top}\mathtt{C}\mathtt{H}^{-1}\,. \tag{2.6}$$

2.1.3 Lens distortion

Camera optics induce a form of image distortion due to the physical properties of the camera lenses. There are different types of distortions, but in the context of this thesis, there is one type that we found particularly noticeable in sports content: *radial distortion*. Depending on the field-of-view of the camera, this distortion can be visible in the image, with wider field-of-view values exhibiting stronger effects. In sports, radial distortion is probably the most noticeable distortion type, which causes straight lines to appear curved in images. Its intensity grows with the distance to the center of distortion. Generally, this center of distortion is assumed to be at the image center and thus coincides with the principal point. In this thesis, we use the radial distortion model proposed by Brown [12].

Augmenting our pinhole model with radial distortion, the point $\mathbf{X}$ is expressed in the camera frame: $\mathbf{X}_C = \mathtt{R}(\mathbf{X} - \mathbf{C})$, it is then projected in the normalized image plane $(x, y) = (X_C/Z_C, Y_C/Z_C)$. An image point is distorted according to its squared distance to the image center $r^2 = x^2 + y^2$ and the final distorted image point in pixels (x_d, y_d) can be obtained by:

$$\begin{aligned} x_d &= f(x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6)) + p_x\,, \\ y_d &= f(y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6)) + p_y\,. \end{aligned} \tag{2.7}$$

The coefficients k_i are called the *radial distortion coefficients*. Common radial distortion effects are pincushion, barrel, or even mustache distortion. There exist other types of distortion, such as tangential or thin prism distortion, but their effect is generally negligible in the context of broadcast images.

The careful reader might note that, by their definition, homographies are unable to capture radial distortion. Consequently, significant distortion requires undistorting the image points before applying a homography-based model to describe the image formation process.

2.2 Estimating camera parameters

With the geometric relationship between the 2D image plane and 3D space defined, the subsequent section focuses on the estimation of the underlying intrinsic and

extrinsic camera parameters. Numerous methods exist for camera calibration. In the following sections, we will review a few ways to calibrate a camera *from a single image*.

2.2.1 From correspondences

The classical method to calibrate a camera or to register a sports field is to obtain correspondences: known 3D elements are localized in the image. For points, this would give sets of point pairs $\{\mathbf{X}_i : (X_i, Y_i, Z_i, 1), \mathbf{x}_i : (x_i, y_i, 1)\}$. Each point pair satisfies the following equation:

$$\mathbf{x}_i \times \mathtt{P}\mathbf{X}_i = 0\,. \tag{2.8}$$

Each correspondence pair gives two linearly independent equations. The projection matrix $\mathtt{P}$ is a 3×4 matrix, in homogeneous coordinates, such that one degree of freedom is dropped by the scale ambiguity. Estimating $\mathtt{P}$ requires fixing eleven degrees of freedom, and as each correspondence gives us two equations, six point pairs are required to estimate $\mathtt{P}$. If the configuration is not degenerate, the homogeneous system of equations obtained from Eq. (2.8) in terms of $\mathtt{P}$'s entries can be solved using the Direct Linear Transform (DLT) algorithm.

The DLT is the easiest way to estimate the projection matrix, but it is also the most demanding in terms of correspondence numbers. Indeed, the projection matrix $\mathtt{P}$ captures more complexity than the model allowed by our pinhole camera model, as we assume square pixels and centered principal point, the calibration matrix $\mathtt{K}$ has only one degree of freedom, plus the extrinsic parameters' six degrees of freedom, totaling seven degrees of freedom. There exist other Perspective-n-Point (PnP) solvers, such as P4Pf [13] which only needs four-point correspondences to estimate the camera's focal length and extrinsic parameters.

Similarly, the homogeneous matrix $\mathtt{H}$ is a 3×3 matrix, defined up to scale, that has eight degrees of freedom. We only need four correspondence pairs to estimate it. The interest of homography estimation also lies in the ease of leveraging lines and points. Besides stacking equations of the form $\mathbf{x}' \times \mathtt{H}\mathbf{x} = 0$, one can also add equations corresponding to line correspondences: $\mathbf{l} \times \mathtt{H}^{\mathsf{T}}\mathbf{l}' = 0$. The ease of exploiting mixed geometric primitives for homography estimation does not extend in a straightforward manner to conics because of the quadratic nature of the Eq. (2.6). This difficulty is further investigated in **Chapter 5**.

However, points and lines can be in degenerate configurations, point or line detections can be noisy, which motivates the use of robust algorithms like RAndom

SAmple Consensus (RANSAC) [14] to find the best camera parameters or the best homography given the quality of the detections.

Starting from the parameters obtained directly from the DLT or from the RANSAC's results, a final non-linear optimization minimizes the geometric distance between measurements and their projection. A traditional approach consists in refining the estimated camera parameters using the Levenberg-Marquardt (LM) algorithm to minimize the geometric reprojection error.

In the context of sports, many camera calibration techniques exploit the fact that sports fields have fixed dimensions and known shapes. Building a sports field detector is thus an approach that enables obtaining 2D-3D correspondences.

2.2.2 Retrieval-based camera calibration

In the previous section, we have seen a classical approach to estimating camera parameters. The classical approach implies the detection of geometric primitives and the establishment of reliable correspondences with known landmarks in the images. Obtaining a sufficient amount of such geometric primitives (often points or lines) is not always possible, especially in close-ups, with texture-less environments such as sports fields. A way to bypass this requirement for correspondences is to exploit prior knowledge of the observed scene and estimate the camera parameters from a history of past camera calibrations. With the recent advances of deep neural networks, robust visual place recognition is now possible, enabling to retrieve images depicting the same location despite significant illumination or weather changes. This is a broad problem in the computer vision community and is mostly studied in the context of visual relocalization. In a simplified retrieval pipeline, a query image is first processed by a deep neural network to extract a compact visual descriptor. These features are compared to a database of features corresponding to the calibrated images of the database, according to a similarity metric.

Formally put, image retrieval-based methods first populate a database $\mathcal{D} : \{(I_i, \mathtt{K}_i, \mathtt{R}_i, \mathbf{t}_i)\}_{i=0}^{N}$, learn a descriptor $\phi(\cdot)$ and retrieve for an image query I_q

$$I_j = \arg\max_i s(\phi(I_q), \phi(I_i)) \,, \tag{2.9}$$

where $s(\cdot)$ is a similarity function like cosine similarity. The final camera parameters are retrieved from $\mathcal{D}$: $\kappa^* = (\mathtt{K}_j, \mathtt{R}_j, \mathbf{t}_j)$. In practice, modern retrieval-based methods rarely rely solely on the nearest neighbor. Instead, multiple candidates are often retrieved and subsequently combined, or used to initialize a geometric refinement stage that further optimizes the camera parameters. In camera calibration for sports

broadcasting, a final refinement step is usually performed to further improve the estimated camera parameters.

Sports broadcast images present a major advantage: their standardized field markings simplify synthetic database generation. Instead of storing real images, semantic segmentation maps of field markings are matched against field model projections across a broad range of camera configurations [7, 15]. This synthetic approach inherently generalizes the database to new or unknown venues, assuming an operational field-marking detector.

Formally, the database is constituted with synthetic images $\mathcal{D}'$: $\{(\pi_i(\mathcal{F}), \mathtt{K}_i, \mathtt{R}_i, \mathbf{t}_i)\}_{i=0}^{N}$, where $\pi_i(\mathcal{F})$ is the projection of the field markings model using the camera parameters $(\mathtt{K}_i, \mathtt{R}_i, \mathbf{t}_i)$. For the retrieval part, having a detector $d(\cdot)$ that performs the semantic segmentation of field markings $\mathcal{F}$, the similarity function is applied to the semantic segmentation of the image and the synthetic maps generated by the projection of the field model using the camera parameters $\max_i s(d(I), \pi_i(\mathcal{F}))$.

2.2.3 Absolute pose regression

In recent years, deep learning has not only been successfully deployed for semantic classification tasks, but direct regression tasks have also seen significant performance gains. With this in mind, why limit the task to retrieval approximation instead of predicting camera parameters directly? This is the approach investigated by Kendall, Grimes, and Cipolla [16] where the camera extrinsic parameters are predicted by a small deep neural network that has been trained on the scene. Formally, Absolute Pose Regression (APR) methods learn a mapping function:

$$\Phi(I) \rightarrow (\mathtt{R}, \mathbf{t}). \tag{2.10}$$

Unfortunately, according to Sattler et al. [17], these neural networks fail to generalize and rather perform an image retrieval task implicitly. As a way to mitigate this, some methods combine the image retrieval task with a relative pose regression task to better generalize.

In the sports domain, absolute pose regression has primarily emerged through hybrid approaches built upon the homography regression network proposed by DeTone, Malisiewicz, and Rabinovich [18]. While originally developed to estimate the homography between two images, making it closer to a relative pose estimation task, this framework was adapted to regress the homography relating a sports field model to an input image.

2.2.4 Perspective and field-of-view fields

Recent works have shown that over-parametrized representations generalize well [19]. This observation has motivated several approaches that formulate camera calibration as the prediction of dense geometric fields rather than a direct regression problem.

Several works have shown great success at framing intrinsics camera parameters estimation as a dense per-pixel prediction problem. Perspective fields and GeoCalib [20, 21], for example, predict for each pixel $\mathbf{x}$ the following geometric cues: up vectors $\mathbf{u}$ and latitudes λ, with λ defined as the angle between the viewing ray and the horizon $\Phi(I, \mathbf{x}) \rightarrow (\mathbf{u}, \lambda)$. Similarly, AnyCalib [22] allows modeling any camera model by predicting the ray corresponding to each pixel with field-of-view fields $\Phi(I, \mathbf{x}) \rightarrow \mathbf{r}$. In a second step, these techniques allow recovering intrinsic camera parameters either by training a neural network [20], by LM optimization [21] or directly by a closed-form solution [22]. While these methods only enable intrinsic camera parameter estimation and optionally some orientation-related information, this partial information can be valuable for sports content enrichment, as will be explored in **Chapter 6**.

2.2.5 Geometry foundation models

Another approach that has gained in popularity recently is the generic prediction of the geometry of the scene from images. Early work in this direction focused on Monocular Depth Estimation (MDE) which predicts depth maps $Z(\cdot)$, which define a depth value for each pixel of the input image: $\Phi(I) \rightarrow Z(\cdot)$ [23, 24]. In the latest metric depth estimation works, the neural networks either directly predict point maps [25] $\Phi(I) \rightarrow (X(\cdot), Y(\cdot), Z(\cdot))$, or alternatively, they enable point map recovery by predicting the focal length [26]: $\Phi(I) \rightarrow (Z(\cdot), f)$ and they can recover $X = \frac{x - p_x}{f} Z(x, y)$ and $Y = \frac{y - p_y}{f} Z(x, y)$, with square pixels and the principal point $\mathbf{p} = (p_x, p_y)$. This evolution is particularly important for camera calibration, as point maps can be interpreted as dense 2D-3D correspondences between image pixels and reconstructed 3D points, from which camera parameters may be estimated. However, the coordinates predicted by these methods are generally expressed in an arbitrary local reference system. While most of the time, in order to enrich sports content in a single view setup, we wish to recover point maps that are expressed in a known, world reference coordinate system.

Recovering scene coordinates in a predefined reference system has been explored by Scene Coordinate Regression networks [27, 28] that predict point maps relative to a scene. However, as these methods are scene-specific, they fall outside the scope

of foundation models.

Related to a shared coordinate system, but still not solving this issue of anchoring the scene in a world reference system, DUSt3R [29] and MASt3R [30] are the first geometric foundation models that allow the estimation of joint metric point maps: they take two views as input and jointly estimate a point map per input image that share a common coordinate system. From these jointly reconstructed point maps, relative camera poses and camera intrinsics can subsequently be estimated. Later works like VGGT [31], Π^3 [32], and MapAnything [33] extend these works to handle several views from the start and estimate point maps and camera poses, and sometimes auxiliary outputs such as depth maps.

These methods are of great interest as they enable 3D reconstruction from images. However their application to the broadcast domain remains challenging because (1) sports scenes are very large, they are underrepresented in common datasets and therefore constitute a clear out-of-distribution scenario and (2) broadcast cameras operate across very wide ranges of field of views compared to common benchmarking cameras which currently primarily target robotics or mobile devices and applications.

2.2.6 Conclusion

In this chapter, we have explored the possible ways to calibrate a camera in an unknown scene in a generic setting and discussed how it applied to sports broadcasting. However, we did not discuss the way to evaluate and the available datasets to perform camera calibration benchmarking. In this regard, sports are very specific, and those two aspects cannot be treated in a general context. This is the subject of our first paper. In the next chapter, we address camera calibration in the context of sports and start by asking ourselves what a good calibration of the camera entails exactly, especially when ground truth is not available.

Chapter 3

Benchmarking Camera Calibration in Sports (1^{st} article)

This first chapter reproduces the following paper: Floriane Magera et al. "A Universal Protocol to Benchmark Camera Calibration for Sports". In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). *Seattle, Washington, USA: IEEE, June 2024, pp. 3335–3346*

Prologue

At the beginning of this thesis, we identified a gap between the level of camera calibration accuracy required by industrial applications and the performance obtained by state-of-the-art methods. Indeed, in applications such as sports officiating, high precision is required to assess the position of player limbs up to a few centimeters from broadcast footage.

At the time, benchmarks for camera calibration consisted of a few datasets that contained only up to a few hundred images. The ground-truth annotations were represented as homographies, which is sensible under two assumptions: (1) radial distortion is negligible and (2) sports fields are planar. Our industrial experience in football was rather demonstrating that neither assumption holds systematically in real broadcast environment. As a consequence, evaluation metrics inherited the same assumptions, making it difficult to determine whether improvements reflected better camera calibration or better agreement with the benchmark model.

In contrast, the industrial benchmarking procedures require dedicated surveying equipment, access to stadiums, and broadcast cameras, without mentioning the

workforce required to operate the cameras and survey the field, grant stadium access, and obtain broadcasting rights. Scaling these tests to an open-source dataset would require tremendous investment.

This chapter is an attempt to tackle these issues by finding a way to evaluate camera calibration such that the cost of creating a dataset is manageable, and such that the shortcomings of the previous benchmarks are resolved. The objective is to create a new benchmarking protocol that is both *reliable* and *scalable*. The first task of this thesis was therefore to ensure there was some common ground to publish and agree on what a good calibration is for sports.

The first contribution of this thesis is thus not a camera calibration method, but rather aims to establish a better way to assess camera calibration quality. Through our substantial investment in the SoccerNet community, this new benchmarking protocol found a direct application to football by the organization of camera calibration and game state reconstruction challenges.

A Universal Protocol to Benchmark Camera Calibration for Sports

Abstract

Camera calibration is a crucial component in the realm of sports analytics, as it serves as the foundation to extract 3D information out of the broadcast images. Despite the significance of camera calibration research in sports analytics, progress is impeded by outdated benchmarking criteria. Indeed, the annotation data and evaluation metrics provided by most currently available benchmarks strongly favor and incite the development of sports field registration methods, i.e. methods estimating homographies that map the sports field plane to the image plane. However, such homography-based methods are doomed to overlook the broader capabilities of camera calibration in bridging the 3D world to the image. In particular, real-world non-planar sports field elements (such as goals, corner flags, baskets, ...) and image distortion caused by broadcast camera lenses are out of the scope of sports field registration methods. To overcome these limitations, we designed a new benchmarking protocol, named ProCC, based on two principles: (1) the protocol should be agnostic to the camera model chosen for a camera calibration method, and (2) the protocol should fairly evaluate camera calibration methods using the reprojection of arbitrary yet accurately known 3D objects. Indirectly, we also provide insights into the metric used in SoccerNet-calibration, which solely relies on image annotation data of viewed 3D objects as ground truth, thus implementing our protocol. With experiments on the World Cup 2014, CARWC, and SoccerNet datasets, we show that our benchmarking protocol provides fairer evaluations of camera calibration methods. By defining our requirements for proper benchmarking, we hope to pave the way for a new stage in camera calibration for sports applications with high accuracy standards.

3.1 Introduction

Camera calibration, also known as camera resectioning, is a necessity in many computer vision tasks. It involves estimating the parameters of a camera model, usually the pinhole camera model, that approximates the physical camera that produces a given image. Applications that require accurate calibration are varied and include virtual reality [34], underwater measurement [35, 36], traffic analysis

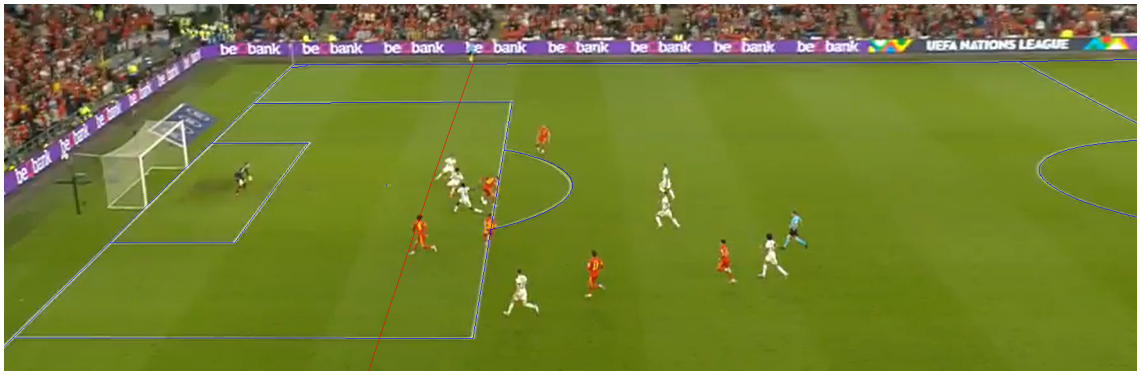


Figure 3.1: Illustration of a successful camera calibration. Lines superimposed in blue are obtained by projecting the markings of a soccer field. A perfect alignment is mandatory to enable high-precision applications such as offside position assessment (in red, a parallel to the goal line to decide on an offside situation). This paper proposes a new protocol, named ProCC, that has two advantages: (1) it is applicable to any sport, and (2) its evaluation is based on a new metric which is agnostic to the chosen camera type and model.

and surveillance [37–39], vehicle localization [40] or speed estimation [41], person re-identification [42], 3D reconstruction [43], *etc*. In this paper however, we focus on calibration of cameras used in sports event broadcasting as shown in Figure 3.1, and present a new protocol for the evaluation of the camera calibration quality that is usable for any camera type (static/moving, PTZ, fish-eye, wide angle) and model (pinhole, with/without radial distortion, *etc*.).

Camera calibration for sports. In the context of sport, computer vision systems are used for multiple purposes, such as augmented reality graphics [44], 3D ball trajectory reconstruction [45] or tracking [46], refereeing [47–49], or the computation of statistics regarding position, speed, and acceleration of balls, players, bats, sticks, pucks, *etc*. [50–53]. The expectations in terms of precision and robustness of these systems is constantly rising. Both the fact that it is considered as a valid alternative to GPS trackers for player tracking systems and the fact that the FIFA supported the use of the SAOT (Semi-Automated Offside Technology) for the 2022 Football World Cup are good indicators of the trust put in computer vision systems. Sports analysis is an active subject of research, with several recent datasets gathering broadcast images with a wide range of annotations serving for game analysis, such as DeepSportRadar dataset for Basketball [54], or the SoccerNet datasets for soccer

[55–58]. The SoccerNet dataset which was initially created for action spotting tasks [57, 59] covers a wide range of tasks for soccer analysis. In particular, the game state recognition task, one of the latest challenges proposed by the SoccerNet team, heavily relies on the ability to calibrate the camera to produce a strategic minimap of the players' localization [8]. This challenge paves the way to a 3D reconstruction that will require to be able to calibrate any camera along the field, including close-up cameras in a multi-view setup.

Camera calibration in the context of sports benefits from the presence of the sports field whose dimensions are specified by the rules of the game (see for example [60, page 32] for soccer, [61] for basketball, [62] for volleyball, and [63] for ice hockey). Their well-known shape and their presence in most sports images make them convenient calibration patterns. Using the field model as a calibration pattern, camera calibration techniques can thus express the parameters by a projection matrix, and sports field registration techniques express the mapping between the field plane and the image by a perspective transform, better known as a homography [64]. Note, however, that the rules of games allow some tolerance on field dimensions, so calibration systems have to cope with some dimension uncertainties. In this regard, goal posts are supposed to be perfectly sized, which makes them adequate to serve as calibration landmarks.

As the terms of "camera" calibration and "sports field" registration can be sometimes used interchangeably in the literature, we deem important to emphasize the difference between them. A sports field registration technique aims to estimate a homography, *i.e.* the transformation between the 3D sports field plane and the image. This mapping is not defined outside these two planes, which is insufficient for all 3D applications that involve elements outside the field plane. This is a blind spot in all papers that mention 3D applications such as player and ball tracking, tactics analysis, augmented reality (AR), *etc*. On the other hand, a camera calibration technique provides a mapping between the 3D world (not only the field plane) and the image, which makes it suited for the aforementioned applications. Sports field registration is an approximate attempt or a first step towards camera calibration, and this is why, in this paper, we consider the homography as a camera model, despite its practical limitations. However, we show that our protocol is applicable to any model, broadening the possibilities in terms of camera models.

Outline. The rest of the paper is organized as follows. We present the related work on camera calibration for sports, including a review of existing datasets, in Section 3.2. Section 3.3 describes the key elements of our benchmarking protocol, named ProCC, the new metric JaC[1] that we introduced in SoccerNet-calibration,

[1]We have renamed the metric from AC, used in SoccerNet-calibration, to JaC to avoid any

as well as a more relevant ground-truth type, to express the camera calibration quality. Our benchmarking protocol, including the annotation procedure and its metric, are the main contributions of this paper. Section 3.4 applies our protocol to the evaluation of camera calibration for sports broadcast events. This section also includes an experimental analysis of the protocol for different camera models, illustrating the universality of our protocol and showing that previous protocols have limitations that our new protocol can overcome. Section 3.5 discusses the results and explains why current calibration techniques developed in the literature overlook the broader capabilities of camera calibration in bridging the 3D world to the image. Finally, Section 3.6 presents a brief conclusion.

3.2 Related work

In the current state of the art, most of the evaluation procedures are based on homography annotations (to the best of our knowledge, basketball [65, 66] and athletics [67] are the only exceptions) which restricts the evaluation to the sports field plane, even if some techniques compute camera parameters. Indeed, all the following camera calibration techniques in the literature [7, 15, 68, 69] use the pinhole model, whose camera parameters can be converted to the homography that maps the image to the sports field plane.

3.2.1 Datasets to benchmark camera calibration

The latest techniques in camera calibration for sports all leverage the presence of a sports field to understand the mapping between the image and the world.

By browsing the scientific literature, we found a dozen datasets on which researchers evaluated their methods. Table 3.1 shows that most datasets are only relevant for sports field registration, as their annotations consist of homographies. For some datasets, it seems that the ground truth has not been annotated once and for all, such that several researchers have created their own annotations, which does not ensure the comparability of their results [70].

The nature of the annotations, the relatively small sizes of the datasets for some sports, as well as the difficulty of getting access to the data, make the creation of reliable camera calibration algorithms difficult.

confusion with the notion of accuracy as used in binary classification.

Name	Sport	Open data	Size	Annotations	Techniques
WorldCup 14	Soccer	Yes	395	Homographies	[5, 7, 15, 50, 68–78]
TS-WorldCup	Soccer	Yes	3,812	Homographies	[75]
SoccerNet-calibration	Soccer	Yes	21,132	Field markings	[69]
CARWC	Soccer	Yes	4,207	Homographies	[79]
SportLogiq	Ice Hockey	No	1.67M	Not specified	[70, 71, 73]
SportsFields by Amazon	Multi-sports	No	2,967	Homographies	[74]
Volley ball	Volley ball	Yes	470	Homographies	[15, 70]
College Basketball	Basketball	No	640	Homographies	[7]
DeepSportRadar	Basketball	Yes	728	Pinhole model	[54]
3DMPB	Basketball	Yes	10k	Pinhole model	[66, 80]
Athletics	Athletics	Yes	10k	Pinhole model	[67]

Table 3.1: Main current datasets dedicated to camera calibration for different sports. In this table, we also mention if the annotations are available, the number of images (Size) for which annotations are provided, if the annotations are specific to homographies or the pinhole model, and list some camera calibration techniques using a given dataset. From this table, one can see that the WorldCup 14 dataset is the most popular dataset for benchmarking, despite its limitations, as explained in this paper.

3.2.2 Evaluation of camera calibration in sports

Today, most authors evaluate their performance using an intersection over union metric, more specifically the $\text{IoU}_{\text{whole}}$ metric, that was first introduced by Homayounfar *et al.* [71] for the performance evaluation of their algorithm on soccer and ice hockey content. This metric measures the intersection over union between the real-world rectangle defined by the field template and its successive projection then deprojection using both the annotated homography and the camera parameters in the 3D world. This metric was extended by Sharma *et al.* [72] into another metric, denoted by IoU_{part}, to only account for the parts of the field that are visible in the image. These two metrics, which are illustrated in Figure 3.2, were applied to other sports such as volleyball [15, 70], basketball [7], tennis and American football [74].

In 2022, we launched the SoccerNet-calibration challenge, which is an attempt to further improve calibration techniques, but despite the organization of two SoccerNet-calibration challenges [9, 10], the benchmark of the World Cup 14 (WC14) dataset remains popular and the most commonly used benchmark. Therefore, we have collected the results of the current state-of-the-art methods, organized by their year of publication, regarding the WC14 benchmark in a consolidated leaderboard given in Table 3.2. However, the evaluation of one method stands out: Theiner *et al.* [69] evaluated their method both on the WC14 and SoccerNet-calibration datasets and, even further, they already showcased some discrepancies between our protocol ProCC and the WC14 evaluation protocol when applied to the same dataset.

Reference	$\text{mIoU}_{\text{whole}}$	$\text{medIoU}_{\text{whole}}$	$\text{mIoU}_{\text{part}}$	$\text{medIoU}_{\text{part}}$
Homayounfar *et al.* [71]	83	-	-	-
Sharma *et al.* [72]	-	-	91.4	92.7
Chen and Little [15]	89.2	91.0	94.7	96.2
Jiang *et al.* [73]	89.8	92.9	95.1	96.7
Sha *et al.* [7]	88.3	92.1	93.2	96.1
Citraro *et al.* [68]	90.5	91.8	-	-
Cioppa *et al.* [5]	79.8	81.7	88.5	92.3
Li *et al.* [77]	92.1	94.3	95.1	96.7
Tsurusaki *et al.* [78]	-	-	**97**	-
Nie *et al.* [74]	91.6	93.4	95.9	97.1
Shi *et al.* [70]	**93.2**	**94.9**	96.6	**97.8**
Chu *et al.* [75]	91.2	93.1	96.0	97.0
Zhang and Izquierdo [76]	90.0	92.8	95.3	96.9
Theiner and Ewerth [69]	-	-	95.3	96.6
Maglo *et al.* [50]	92	94.1	96.3	97.4

Table 3.2: Consolidated leaderboard by collecting published values on the World Cup 14 benchmarking dataset. The values are the mean or median intersections over union (denoted respectively by mIoU and medIoU), on the whole field or on the visible part of it, as mentioned in the references. In this table, references are organized by their year of publication, and the best values are given in bold.

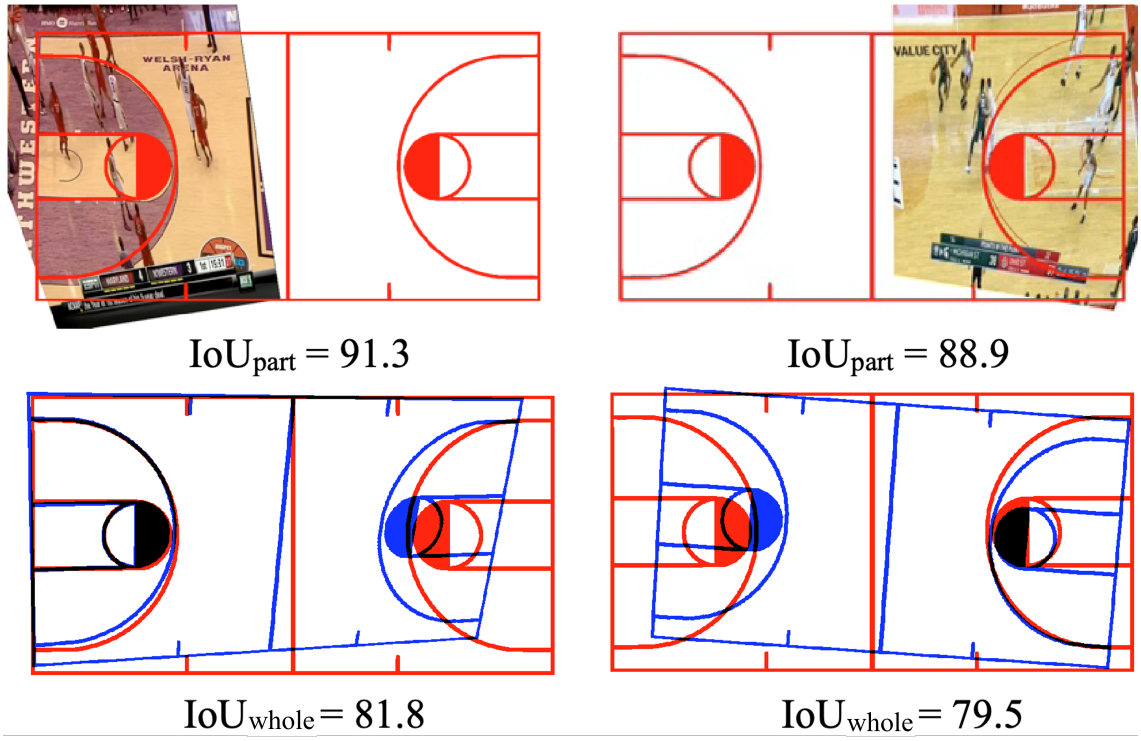


Figure 3.2: Visualization of the IoU_{whole} and IoU_{part} metrics. Illustration taken from [7] (©IEEE, 2020).

Furthermore, Nie *et al.* [74] and Chu *et al.* [75] provide an exhaustive evaluation by adding reprojection and projection errors. The projection error measures the average distance in meters between the projection of pixels sampled in the field image, using the inversion of both the estimated homography and the ground-truth homography. The reprojection error measures the average distance between reprojected points in the image using the ground-truth homography and the predicted homography, this distance is then normalized by the image height.

All the aforementioned metrics rely on the annotated homographies of the dataset, which are a limited and simplified interpretation of the observable field markings in the image. Moreover, with the small size of the World Cup 14 dataset and the small improvements in performance obtained in the last few years, concerns have been raised about the relevance of this dataset. Indeed, Chu *et al.* [75] proposed a new dataset called TS-WorldCup to increase the dataset size and allow for tracking evaluation, Claasen *et al.* [79] further corrected both the World Cup 14 and the TS-WorldCup annotations in a revised version named CARWC, which proposes homographies that are more precise. While the latter two improvements address some limitations of the current benchmarking protocol, we wish to go one step further by proposing to get rid of homographies as ground-truth data for camera calibration

in sports. In our opinion, one major problem with the current benchmarks is the imperative of a restricted and limited camera model for fulfilling the requirements of professional use. Indeed, around the field for a top-tier sports game, there can be up to 50 cameras, of a wide variety of quality: from wide-angle to super slow-motion cameras or fish-eye cameras, one model does not fit all of them. To circumvent this issue, we propose a new benchmarking protocol that allows for the use of any camera model. Furthermore, if we were to actually choose only one camera model for a set of broadcast cameras, our protocol would allow deducing which model would be the best fit according to our model-agnostic metric —or usual metrics such as the reprojection error— instead of choosing an axiomatic, arbitrary camera model from the outset.

3.3 A model-agnostic camera calibration benchmark

In this section, we describe our new benchmarking protocol that is based on two main pillars: annotations and a metric. Both are designed by assuming that a good camera calibration algorithm will be able to produce results that allow a minimal reprojection error, which is, in fact, the only reasonable assumption when there is no actual ground-truth knowledge about the broadcast cameras that captured the images of the datasets.

3.3.1 Annotations: beyond homographies

As an essential requirement for evaluating camera calibration is to handle different camera models, we must first change the type of the annotations and, subsequently, provide a revised evaluation metric.

In contrast to camera calibration techniques that rely on a specific pattern that will only be seen once and self-calibration techniques that do not require any calibration target and instead usually rely on reference frames, sports images have the unique advantage of always displaying at least partially a specific pattern, which is the sports field. We suggest leveraging this advantage, which is already done in most benchmarking and calibration techniques, but also pushing forward, by using as ground truth a lower-level interpretation of the sports field image: semantic point annotations of the sports field markings, which are annotations that are valid for any type of camera.

A sports field can be decomposed into a set of simple geometrical elements that are points, lines, circles, or ellipses. For each of these simple constituent elements of the field, we propose to assign a unique semantic label. In general, the field is only

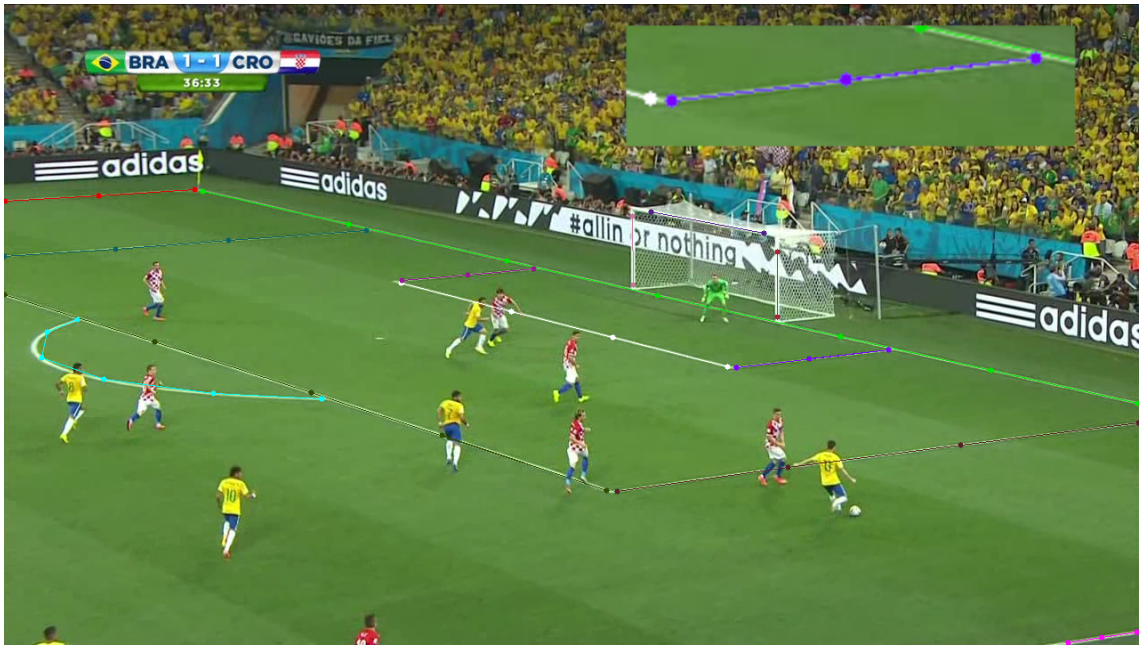


Figure 3.3: Illustration of annotations on a SoccerNet-v3 [81] image as used for the camera calibration challenge. As shown in the zoomed snapshot, annotations consist of points and the labels of the objects they belong to.

partially visible in the image, so for a given image, only a few elements need to be annotated. Practically, we annotate points along each element and thus, obtain a set of points for each semantic element of the field. To illustrate this, in SoccerNet, the annotations consist of polylines, *i.e.* sequential lists of 2D points along each soccer field element. Ideally, annotations should be regularly spaced, and the density of annotated points might be increased depending on the curvature of the field element as it appears in the image. Each element of the soccer field corresponds to a class, and in total, we count 26 semantic classes. An example of annotation is shown in Figure 3.3; in this figure, the points are provided by human annotators, while polylines are superimposed by an annotation tool.

By employing a semantic annotation of soccer field elements represented by a polyline, we meet the first requirement of a universal evaluation protocol, in the sense that the ground truth is independent of the used camera model. The main advantages of this annotation type in the image domain are its independence from the task to be solved and the ease to validate it by the human eye. As the annotations do not necessarily cover all field markings pixels, we will show that our metric addresses the issue that while the annotated points are part of field elements, it is possible that neighboring pixels also belong to the same element, despite being unannotated.

Finally, the use of semantic annotations opens up possibilities for more complex calibration scenarios. For example, if a system can triangulate a 3D point and knows

its projection in the image, it may use it for calibration, even if the point belongs to a moving object such as a player. This opens the way to close-up camera calibration in a multi-view setup.

We now show how to evaluate a camera calibration based on this ground truth.

3.3.2 Evaluation metric

Given a field model and the corresponding annotations in the image, we can evaluate the quality of camera parameters by assessing how well the camera projection of the 3D field superimposes on the annotations. A camera model definition includes the mathematical imaging function that allows for the computation of the 2D image of a 3D object. This function is also called the *projection* function. In our protocol, since the evaluation must be independent of the model, the only requirement for a camera model is that this function must be defined and, when it is given a 3D point $\mathbf{X}$, it must be able to provide its corresponding 2D location $\mathbf{x}$ in the image plane. We denote the projection function of a camera model by π and, therefore, $\mathbf{x} = \pi(\mathbf{X})$.

The metric that we propose is inspired by the reprojection error, while aiming to address its shortcomings. The reprojection error usually measures the Euclidean distance between the observation (*e.g.* the annotation) of an object and the projection of that 3D object model using the estimated camera model. The mechanism of projection is the source of several difficulties. First, depending on the chosen camera model, the reprojection error may be undefined when the projection of the 3D object does not land in the image and, more generally, is unsuited to handle hallucinated or missing objects. Second, a projection may not exactly match an annotation, but may still be superimposed on visible field marks, since the field markings have a thickness that can make them visible for more than one pixel in width. Finally, it is inconvenient to grasp the meaning of an unbounded metric that, in our case, can vary between zero and infinity. For these reasons, by thresholding the reprojection errors, we obtain another metric that can be intuitively understood as the proportion of field elements that are correctly imaged, bridging the camera calibration evaluation with an object detection problem. Indeed, the underlying intuition behind our metric is that the quality of the camera parameters reflects how well the parameters can reproject objects close to their image.

Since we defined that each field marking element is annotated with as many points as necessary to constitute a fitting polyline, we mark a field element as correctly detected if its reprojection in the image is close enough to each of the annotated points of the polyline. The closeness criterion is defined based on a threshold value τ in pixels. More formally, a field element is defined by a polyline L, which is an

ordered list of 3D points. Its projection $\pi(L)$ gives a 2D polyline (*i.e.* an ordered list of 2D points) which also defines a list of line segments S by considering pairs of the consecutive 2D points. The distance between an annotated point $\mathbf{x}_i$ and the projection of the field element $\pi(L)$ is given by the minimal Euclidean distance between the point $\mathbf{x}_i$ and the segments S constituting $\pi(L)$.

In the following, we define the function $d(.)$ computing the distance between a point and a line segment. More specifically, given the projection $\mathbf{c}$ of the point $\mathbf{x}$ on the line passing through the segment extremities $\mathbf{a}$ and $\mathbf{b}$, we define the distance between the point $\mathbf{x}$ and the line segment $S_{\boldsymbol{ab}}$ defined by $\mathbf{a}$ and $\mathbf{b}$ with the traditional Euclidean distance function as follows:

$$d(\mathbf{x}, S_{\boldsymbol{ab}}) = \begin{cases} \|\mathbf{x}-\mathbf{a}\|_2 & z \leq 0\,, \\ \|\mathbf{x}-\mathbf{c}\|_2 & 0 < z < 1\,, z = \dfrac{\mathbf{c}-\mathbf{a}}{\mathbf{b}-\mathbf{a}}\,, \\ \|\mathbf{x}-\mathbf{b}\|_2 & z \geq 1\,, \end{cases} \tag{3.1}$$

where the variable z is introduced to account for the fact that the distance to the line segment is equal to the distance to one of its extremities if the point projection c lands outside the segment. Finally, we say that a field element L is correct if all the points $\mathbf{x}_i$ annotated for this element are less than τ pixels away from the projection of said field element L. In mathematical terms, we then have that:

$$\min_{S \in \pi(L)} d(\mathbf{x}_i, S) < \tau\,, \forall \mathbf{x}_i\,. \tag{3.2}$$

The correctly detected field markings are counted as true positives TP_τ. Both the hallucinated field markings and wrongly detected field elements whose reprojection lands further than τ pixels from the annotations are false positives FP. Missing field elements are counted as false negatives FN. Finally, we define our JaC_τ metric, the Jaccard index for camera calibration or JaC, to evaluate the calibration "accuracy" as follows:

$$\mathrm{JaC}_\tau = \frac{\mathsf{TP}_\tau}{\mathsf{TP}_\tau + \mathsf{FN} + \mathsf{FP}}\,. \tag{3.3}$$

We propose to use this metric, parametrized by the reprojection error τ, for the best comprehension of the camera parameters quality.

3.4 Results

For practical reasons, our experiments focus on soccer because this sport has the most available techniques and data. Furthermore, we are restricted to soccer due to a lack of semantic annotations for other sports, despite that ProCC is applicable to any sport for which there are measurable 3D references. Hereafter, we describe experiments on the WC14, CARWC, and the SoccerNet datasets to demonstrate our benchmarking protocol.

3.4.1 Description of the datasets

For all the datasets, the experiments employ soccer field markings and goal posts that are decomposed into distinct classes.

SoccerNet. The SoccerNet dataset contains 21,132 images that were recently annotated with soccer field markings elements, as well as goal posts elements, which provides 3D correspondences, thus outside the field plane [81]. In total, the annotations consist of 167,589 field marking lines or circles and 53,577 goal posts elements that are grouped into 26 classes for soccer fields. Each element is annotated with its two extremities for rectilinear elements and by as many points as needed to fit curves for non-rectilinear elements.

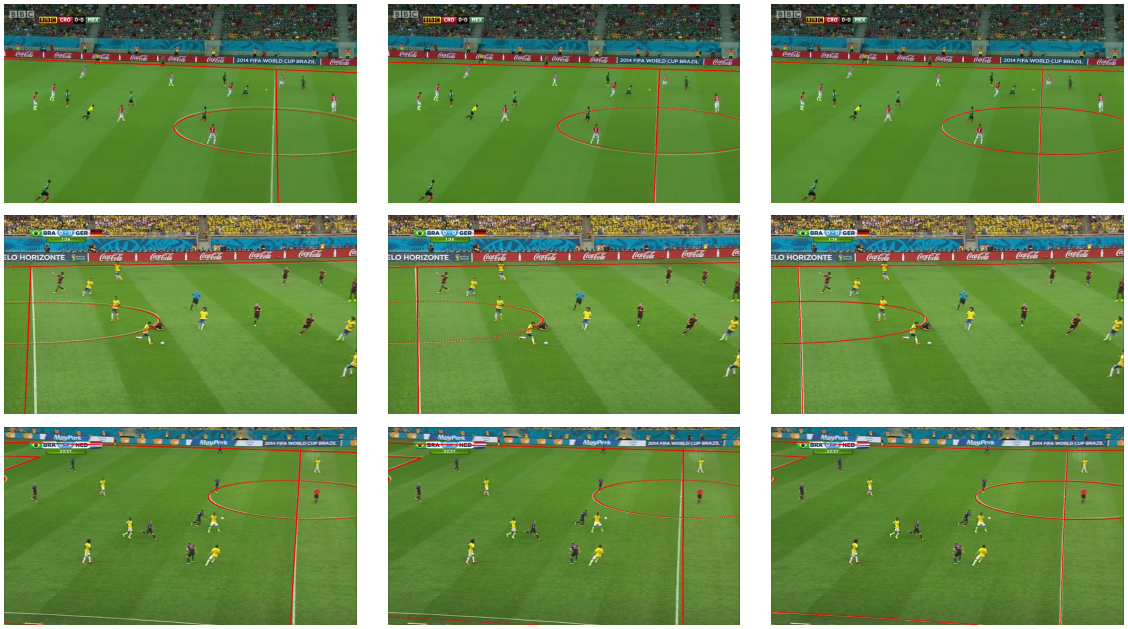
World Cup 14. For the sake of comparison, the World Cup 14 test set has been manually annotated following our convention. Indeed, we annotated the 186 images with points along soccer field elements, resulting in a total of 1,681 soccer field elements annotated with several points.

3.4.2 A better camera model for broadcast cameras

The goal of our experiments is twofold: validate our evaluation protocol, and demonstrate that there is a better camera model for broadcast cameras. This is why we first compare both qualitatively and quantitatively ground-truth homographies of the WC14 and CARWC datasets with estimated camera parameters following a richer camera model.

In our experiments, we selected a threshold τ of 5 pixels for our JaC_5 metric, which is a value that is suited for distinguishing between methods given the current quality standards of the different camera calibration methods that exist in the literature. When we want to differentiate quite precise methods, we tune the pixel threshold to 2 pixels.

As illustrated in Figure 3.4 and explained in the above, the reprojection of the



WC14 annotations | CARWC annotations | Xeebra

Figure 3.4: Comparison of reprojected field elements based on different types of annotations: in the first column, the WC14 homographies are used to project the soccer field model. The second column corresponds to the CARWC annotations. Both these annotations fail to provide correct reprojections. Finally, the third column displays results obtained with the Xeebra product.

soccer field model using *ground-truth homographies* (see columns 1 and 2) fails to match the images of the field markings. This is an indication that the use of these annotations is suboptimal if we agree that a good camera calibration algorithm should be able to match the image of known 3D objects such as soccer fields, which is a reasonable and tractable solution in the absence of actual ground truth for the camera parameters.

In Table 3.3, we compare different camera models and highlight the limitations of using homographies as ground truth. The manual annotations and JaC_5 metric allow us to quantify the quality of the annotations of the WC14 and CARWC datasets compared to the camera parameters estimated with the Xeebra product. Xeebra [6] is a professional product for Video Assistant Referee, whose Offside Technology features are certified by the FIFA Quality Program [82]. With these experiments, our evaluation protocol confirms the concerns raised [69, 75, 79] about the quality of the WC14 ground truth. We obtain different results for the WC14 homographies evaluation than Theiner *et al.* [69] which is explained by the fact that we used our annotations. We also establish that the consolidated annotations provided by Claasen *et al.* [79] in CARWC are a welcome enhancement in terms of precision.

Camera model	JaC_5 (↑)	Reprojection error (↓)
Homography (WC14 annotations [71])	67.4	3.07
Homography (CARWC annotations [79])	79.1	1.79
Pinhole camera parameters with one radial distortion coefficient ([6])	92.5	1.44

Table 3.3: Comparison of camera models on the World Cup 14 dataset.

However, despite visible enhancements provided by the CARWC annotations to the WC14 dataset, the homography models fail to reproduce the images of the field markings, while we obtain better results by estimating camera parameters with Xeebra. This experiment demonstrates that the qualitative insights shown in Figure 3.4 are further supported by both the JaC_5 metric and the reprojection error, comforting the idea that our evaluation protocol is relevant and fulfills a need for correct evaluation. Moreover, the middling results of the careful CARWC annotations suggest that the problem does not lie with the quality of the provided annotations, but rather with the nature of the annotation. Indeed, we attribute the better results of the camera parameters of Xeebra to its inclusion of radial distortion parameters. In an attempt to demonstrate both our protocol's ability to assess the quality of different camera models and to further prove our last hypothesis, we evaluated Xeebra's results on SoccerNet as well.

In the next experiment, as can be seen in Table 3.4, we establish that on the SoccerNet test set, which is a much larger and thus more diverse and challenging dataset, it is a necessity to consider the radial distortion. Indeed, when radial distortion is considered in the optimization of the camera parameters, better results are obtained according to our protocol.

In these experiments, we have shown that our evaluation protocol obtains results in agreement with qualitative and quantitative evaluations, and that it allows to evaluate different camera models, which enables us to support our second contribution, stating that soccer broadcast cameras are better modeled with radial distortion.

3.5 Discussion

With the previous results, we have shown that better camera modeling allows for better precision in the image. In this section, we want to stress the impact of such improvements on the real-world applications of camera calibration algorithms.

	JaC_5 (↑)	JaC_2 (↑)	Reprojection error (↓)
P	78.7	40.2	4.51
R	83.1	54.3	4.01

Table 3.4: Comparison of camera models on the SoccerNet-calibration dataset. P corresponds to the simplified pinhole camera model, and R corresponds to the simplified pinhole model extended with one radial distortion coefficient.

Considering that many sports field registration methods emphasize the relevance of their research in applications like player tracking, it is essential to highlight the pivotal role of selecting an appropriate camera model for efficient player tracking on a sports field. In Figure 3.5, we show that, depending on the chosen camera model to link the image and the physical world, the differences in the estimated 3D positions can often exceed one meter. Such variations will inevitably impact the accuracy of player tracking throughout an entire sports game. Hence, we propose initiating a discussion on optimal modeling strategies before venturing into applications that may not be practically feasible. With our newly introduced benchmarking protocol, we enable a systematic evaluation that facilitates the exploration and analysis of the best models in this context.

Still, our protocol presents some shortcomings. The inherent limitations of our protocol lies in the quantity of field elements present in the images. Yet, the quality of the evaluation gets better when the number of elements increases and when the field markings are well distributed in the image. Indeed, camera models may become over-parameterized (and be overfitted) in the case of images that show only few field elements. Another limitation is the fact that it might be possible that very different camera parameters obtain the same results due to the symmetry of the sports field template, for instance; the uniqueness of the results is not guaranteed. However, introducing a means of managing all kinds of ambiguities, which in practice are rather rare, would make the evaluation protocol considerably more complex. A possible yet costly solution to this issue is to acquire real ground truth for the camera parameters, which requires new datasets and much greater means to equip the production cameras with sensors.

3.6 Conclusion

In conclusion, this study demonstrates the inadequacy of current sport field registration benchmarks due to suboptimal ground truth. Considering the availability of superior camera models, we advocate abandoning metrics reliant on homographies,

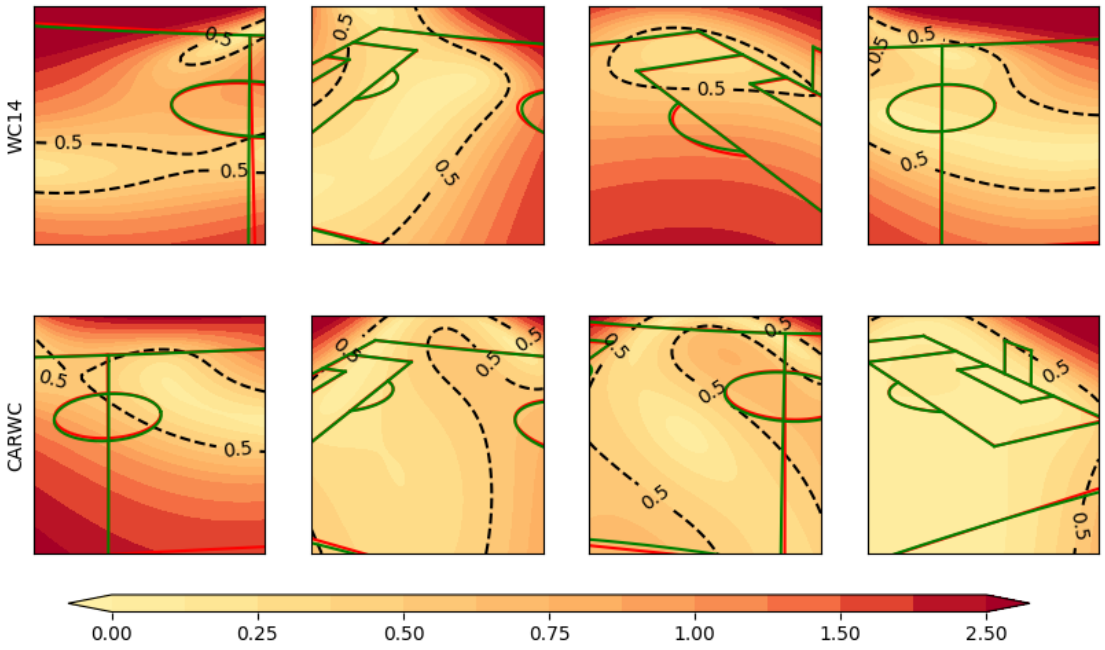


Figure 3.5: Illustration of the disagreement between estimations from two camera models: the ground-truth homography (see the red wireframes) and the richer model combining the pinhole and radial distortion used in Xeebra (see the green wireframes). The small difference between the two reprojected wireframes is misleading, as the superimposition of contour plots on the field shows that some parts of the sports field are seen over 2.5 meters apart by these two camera models. We have also plotted the contour line for a difference of 50 cm (see the dashed lines), which is the minimal accuracy standard demanded by the FIFA for Offside Technologies, to highlight why methods developed with older calibration protocols such as the WC14 or CARWC would fail to meet professional requirements.

and propose a new type of annotation, combined with proper metrics, leading to the definition of a camera model-agnostic evaluation protocol. Furthermore, our protocol proves that richer camera models, such as the pinhole model augmented with radial distortion instead of homographies obtain better results on broadcast cameras, which emphasizes the need to evaluate camera calibration algorithms independently of the camera model. Finally, it should be mentioned that our evaluation methodology applies to multi-view camera systems, as one can imagine that 3D elements visible in the calibrated cameras can be used, like any annotated data, as key points to calibrate other cameras. By doing so, not only have we have proven that our protocol improves camera calibration in sports, but also that there are new opportunities that are beyond the reach of methods solely based on field registration.

Chapter 4

Broadcast Camera Tracking (2^{nd} article)

This chapter reproduces the following paper: Floriane Magera et al. "BroadTrack: Broadcast Camera Tracking for Soccer". In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). *Tucson, Arizona, USA: IEEE, Feb. 2025, pp. 6177–6187*

Prologue

After introducing a way to evaluate camera calibration, which both met industrial needs and enabled the evaluation of more complex camera models than previous benchmarks, it was thus possible to prove that camera calibration would benefit from richer models, and to contribute to the state of the art with a new camera calibration method.

For this second paper, we noticed a mismatch between evaluation protocols and real-world usage: while camera calibration benchmarks were single-frame oriented and thus were designed to offer a collection of wide, diverse camera setups, the purpose of camera calibration is often targeting temporal applications, like player tracking. With the introduction of the SoccerNet Game State Reconstruction challenge [8], we had an opportunity to change this paradigm of benchmarking single-view camera calibration and actually study broadcast camera tracking.

This research is also motivated by the observation that applying single-frame camera calibration to each frame of a video introduces jitter and degrades tracking results, while by design the smoothness of the camera motion is ensured by both the camera operator and a professional camera tripod. Indeed, professional tripods are designed to resist abrupt motion changes through fluid drag mechanisms, meaning

that fast panning or tilting requires substantial operator effort. A camera operator will always apply low frequency movements to the camera; alternating pan direction or increasing and decreasing the focal length frame to frame is unlikely.

In this work, we try to exploit all the knowledge we have about broadcast cameras to propose a physical tripod model, which differs from the naive version of a fixed rotation center, as would be modeled with a PTZ camera, and exploit temporal consistency to produce camera parameters that evolve smoothly over time.

This resulted in a new state-of-the-art camera calibration algorithm, which allows sounder physical camera extrinsic parameters and demonstrates that the biggest accuracy improvement gain comes from modeling radial distortion.

BroadTrack: Broadcast Camera Tracking for Soccer

Abstract

Camera calibration and localization, sometimes simply named camera calibration, enables many applications in the context of soccer broadcasting, for instance regarding the interpretation and analysis of the game, or the insertion of augmented reality graphics for storytelling or refereeing purposes. To contribute to such applications, the research community has typically focused on single-view calibration methods, leveraging the near-omnipresence of soccer field markings in wide-angle broadcast views, but leaving all temporal aspects, if considered at all, to general-purpose tracking or filtering techniques. Only a few contributions have been made to leverage any domain-specific knowledge for this tracking task, and, as a result, there lacks a truly performant and off-the-shelf camera tracking system tailored for soccer broadcasting, specifically for elevated tripod-mounted cameras around the stadium. In this work, we present such a system capable of addressing the task of soccer broadcast camera tracking efficiently, robustly, and accurately, outperforming by far the most precise methods of the state of the art. By combining the available open-source soccer field detectors with carefully designed camera and tripod models, our tracking system, BroadTrack, halves the mean reprojection error rate and gains more than 15% in terms of Jaccard index for camera calibration on the SoccerNet dataset. Furthermore, as the SoccerNet dataset videos are relatively short (30 seconds), we also present qualitative results on a 20-minute broadcast clip to showcase the robustness and the soundness of our system.

4.1 Introduction

Sports content has the advantage of displaying sports fields, which have strictly regulated shapes and dimensions. This particularity makes sports camera calibration and localization possible in the wild, without the need for any other prior knowledge but the ability to correctly detect these field markings. This advantage is well-exploited, as all camera calibration techniques detect field markings, and single-frame camera calibration methods flourish, with ever-increasing accuracy. However, two facts that tend to be overlooked are (1) sports field markings can be really sparse in some areas, and (2) wide-angle broadcast cameras have typically 25× zoom

lenses, which can sometimes lead to quite narrow views. Thus, while the open-source datasets keep the task conveniently framed, there is an unmet need to address the case where few or even none of the field markings are visible for real-world applications. By carefully modeling broadcast cameras, we first include lens distortion effects — something that is systematically overlooked in sports field registration techniques that estimate homographies—, and which we show has a great impact on the accuracy of the results. Then, by including constraints on the camera movements that a tripod actually allows, we demonstrate that with more than accurate reprojections in the image, our models recover consistent position and rotation values, leading to effectively smooth tracking. Besides its novel state-of-the-art performance, our tracking system is robust and includes a reinitialization strategy that is thoroughly validated on the SoccerNet-calibration dataset. Finally, our system is usable in practice, as it works from the first frame and performs at a speed of 16 frames per second (fps) for HD (1920×1080 pixel resolution) images on a server equipped with two RTX 4090 GPUs, without any optimization; by experience, we know that a proper optimization of our code will make it real-time.

To validate our technique, we use the SoccerNet datasets [56, 57, 81]. The SoccerNet initiative has allowed new tracks of research for many problems related to soccer video understanding, including camera calibration and localization, by providing task definitions, corresponding public datasets, and metrics. Among others, the SoccertNet-calibration dataset (denoted as sn-calibration hereafter) is the first open-source dataset that does not use homographies as annotations, but rather the soccer field markings and goal posts. Besides the calibration task, which is about single-frame camera calibration, the novel game state reconstruction task [8] aims to continuously estimate the bird's-eye view of the soccer field with the players position. The dataset of the game state reconstruction task, denoted by sn-gamestate in the following, consists of soccer videos. Its annotations follow the same conventions defined for sn-calibration. Solving the game state task requires precise camera calibration and localization, consistently over time, which is a property also required for most applications regarding player performances and game analysis.

Contributions. By leveraging specific knowledge about broadcast cameras, we produce an efficient and accurate tracking system for soccer. We summarize our contributions as follows: **(i)** We propose a tracking system, named BroadTrack for the calibration of moving cameras that is both efficient and state-of-the-art, **(ii)** We define a camera model that is tailored for broadcast applications, and **(iii)** We release the source code of our tracking system at `https://github.com/evs-broadcast/BroadTrack`.

4.2 Related work

Common approaches to solving camera calibration and localization problems include structure from motion (SfM) [83] and simultaneous localization and mapping (SLAM) [84], both relying on the camera parallax to derive properties of the 3D world in which the camera moves. A typical pipeline could be described as follows: features are detected in the video frames, matched together across time, and finally triangulated. The resulting 2D-3D correspondences are fed in a PnP solver to derive the camera parameters. Compared to the usual problems that both SfM and SLAM solve, broadcast applications require custom approaches. Indeed, real-world applications often require a metric reconstruction of the scene, but for sports applications, the world reference system should not be placed regarding a camera; rather, the cameras should all be expressed relatively to the sports field.

Detectors. Unlike common camera localization or calibration schemes, sports camera calibration algorithms do not usually rely on 2D matching of features between frames, but rather on the detection of sports field elements that provide direct metric correspondences, which is an advantage for tracking, in effect reducing the risk of drifting. Most methods detect field markings in the image, either by binary segmentation or Hough line transforms [85–87], and sometimes even deriving vanishing points associated to horizontal and vertical lines [71, 88]. Lately, thanks to the performance of deep neural networks, higher levels of semantic interpretation were attained with sports markings being used as zone delimiters [7, 89], dividing the sports field into segmentation zones, or with some methods identifying specific markings as unique classes. Finally, some methods derive directly sports field lines intersections because most of the methods relying on solvers for camera parameters solely use point correspondences.

Virtual augmentation of correspondences. A global problem for all single-frame camera calibration and localization methods is the sparse nature of sports field markings, which are rarely uniformly distributed, leading to a lack of visual support in the image as there are few visible markings. Several methods alleviate this by detecting virtual correspondences on the field, or by deriving additional geometric cues from existing elements. For example, some works extend line segments to get new intersections [78, 90–92], derive tangents to circles or circle keypoints [90–93], detect grass mowing patterns [94], or even learn to detect grids of keypoints on the sports field surface [50, 74, 75, 79, 95]. These methods leveraging grid-like keypoints on sports fields have the drawback of first relying on homographies to derive the grid

projection in the image, which, as shown in previous works [1], are not the best fit to model the projective transformation.

Dictionary methods. Other works leverage directly prior knowledge about the broadcast camera, such that the number of correspondences in the image does not matter. These methods construct a dictionary of plausible broadcast camera views, and then retrieve camera parameters based on the similarity between the dictionary view and the camera view [7, 15, 72, 89, 96, 97]. These methods tend to be slow due to the search time in the dictionary, or if the dictionary is sparsely populated, relies on the ability of a Spatial Transformer Network to regress the parameters of the homography that maps the dictionary view to the current view [7, 89].

Homography regression. Inspired by PoseNet [16] and learned homography estimation [18], previous works directly regress homographies [70, 73, 98]. While the work of DeTone *et al.* [18] estimates the homography between a pair of images, its transposition to sports field registration requires the regressed homography to capture the transformation between a broadcast image and the synthetic bird's-eye view of the sports field. The huge perspective difference between a bird's-eye view of the sports field and one broadcast image probably explains why these methods rarely work in a single forward pass, limiting their use in actual broadcast scenarios.

Tracking. Due to the limited number of datasets with broadcast videos, few methods took interest in the temporal consistency of the camera calibration and localization. Among these methods, we define two main categories: (a) the ones based on homographies between successive frames, and (b) the ones using traditional tracking filters. These categories are further detailed hereafter.

(a) Homography between successive frames: A common strategy to perform tracking is to extract features to obtain point correspondences, or even line ones [99], and robustly estimate a homography between pairs of frames with RANSAC [50, 100]. The underlying assumption made when computing homographies between successive frames is that the camera is purely in rotation, an assumption that other works refute for soccer [101]. Note that a homography can be computed between frames if the correspondences are taken only on the sports field plane [79, 99]. In the case of rather small sports fields, like tennis, Farin *et al.* [87] assume that the camera velocity is constant, and thus assume that the homography between the future image pair will be equivalent to the homography mapping the present image pair.

(b) Common temporal filters: A range of methods also apply common tracking

algorithms such as the Kalman Filter, starting with Claasen and de Villiers [79], who use a Kalman filter to track point correspondences on the sports field, which are later used to track the homography using an extended Kalman filter. Beetz *et al.* [102] use an iterative extended Kalman filter to track their camera parameters, while Citraro *et al.* [68] use a particle filter to model the motion of the camera. Lu *et al.* [103] propose a SLAM algorithm tailored for a PTZ broadcast camera by building a map of 3D rays rather than 3D points, but consider the camera focal point fixed and neglect lens distortion.

Besides NBJW [90] which is the SOTA on the sn-calibration dataset, the methods that inspire us in terms of broadcast modeling are the ones that do model radial distortion [69, 86, 102, 104] and the tripod of the camera [101, 103, 105].

Available data. For completeness, we addressed both camera calibration and sports field registration techniques. However, in terms of datasets, given the recent concerns about the ability of homographies to properly model broadcast cameras[1, 69], we do not use any of the datasets that provide homographies as pseudo ground truth [71, 75, 79]. In this work, by combining high-level semantic analysis of the sports field marking detection and broadcast camera knowledge, we derive a new tracking system that achieves SOTA results on the sn-gamestate dataset.

4.3 Method

First, we define our models for the camera and for the tripod (Section 4.3.1), then we outline our tracking system (Section 4.3.2), which comes with a reinitialization algorithm (Section 4.3.3).

4.3.1 Broadcast camera model

To model our camera, we start from the pinhole model with the *calibration matrix* $\mathtt{K}$, defined as follows [11]

$$\mathtt{K} = \begin{bmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix} . \tag{4.1}$$

As stated in the standards of telecommunications for HD [106] and UHD content [107], broadcast pixels are squares, such that the focal length is the same for both axes, and the skew parameter can be ignored. We further assume that the principal point

is located at the image center, so that the set of intrinsic parameters reduces to one parameter, f.

The camera pose is modeled by its focal point position $\mathbf{C} = (C_x, C_y, C_z)$ and a rotation matrix $\mathtt{R}$ parameterized by the *pan* ϕ, *tilt* θ, and *roll* γ angles, defined accordingly in the Euler angles convention: $\mathtt{R} = \mathtt{R}_{\boldsymbol{z}}(\gamma)\mathtt{R}_x(\theta)\mathtt{R}_{\boldsymbol{z}}(\phi)$. In the absence of lens aberrations, the projection matrix of our camera would be formulated by the following equation[11]: $\mathtt{P} = \mathtt{K}\mathtt{R}\left[\begin{array}{cc} \mathtt{I} & -\mathbf{C} \end{array}\right]$. However, it is not uncommon for broadcast cameras to display some radial distortion, a deformation effect that occurs due to the curved nature of the camera lens. While it may be modeled in the pixel space [69, 104], we follow the standard way and apply this deformation in normalized coordinates [11, 108].

Given a point in the normalized image plane $\overline{\boldsymbol{x}} = (\overline{x}, \overline{y})$, the distortion function $\mathcal{L}(r) = 1 + k_1 r^2$ transforms $\overline{\boldsymbol{x}}$ according to its distance to the origin of the image plane $r = \|\overline{\boldsymbol{x}}\|_2$ which finally gives the image point expressed in pixel coordinates:

$$\boldsymbol{x} = f\,\mathcal{L}(r)\overline{\boldsymbol{x}} + \mathbf{p}\,. \tag{4.2}$$

We use the model proposed by Brown-Conrady [12], and discard all higher orders of radial distortion and tangential distortion as we experimentally find them superfluous. The set of unknown parameters for our camera is $\kappa : \{f, k_1, \phi, \theta, \gamma, C_x, C_y, C_z\}$, and we define the function $\pi_\kappa : \mathbf{X} \to \mathbf{x}$ that projects a 3D point to its respective image point. In the next section, we further analyze the specificities of broadcast cameras, and extend our model to consider the tripod on which a broadcast camera is installed, and the constraints that it creates on the camera parameters.

Pan-tilt head and tripod

Compared to the usual cameras used in the computer vision literature for robotics or AR with mobile phone captures, broadcast cameras are another kind of beast. A typical broadcast camera for soccer has a lens that can zoom up to 25 times, which consists of an arrangement of optics that can weigh up to a few kilograms [109]. When we use the pinhole model to represent the arrangement of optics inside those lenses, we approximate it with a single optic. While Chen *et al.* [101] assume that the camera focal point is at a fixed location inside the camera lens, we argue that this virtual projection center may not be located inside the camera and that it should evolve along the optical axis as the camera zooms. Therefore, our only assumption about the position of the focal point regarding the physical camera is that the focal point is located on the optical axis of the camera. We choose to rather constrain the

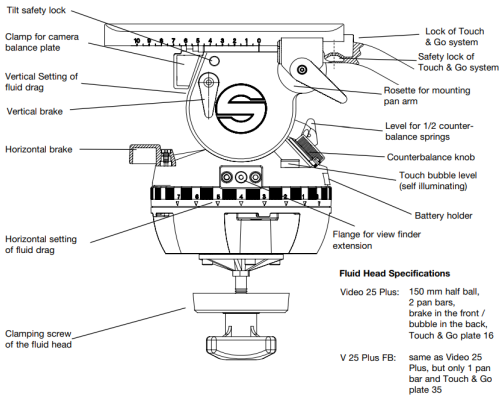


Figure 4.1: Usual pan-tilt head for 25× lens broadcast camera. (taken from [110], others pan-tilt heads can be found online [111]).

optical axis position regarding the tripod.

A camera comes in several parts: the tripod, the pan-tilt head, and the camera itself. The challenging part for our extrinsic parameters model is the pan-tilt head. This is the part that is mounted on the tripod on which the camera is rigged. The pan-tilt head allows the camera to rotate smoothly while allowing the camera operators to make quick sweeps if they lose track of the action. By construction, it allows the camera to rotate along two axes, *i.e.* to pan and tilt the camera. To better visualize the rotation axes, a common professional pan-tilt head is depicted in Figure 4.1.

According to this blueprint, we make the assumption that the pan and tilt rotation axes intersect in a point $\mathbf{T}$, which remains fixed during the whole game. Our model states that there is a point $\mathbf{O}$ in the camera which belongs to the optical axis of the camera, and which remains at a fixed distance δ of the rotation center $\mathbf{T}$ as the camera moves. Let $\mathbf{r}_i$ be the ith column of the orientation matrix of the camera $\mathtt{R}^\mathsf{T}$, the vector $-\mathbf{r}_2$ is the upvector of the camera, while the vector $\mathbf{r}_3$ defines its optical axis direction. If we further assume that the camera is centered on the pan-tilt head, this point $\mathbf{O}$ is then determined by the upvector $-\mathbf{r}_2$ and its distance to the tripod rotation center $\mathbf{O} = \mathbf{T} - \delta\mathbf{r}_2$. As the focal point of the camera changes with the zoom level of the camera, the position of the focal point is finally given by $\mathbf{C} = \mathbf{T} - \delta\mathbf{r}_2 + \lambda\mathbf{r}_3$. The parameters $\{\mathbf{T}, \delta\}$ are fixed over time, while λ varies with the camera view. A visualization of the model is shown in Figure 4.2.

4.3.2 Complete description of our tracking system

Our system comprises different steps, described hereafter: the detection of the field markings (Section 4.3.2), which is augmented using optical flow (Section 4.3.2),

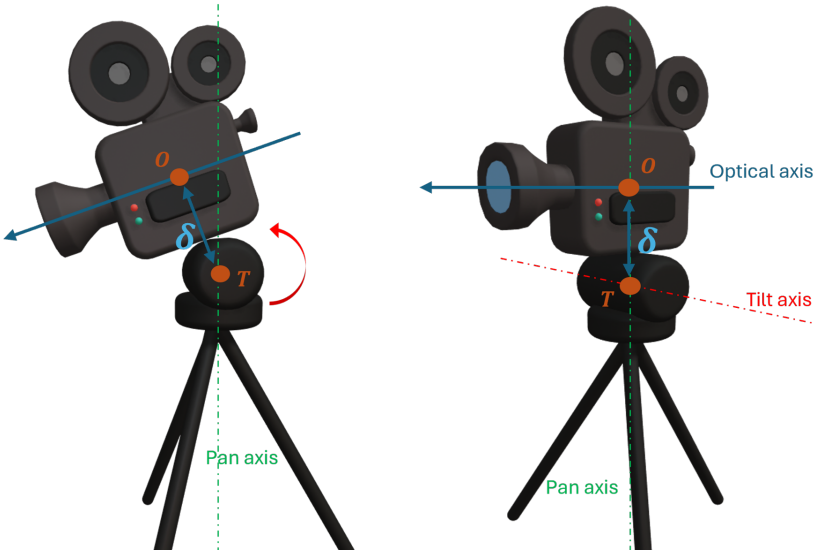


Figure 4.2: Tripod model. The center of rotation **T** remains fixed as the camera rotates, and the point **O** which belongs to the optical axis of the camera remains at a fixed δ distance of **T**.

a parameter update procedure (Section 4.3.2), and an evaluation of the tracking confidence (Section 4.3.2).

Sports field detection

Let $\mathcal{S}$ be the set of semantic classes constituent of soccer field markings, such as "left goal line", "center circle", *etc*. Each one can be viewed as a simple geometric element. The laws of the game [60] further specify their position and dimensions, such that we can define the soccer field template $\mathcal{F} : \mathcal{S} \rightarrow \mathtt{E}$, which maps each semantic class $c \in \mathcal{S}$ to its corresponding 3D geometric element $e \in \mathtt{E}$. For convenience, the projection of the soccer field markings in the camera view κ will be denoted by $\pi_\kappa(\mathcal{F})$.

To obtain 2D-3D correspondences, we only need to detect and identify soccer field elements in the videos. We leverage existing open-source sports field detectors such as the keypoint detector of Gutiérrez-Pérez and Agudo [90] that detects the intersection points of the soccer field markings. While our reinitialization algorithm uses the detected points, the tracking leverages the denser information retrieved from semantic segmentation of field markings. We rely on the semantic field markings detector of Theiner and Ewerth [69]. Semantic segmentation of field markings provides blobs of pixels towards 3D line or circle equation correspondences. As segmentation blobs provide a robust, but maybe too rich source of information, we synthesize it with the mean shift algorithm [112] to fit a set of points that approximate the segmented blob, thus generating a set of image points $\mathbf{x}_c^1, ...\mathbf{x}_c^n$ per visible soccer field marking

class c of $\mathcal{S}$.

Optical flow

As broadcast cameras can zoom in on areas of the soccer field that are sparse in terms of markings, optical flow correspondences are necessary to prevent drifting. Furthermore, these correspondences can be sampled uniformly in the image, which helps to distribute the visual support in the whole image, unlike field markings which can be condensed into small areas. We use the pyramidal version of Lucas-Kanade's algorithm [113] to retrieve N_o point matches $\left\{\mathbf{x}_{t-1}^i \leftrightarrow \mathbf{x}_t^i\right\}_{i=0}^{N_o}$ between the previous image $\mathcal{I}_{t-1}$ and the current one $\mathcal{I}_t$.

Update through non-linear optimization

Starting from the previous camera estimate κ_{t-1}, we update the camera parameters $\kappa_t : \{f, k_1, \phi, \theta, \gamma, C_x, C_y, C_z\}$ by minimizing the sum of three error functions.

First, given soccer field markings correspondences, if we denote by $\mathcal{S}^*$ the subset of $\mathcal{S}$ that is detected in the current frame $\mathcal{I}_t$, and by $\{\mathbf{x}_c^i\}_{i=0}^{N_c}$ the set of points extracted by mean shift from the segmentation maps for the soccer field marking class c, we minimize the reprojection error between each extracted point and the closest point $\mathbf{p}$ belonging to the projected soccer field element $\pi_\kappa(\mathcal{F}(c))$:

$$\mathcal{L}_\mathcal{F} = \sum_{c \in \mathcal{S}^*} \sum_{i=0}^{N_c} \rho \left(\min_{\mathbf{p} \in \pi_\kappa(\mathcal{F}(c))} \left\| \mathbf{p} - \mathbf{x}_c^i \right\|_2 \right) , \tag{4.3}$$

where $\rho(\cdot)$ is the Cauchy loss, used to filter outliers.

Secondly, given optical flow correspondences, we can filter out the point correspondences landing outside the polygon obtained by the projection of the soccer field side lines. To minimize the noise coming from the players motion, the players are detected with RTMDet [114], and the correspondences are discarded around their bounding boxes. A point correspondence $\mathbf{x}_{t-1}^i$ in frame $\mathcal{I}_{t-1}$ can be mapped to a ray by inverting the projection function: $\mathtt{L}_{t-1}^i = \pi_{\kappa_{t-1}}^{-1}(\mathbf{x}_{t-1}^i)$. As we only keep correspondences on the surface of the sports field, its intersection with the plane $Z = 0$ allows us to retrieve a 3D point $\mathbf{X}_{t-1}^i = \mathtt{L}_{t-1}^i \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}^\mathsf{T}$. As a result, for all detected optical flow correspondences, we minimize the Cauchy loss of the reprojection error:

$$\mathcal{L}_{OF} = \sum_{i=0}^{N_o} \rho \left(\left\| \pi_\kappa(\mathbf{X}_{t-1}^i) - \mathbf{x}_t^i \right\|_2 \right) . \tag{4.4}$$

The third term of our objective function is a constraint that ensures that the

camera rotation and focal point position satisfy our tripod model. We derive the position of the point $\mathbf{O}^*$ as the closest point to $\mathbf{T}$ belonging to the camera optical axis:

$$\mathbf{O}^* = \mathbf{C} + \frac{\langle (\mathbf{T} - \mathbf{C}), \mathbf{r}_3 \rangle \, \mathbf{r}_3}{\|\mathbf{r}_3\|_2^2}, \tag{4.5}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product. Since the distance between this point and the tripod rotation center should be δ meters, we define the following loss term:

$$\mathcal{L}_T = \delta - \|\mathbf{O}^* - \mathbf{T}\|_2 . \tag{4.6}$$

The previous losses are scaled and summed into a final objective function $\mathcal{L} = \mathcal{L}_\mathcal{F} + \mathcal{L}_{OF} + \omega \mathcal{L}_T$, with ω being a scalar hyperparameter, to optimize the camera parameters κ using the Levenberg-Marquardt algorithm.

Online confidence evaluation

Given the semantic segmentation of the soccer field markings S, we produce a binary map B of all the soccer field markings in the image: $B(x, y) = 1, S(x, y) = c, \forall c \in \mathcal{S}$. Considering the projection of the soccer field template $\mathcal{F}$ in the image $\pi_\kappa(\mathcal{F})$, our confidence score is the Jaccard index between the two generated masks:

$$s = \frac{|B \cap \pi_\kappa(\mathcal{F})|}{|B \cup \pi_\kappa(\mathcal{F})|} . \tag{4.7}$$

This score is used to detect when a reinitialization of the tracker is needed, and when the optical flow correspondences are to be discarded to prevent them from damaging the optimization, as it relies on the previous camera parameters κ_{t-1}.

4.3.3 Reinitialization algorithm

For initialization, or when the confidence score becomes low, we propose a reinitialization algorithm. With the capabilities of deep neural networks today, soccer field markings detection is not a very difficult task. Usually, the tracking starts drifting when the sports field markings become very sparse in the image, which is why we include a reinitialization algorithm that only needs two point correspondences.

Given two point correspondences between the image and the top-view sports field model, our reinitialization algorithm estimates the focal length, the pan, and the tilt of the camera. The focal point is set to the tripod rotation center, and the roll value is set to zero.

Then, given the image points $\mathbf{x}_i = (x_i, y_i)$ and their corresponding world points

$\mathbf{X}_i = (X_i, Y_i, Z_i)$, we derive the camera parameters $\{f, \phi, \theta\}$. The focal length f is estimated with the equations originally derived by Gedikli [115], and summarized in Appendix 6.1 of [105]. The pan and tilt angles, respectively ϕ and θ, are estimated by an iterative algorithm assuming that ϕ only affects the projection along the x-axis, and that θ only affects the projection along the y-axis. We initialize ϕ and θ by steering the optical axis towards the barycenter of the detected world points $\mathbf{X}_i$. Then we estimate updates depending on the correspondences:

$$d\phi = \sum_i \frac{1}{2} \tan^{-1}(x_i - \pi_\kappa \left(\mathbf{X}_i\right)_x, f), \tag{4.8}$$

$$d\theta = \sum_i \frac{1}{2} \tan^{-1}(y_i - \pi_\kappa \left(\mathbf{X}_i\right)_y, f), \tag{4.9}$$

with $\pi_\kappa \left(.\right)_x$ denoting the x coordinate of the point projection. The pan and tilt are thus iteratively refined as $\phi_{k+1} = \phi_k + d\phi_k$ and $\theta_{k+1} = \theta_k + d\theta_k$, with k being less or equal to 5.

If there are more than two correspondences in the image, we use RANSAC to filter out potential outliers. As a last step, the optimization described in Section 4.3.2 is performed without the $\mathcal{L}_{OF}$ error function.

4.4 Results

To validate our tracking system, we conducted several experiments. In Section 4.4.1, we first report the performance of BroadTrack on sn-gamestate according to the metrics proposed with the sn-calibration dataset. Then, in Section 4.4.2, we validate our reinitialization algorithm on the 2023 version of the sn-calibration challenge. To further validate the effects of optical flow, tripod constraint, and lens distortion, we provide an ablation study in Section 4.4.3. Finally, since the metrics are meant for single-frame camera calibration and only evaluate the quality of the reprojection in the images, in Section 4.4.4, we qualitatively illustrate the suitability of the camera parameters derived by BroadTrack.

4.4.1 Tracking system on SoccerNet-gamestate

We evaluate our system on the sn-gamestate dataset, which consists of sequences of 30 seconds, where each frame is annotated with point correspondences along the soccer field markings and goal posts. The test set contains 49 sequences of 750 images each, extracted from 3 games of the Swiss Football League.

	$JaC_5(\uparrow)$	$JaC_{10}(\uparrow)$	MRE($\downarrow$)	CR($\uparrow$)
TVCalib[69]	19.88	50.42	12.4	99.93
NBJW[90]	37.14	68.24	10.28	93.67
PTZ-SLAM[103]	25.87	45.28	27.64	26.67*
Ours (fixed **C**)	50.97	74.93	5.39	100
BroadTrack	**56.88**	**79.79**	**5.02**	**100**

Table 4.1: Comparison metrics on the sn-gamestate dataset. MRE stands for Mean Reprojection Error and is measured in pixels; CR stands for completeness rate in percent. All results are reported for HD, 1920×1080 frames. For the PTZ-SLAM method, the code provided by the authors crashes after processing about 200 frames; hence the low completeness rate (see *).

For our experiments, we use the JaC metric [1] (previously denoted by AC in the sn-calibration challenges [9, 10]), which expresses the percentage of soccer field elements that are correctly reprojected in the image, the correctness being tuned by a tolerance parameter τ in pixels. We report this metric for both 5 and 10 pixels, which are quite challenging threshold values for the HD frames of the dataset. For comparison, note that the SoccerNet camera calibration challenge evaluates this metric for 5 pixels, but for 960×540-sized images, a quarter of this dataset resolution. We also give the mean reprojection error and the completeness rate, the latter indicating the proportion of the dataset for which the technique produced camera parameters.

To estimate the tripod rotation center $\mathbf{T}$, we run our tracking system without the tripod constraint, and then optimize the tripod position and distance to the optical axis δ given the estimated camera parameters. From the keypoints detected by the neural network of Gutiérrez-Pérez and Agudo [90], we only keep the ones that are actual marks, line intersections, or line and circle (arc) intersections, *i.e.* we discard all keypoints that are derived from other geometric cues. The reinitialization algorithm is used to get the parameters of the first frame, and during the tracking, the reinitialization is performed when the score confidence score s falls under 0.5, which leads to frequent reinitialization.

During our evaluation, out of the 49 sequences of 750 frames in the test set, the reinitialization had to be used 33 times, and the loss of tracking lasted for 15 frames on average. From the result reported in Table 4.1, BroadTrack outperforms all available open-source methods. We also experimentally confirm that the modeling of a broadcast camera as a PTZ camera with a fixed focal point deters the performance of the tracking. This emphasizes the specificity of broadcast cameras, which are not well modeled by PTZ cameras, even if they are rigged on a tripod.

	$JaC_5(\uparrow)$	$JaC_{10}(\uparrow)$	$CR(\uparrow)$
TVCalib [69]	52.9	73.4	66.5
NBJW [90]	73.7	86.7	**77.5**
Ours	**75.25**	**86.8**	69.8

Table 4.2: Two-point reinitialization algorithm evaluation on the sn-calibration dataset of 2023. JaC metrics are reported in percent at 5 and 10 pixels for 960×540 frames. Our reinitialization algorithm equals or achieves the SOTA performance of the NBJW method on a slightly smaller subset of the dataset. To enable fairer comparison with TVCalib, we filter out their calibrations with $JaC_5 = 0$, because, by design, their method never signals failure.

4.4.2 Reinitialization on SoccerNet-calibration

We evaluate our reinitialization algorithm on the sn-calibration dataset. The test set comprises 3,141 images from a wide range of cameras used during soccer broadcast, including wide-angle cameras, and fish-eye cameras. The JaC evaluation of our method is computed only for views that obtain a confidence score $s < 0.2$, which lowers our completeness rate. We also expect this score condition to filter out views for which our tracking system is not especially designed, *e.g.* views from fish-eye cameras. The default focal point localizations are defined for common wide-angle cameras, that is for main, 16 meters, and high behind the goal (HBG) cameras respectively as $\mathbf{C}_{main} = (0, 55, -12)$, $\mathbf{C}_{16m} = (\pm 36, 55, -12)$, $\mathbf{C}_{HBG} = (-65, 0, -15)$, all expressed in meters in the sn-calibration world reference system [116].

From the results reported in Table 4.2, we establish that our algorithm can reach state-of-the-art performance on a smaller part of the dataset. Since the camera diversity of the dataset is higher than the one we design our system for, the high performance of the reinitialization part comforts our strategy of frequent reinitialization.

4.4.3 Ablation study

We conduct our ablation study on the sn-gamestate test set. To run our algorithm without any prior knowledge of the tripod rotation center, we choose a default position as input to our initialization algorithm. We arbitrarily set the camera focal point to a default main location in $\mathbf{C} = (0, 55, -12)$ meters. As shown in Table 4.3, the biggest improvement in performance comes from the inclusion of radial distortion in the camera model, which legitimates our concerns about the previous datasets based on homographies and confirms our choice of not using their annotations as it would lead to an unfair evaluation. It is also worth noticing that

OF	**T**	k_1	$JaC_5(\uparrow)$	$JaC_{10}(\uparrow)$	$MRE(\downarrow)$	$MedRE(\downarrow)$
✘	✘	✘	42.09	74.09	5.74	3.13
✘	✘	✔	54.1	78.35	5.04	2.47
✔	✘	✔	55.96	79.07	**4.85**	2.43
✘	✔	✔	54.9	78.99	4.95	2.44
✔	✔	✔	**56.88**	**79.79**	5.02	**2.37**

Table 4.3: Ablation study on sn-gamestate. JaC metrics are reported in percent at 5 and 10 pixels for HD frames. Mean (MRE) and Median (MedRE) Reprojection Errors are reported in pixels. OF stands for optical flow, $\boldsymbol{T}$ is for our tripod constraint, and finally k_1 represents the inclusion of radial distortion in our model.

the optimization procedure starting from the previous camera estimate does not lead to much of a performance boost when compared with the NBJW method. This demonstrates that their strategy of deriving virtual keypoints pays off, even if it is at the expense of physical modeling of the camera parameters, as discussed in Section 4.4.4. Both optical flow and tripod constraints demonstrate a smaller contribution to the performance of our algorithm according to the reported metrics of Table 4.3. We argue and show in the next section that their contribution is only partially reflected by the single-view metrics used until now. In the next section, we show their benefits through qualitative evaluation and comparisons.

4.4.4 Physical soundness

As explained in Section 4.3.1, professional pan-tilt heads are made to ensure the smooth motion of the camera. This means that we expect pan and tilt values to be particularly smooth over time. We also expect our tripod constraint to make the camera focal points more localized, even if, per se, there is no restriction in terms of distance to the tripod. These intuitions are confirmed by our visualization of Figure 4.3 and Figure 4.4. We notice that, while the focal lengths estimated by BroadTrack are smoother or display lower variations than the other methods, it still shows some high-frequency variations. We explain that because of a vertigo effect, as the scene is far away, some uncertainty in terms of position can be compensated by a focal length adjustment, and conversely, without displaying perspective distortions.

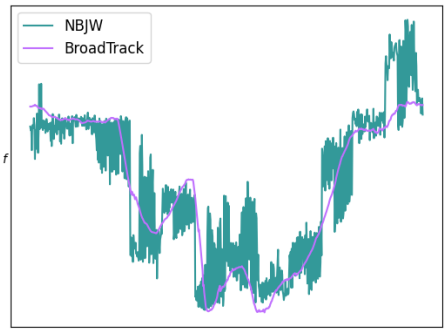
4.4.5 Long-term tracking

To further demonstrate the performance of BroadTrack, we show results on 20 minutes (60,000 frames) of the main camera footage taken from a game of the

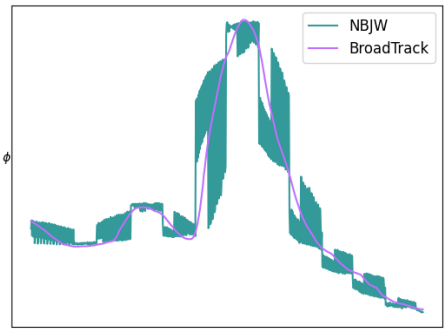
German Bundesliga. As the stadium is much bigger than the ones of SoccerNet, we set the default focal point position $\mathbf{C}_{default} = (0, 90, -18)$, and we use a commercial tool for keypoints and markings detections [6]. To derive the tripod position, we run our system without the tripod constraint on the first 5,000 frames, and perform the optimization procedure as described in Section 4.4.1 to derive $\mathbf{T}$ and δ. This sequence is not annotated; hence, only the overlay of the soccer field projection provides a qualitative assessment. As displayed in Figure 4.5, the projection of the soccer field overlays almost perfectly on top of the actual field markings. BroadTrack maintains the same quality of overlay for the complete video; this illustrative video is given in the supplementary material. Moreover, the reinitialization step is only performed 60 times, and lasts for 4 frames on average. We explain this improvement by the quality of the commercial keypoints and markings detection.

4.5 Conclusions

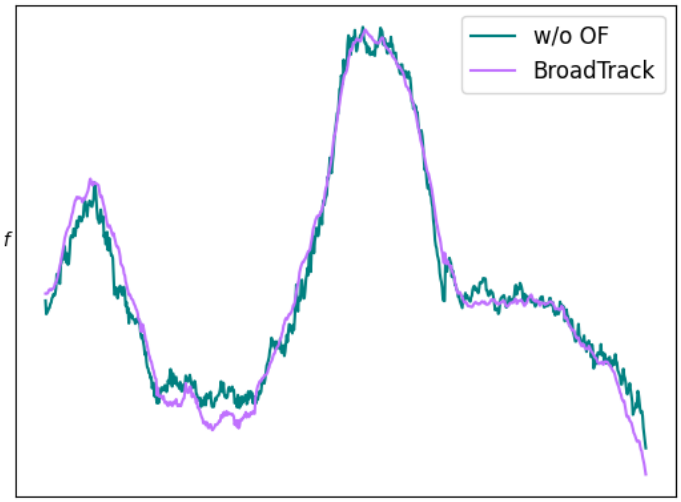
In this paper, we have presented BroadTrack, a new tracking system specially designed for wide-angle broadcast cameras. Through diverse qualitative and quantitative evaluation, we show that our system is both accurate and robust, outperforming available solutions in both aspects. Our results also corroborate the suitability of our broadcast camera lens and tripod models, motivating further exploration and refinement. Future works entail the dynamic incorporation of the tripod constraint, even if BroadTrack obtains convincing results with a roughly appropriate focal point. Another exciting piece of research lies in the vertigo effect noticed with the high-frequency variation of the focal length, as we believe modeling this effect might lead to even better focal point position.



(a) Focal length value variations.

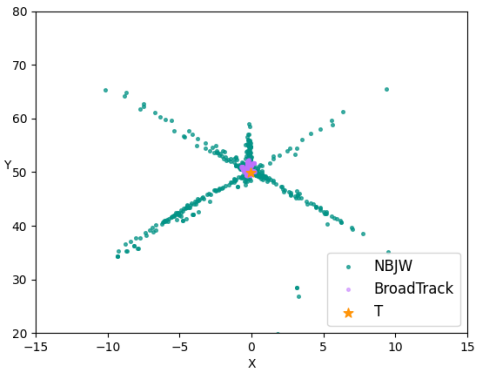


(b) Pan value variations.

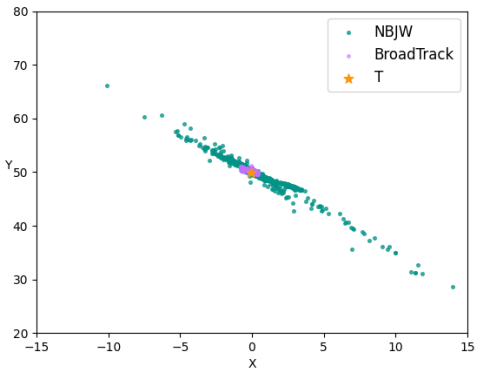


(c) Effect of optical flow on focal length stability.

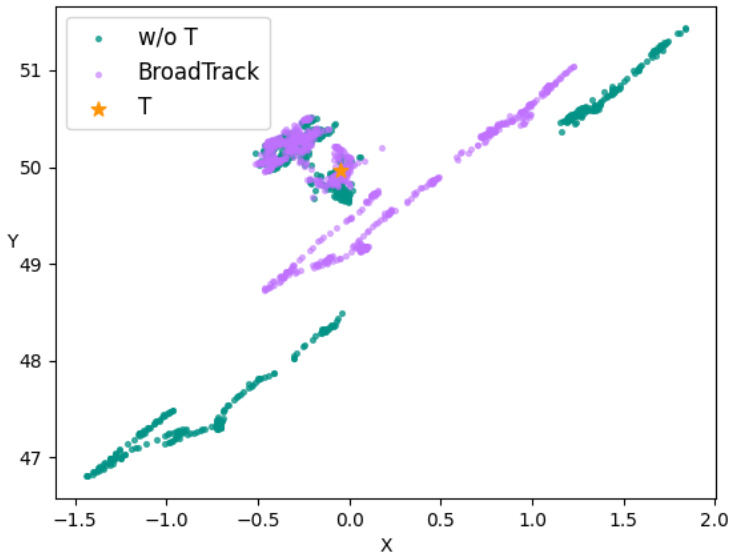
Figure 4.3: Temporal evolution of camera parameters along sequences of the sn-gamestate dataset. Figures 4.3a and 4.3b show that the parameters estimated by NBJW exhibit stronger jitter than those produced by BroadTrack. Figure 4.3c further demonstrates that optical flow contributes to smoothing focal length estimates over time.



(a) Sequence 1.



(b) Sequence 2.



(c) Effect of the tripod constraint.

Figure 4.4: Estimated focal point positions in the XY plane. Figures 4.4a and 4.4b show that the focal point estimated by NBJW can travel up to 20 m within a single sequence, whereas the position estimated by BroadTrack remains concentrated around the tripod location. Figure 4.4c highlights the benefit of enforcing the tripod constraint, which keeps the estimated camera position closer to the inferred center of rotation $\mathbf{T}$.



Figure 4.5: Qualitative evaluation of the 20 minutes sequence from the Bundesliga. One out of 10,000 images is displayed. Soccer field markings are reprojected in red using the estimated camera parameters κ.

Chapter 5

Leveraging Circular Markings (3^{rd} article)

This chapter reproduces the following paper: Floriane Magera et al. "Can Geometry Save Central Views for Sports Field Registration?" In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). *Nashville, Tennessee, USA: IEEE, June 2025, pp. 6060–6069*

Prologue

In the previous works, we worked towards achieving the highest accuracy possible for camera calibration in sports. In the remainder of this thesis, we instead seek to push the limits of camera calibration under challenging conditions.

This work marks the first step of this shift. It starts from the observation that circular markings are frequently overlooked in camera calibration pipelines, despite being, in some cases, the main source of information present in the image. This is particularly striking for zoomed-in central views of the playing surface in sports like football, hockey, or basketball.

This observation motivated the investigation of neglected geometric cues present in these challenging images: we investigated (1) how circle-conic correspondences could be exploited and furthermore, (2) whether the thickness of circular markings could provide a valuable source of information.

The outcome of this investigation is an explicit geometric approach that derives novel point-to-point and line-to-line correspondences from the detection of circular markings in the image. Moreover, by casting circle-to-conic correspondences into multiple correspondences of simpler geometric primitives, our approach can be integrated seamlessly into existing camera calibration pipelines.

Can Geometry Save Central Views for Sports Field Registration?

Abstract

Single-frame sports field registration often serves as the foundation for extracting 3D information from broadcast videos, enabling applications related to sports analytics, refereeing, or fan engagement. As sports fields have rigorous specifications in terms of shape and dimensions of their line, circle and point components, sports field markings are commonly used as calibration targets for this task. However, because of the sparse and uneven distribution of field markings, close-up camera views around central areas of the field often depict only line and circle markings. On these views, sports field registration is challenging for the vast majority of existing methods, as they focus on leveraging line field markings and their intersections. It is indeed a challenge to include circle correspondences in a set of linear equations. In this work, we propose a novel method to derive a set of points and lines from circle correspondences, enabling the exploitation of circle correspondences for both sports field registration and image annotation. In our experiments, we illustrate the benefits of our bottom-up geometric method against top-performing detectors and show that our method successfully complements them, enabling sports field registration in difficult scenarios.

5.1 Introduction

In the realm of sports analytics, sports field registration methods are key to interpreting the game from the broadcast feeds. Accurately projecting any 3D real-world sport field point onto the 2D image plane enables numerous sport analytics applications. Indeed, extracting statistics about player performances [117], actions such as goals, offsides, or fouls [47], or the game state [8] relies on the ability to locate the ball and players on the sports field [118].

An advantageous feature of sports fields is their fixed dimensions [61, 63, 82], which makes them convenient calibration targets. However, the sports field markings might be sparsely distributed in some areas, making the single-frame camera calibration particularly challenging, as shown in Fig. 5.1. For this reason, sports field registration, *i.e.* homography estimation methods, are often considered as a proxy to camera calibration in sports. In fact, only four non-collinear point correspondences are

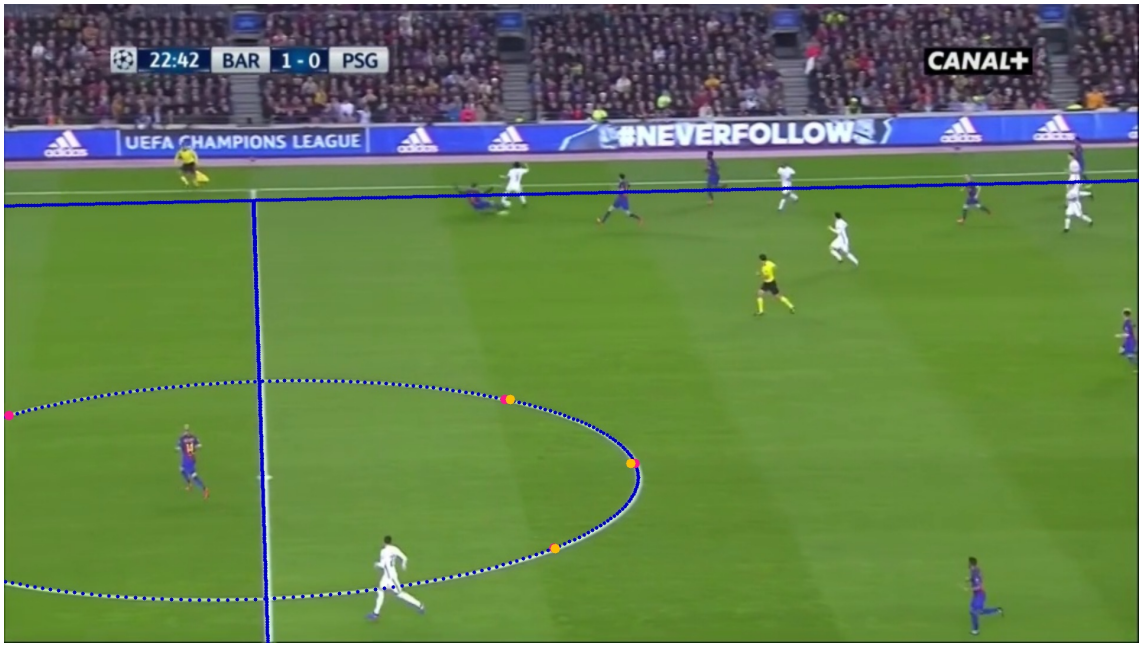


Figure 5.1: Single-frame camera calibration is challenging when only a limited amount of field markings are visible in the scene. For example, in this image, camera parameters cannot be derived from the line and point correspondences without incorporating priors about the camera. This is why current methods generate virtual keypoints (PnLCalib [119] in yellow) which here, however, are not sufficient to allow for a registration of the sports field. Our geometric approach derives the pink points from the fitted ellipse and the detected center, which allow for the estimation of the homography (blue).

needed to estimate the homography matrix that maps the sports field top-view template to its image, while four to six points are required to estimate the projection matrix of a pinhole camera depending on the PnP solver used [13].

Nevertheless, four non-collinear point correspondences may not always be found directly in the image, as broadcast cameras often happen to zoom in on an action taking place nearby a mostly isolated circular marking on the sports field, usually around the center circle (as in Fig. 5.1). Fortunately, we show that there are several options to handle those difficult views. In this work, we propose a solution that enables the registration of such zoomed-in camera views. For that, we take a geometrical, constructive approach that leverages the properties of circle projections to enable the retrieval of point and line correspondences.

We will also explore a more difficult case, when only a small part of the circle is visible, by exploiting the constant thickness of the sports field markings on the ground.

Contributions. We summarize our contributions as follows: **(i)** We introduce a

novel method (relying on three variants depending on image support) to derive the set of primitives –points and lines– that are necessary to estimate camera parameters for the challenging central view case. In this way, we are broadening the spectrum of scenes that can be calibrated. **(ii)** Besides enabling the calibration of difficult views, our method offers a precise way to annotate point correspondences along circles. Finally, **(iii)** we investigate the potential of our method on the SoccerNet dataset [81] and on challenging cases through qualitative and quantitative experiments.

5.2 Related work

With the recent SoccerNet Game State Recognition challenge [8] and the Roboflow tutorial on soccer analytics [120], a range of applications creating a radar view of the game with player tracking has flourished, with varying degrees of accuracy. A common flaw of these solutions is the visible shift of the players' position when the camera targets the central view of the soccer field, indicating a problem in soccer field registration for central views. All single-frame methods for camera calibration in sports heavily rely on the visibility of sports field markings in the camera view. Most methods leverage either the detection or the segmentation of such markings to derive the homography that maps the camera view to the template of the sports field (depicted in Fig. 5.2). Difficulties arise from the fact that sports field markings are typically composed of a sparse collection of heterogeneous primitives, such as points, lines, and circles, unevenly distributed across the playfield. This issue makes camera calibration in sports challenging for narrow-angle camera views, especially for central views where only the halfway line and center circle are visible.

Over the years, several methods have tried to solve the problems of sports field registration or camera calibration in sports by addressing the potential lack of visible markings in the image, as described hereafter.

Dictionary-based methods. A common approach is to enumerate all plausible camera views of the sports field markings in a dictionary. At detection time, the closest sample from the dictionary is retrieved to initialize the camera parameters and these parameters are later refined based on the field markings detection in the image [7, 15, 72, 96, 97]. These solutions usually assume that the set of camera parameters or homographies encountered in practice are well-known. They also tend to be quite slow in practice, as the size of the dictionary is proportional to the range of cameras to be calibrated *e.g.*, 450 samples [7].

Regression methods. Starting from a first homography estimated with a deep neural network [70, 73, 98], *i.e.* learned homography priors, or from explicit camera

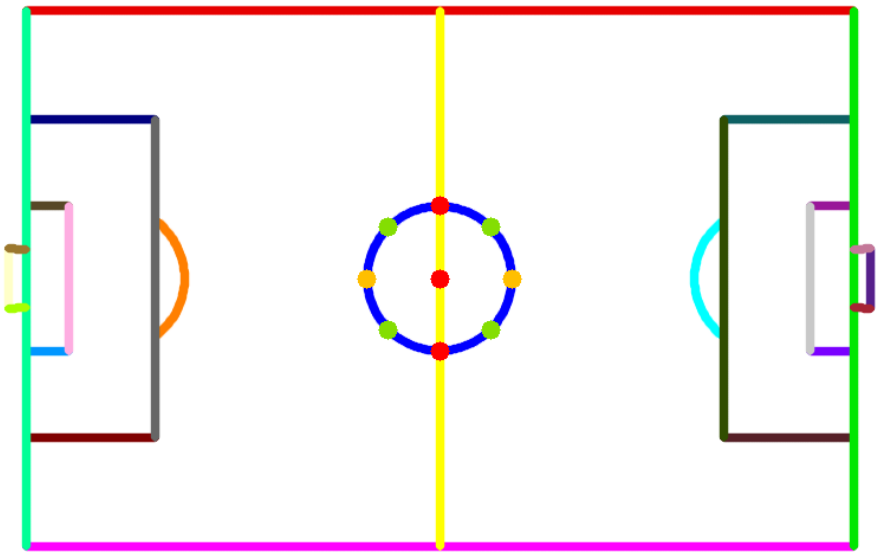


Figure 5.2: Top-view of the soccer field layout, as decomposed into different semantic elements. These elements form a set of markings include points, lines, and circles that can be used to calibrate a camera.

parameters priors [69], these methods refine their initial estimate using deep neural networks. While the initial work on which some of these methods build estimates the relative homography between frames [18], a direct sports field registration through homography regression is closer to an APR task.

In their work, Sattler *et al.* [17] show that convolutional neural networks trained to solve the APR task fail to generalize beyond their training sets, and rather perform pose approximation via image retrieval than actual pose estimation. Given that sports datasets are scarce, consisting of few images, and generally only depict top-tier games, we fear these methods will not generalize well and will only work for top-tier main cameras.

Virtual augmentation of keypoints. Some solutions augment identifiable sports field landmarks with additional virtual keypoints which are then used to train neural networks. These extra keypoints may be: (1) a grid of regularly spaced keypoints of the whole sports field plane [50, 74, 75, 79], or (2) points on sports field circles, typically regularly spaced *e.g.*, every 45 degrees [90, 92, 119, 121].

Our work specifically addresses the potential shortcomings of these two approaches. First, those rely on the generalization capabilities of deep neural networks to detect points that do not have anything distinguishable but their location; there is no pattern to be learned for the keypoints mentioned in (1), they correspond to texture-less grass, and for the keypoints mentioned in (2), they correspond to a specific location

along a curve, whose appearance is a function of the perspectivity. The ability of a neural network to detect those relies on its ability to grasp the perspectivity of the image, which rather relates to an APR task than a detection or matching task. We hypothesize that the generalization capabilities of the neural networks detecting virtual keypoints may suffer from the same problem as the above-mentioned regression methods and thus be restricted to top-tier main cameras. Second, these approaches need to register the sports field on these images to create annotations, which is always done through homography estimation. And, as previously mentioned, only images already containing a sufficient number of primitives can be annotated, while these additional points are needed the most when we have fewer than four point or line correspondences. There is thus a chicken-and-egg problem here.

Specific solutions for central views. Some previous work directly address circle correspondences in their method: Gupta *et al.* [100] leverage ellipses through their pole-polar relationship with already found point or line matches, thus enforcing geometric constraints to estimate homographies for hockey rink registration. Alvarez and Caselles [122] derive a closed-form solution for homography estimation between a soccer field and an image using correspondences of the center circle, the halfway line, and up to three points belonging to this line. Our method is similar in the sense of geometric exploitation of circle correspondences but targets a wider scope than homography estimation: the correspondences that we output also allow camera parameters estimation and keypoint annotation.

Approximations with the great axis of the ellipse. Other similar works make relatively strong assumptions about the type of perspectivity at hand. Andrews *et al.* [93] derive the vanishing point of the middle line and its parallels, then link it to the extremities of the ellipse great axis to derive additional point and line correspondences. Cuevas *et al.* [91] match the great ellipse axis extremities with the circle points in orange as displayed in Fig. 5.2, which is also the approach chosen in the Roboflow tutorial [120]. However, the projection of the circle center does not necessarily land on the ellipse center. Hence, the matching between the extremities of the great ellipse axis and the orange points in Fig. 5.2 is an approximation whose negative effect is depicted in Fig. 5.3.

Generic ellipse-based geometric methods. Beyond sports applications, generic solutions for homography estimation or camera calibration with ellipse correspondences have been studied. Chum and Matas [123] linearize ellipse correspondences using a first-order Taylor expansion of the homography mapping around two or more ellipse correspondences. Chen and Wu [124] derive the focal length and extrinsic camera parameters from two coplanar circle correspondences. Huang *et al.* [125]

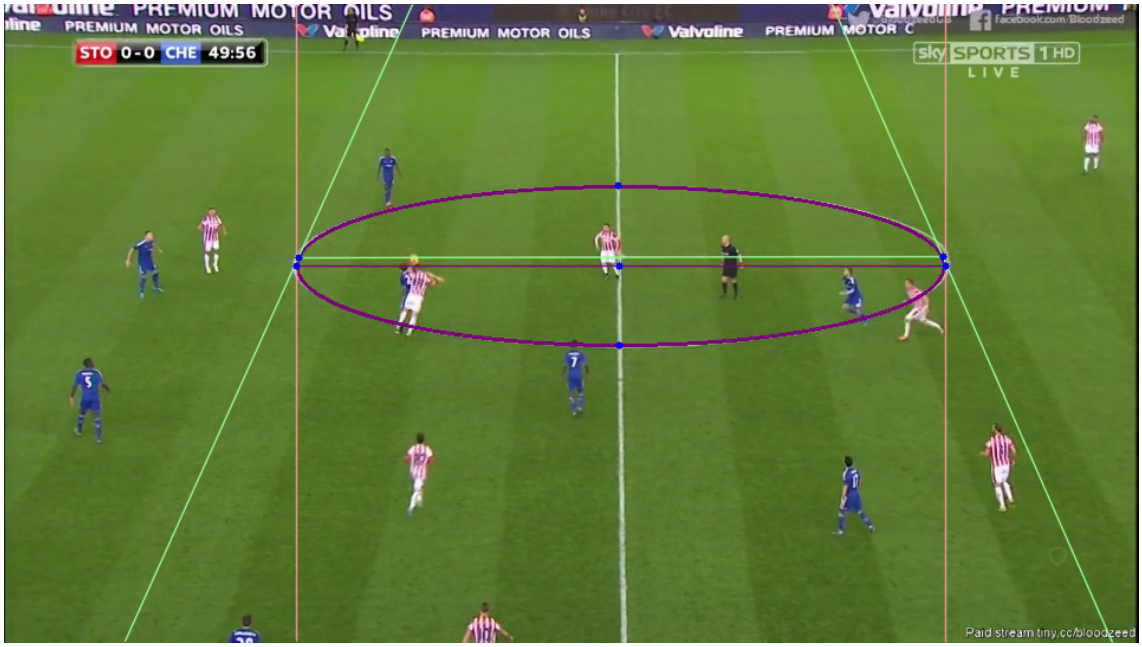


Figure 5.3: Visualization of the fitted ellipse (purple), its tangents from the great axis extremities (orange), or from the points extracted by our proposed method (green), that come from the line crossing the actual image of the circle center. The green tangents better capture the perspectivity as parallels should intersect at one point on the vanishing line of the supporting plane.

leverage the properties of the common self-polar triangle of separate ellipses to derive homographies. Finally, Huang *et al*. [126] recover the camera intrinsic parameters by leveraging concentric circles correspondences in three views.

Inspired by generic solutions using ellipse correspondences, we investigated the use of a fully geometric bottom-up approach to derive point and line correspondences from circle ones, and their application to camera calibration to address the challenging central views in soccer.

5.3 Method

Preliminaries. The goal of our method is to convert a circle correspondence into several point and line correspondences that can be integrated in the DLT usually done to estimate a homography or in a PnP solver to extract camera parameters. Since many solutions to sports field markings segmentation and detection exist in the literature, we take off-the-shelf detectors, which are already capable of segmenting circular markings and detecting remarkable keypoints on sports fields, including the center of these circles if they are an integral part of the markings.

The sole detection of a circle does not suffice for calibration. Hence, starting from

the detection of a circle correspondence, we developed three cases. The first two combine the circle correspondence and the detected projected circle center to: **Case 1** (Section 5.3.1) a line correspondence, **Case 2** (Section 5.3.2) a point correspondence, and the last one, **Case 3** (Section 5.3.3) is a fallback in case of unknown projection of the circle center. In the following, we start by defining the geometrical notions that will be necessary to derive a sufficient amount of linear constraints to estimate the homography.

Geometric principles in the top-view. Let us consider the supporting plane of the sports field π from the 2D projective plane $\mathbb{P}^2$ (notations and equations are taken from [11]). In the reference system of π, world geometric elements are described by their 2D homogeneous coordinates. Let $\mathtt{C}$ be the world circle on the sports field, which is centered at location $\mathbf{o} = (o_x, o_y, 1)^\mathsf{T}$, and has a radius r.

A point $\mathbf{x}$ expressed in homogeneous coordinates $(x, y, 1)^\mathsf{T}$ belongs to the circle if

$$\mathbf{x}^\mathsf{T}\mathtt{C}\mathbf{x} = 0\,, \tag{5.1}$$

where the matrix representation of $\mathtt{C}$ is given by

$$\mathtt{C} = \begin{bmatrix} 1 & 0 & -o_x \\ 0 & 1 & -o_y \\ -o_x & -o_y & o_x^2 + o_y^2 - r^2 \end{bmatrix}. \tag{5.2}$$

A useful property of conics is the pole-polar relationship defined by a conic. It allows to map a point $\mathbf{p}$, the pole, and the *polar* (line) $\mathbf{l}$ with respect to conic $\mathtt{C}$ by $\mathbf{l} = \mathtt{C}\,\mathbf{p}$ [11, page 58]. The polar of the circle's center is the line at infinity of the supporting plane π. Indeed,

$$\mathtt{C}\,\mathbf{o} = \begin{bmatrix} 1 & 0 & -o_x \\ 0 & 1 & -o_y \\ -o_x & -o_y & o_x^2 + o_y^2 - r^2 \end{bmatrix} \begin{pmatrix} o_x \\ o_y \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -r^2 \end{pmatrix} = -r^2 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = -r^2\,\mathbf{l}_{\text{inf}}. \tag{5.3}$$

An important property of pole-polar relationships is their preservation through projective transformations. This means that the estimation of the imaged center of the circle in our image is key to deriving the projection of the vanishing line of π in the image.

Therefore, starting from the ellipse $\mathtt{E} = \mathtt{H}^{-T}\mathtt{C}\,\mathtt{H}^{-1}$ that corresponds to the projection of the circle, and the projection of its center: $\mathbf{c} = \mathtt{H}\,\mathbf{o}$, we estimate the vanishing

line:

$$\mathbf{l}_{\text{vanishing}} = \mathtt{E}\,\mathbf{c}. \tag{5.4}$$

In the next subsection, we explain how to derive a sufficient number of correspondences based on one more primitive to estimate a homography matrix.

5.3.1 Case 1: Line correspondence

Given a known line $\mathbf{l}_1$ detected in the image, we can estimate the vanishing point $\mathbf{v}$ linked to all parallels to $\mathbf{l}_1$ by

$$\mathbf{v} = \mathbf{l}_1 \times \mathbf{l}_{\text{vanishing}}. \tag{5.5}$$

Once that $\mathbf{v}$ is known, it is easy to find the parallel to $\mathbf{l}_1$ passing through the projected circle center $\mathbf{c}$:

$$\mathbf{l}_2 = \mathbf{v} \times \mathbf{c}. \tag{5.6}$$

As illustrated in the left part of Fig. 5.4, this line $\mathbf{l}_2$ intersects the ellipse in two points $\mathbf{a}$ and $\mathbf{b}$, which are our first two point correspondences. Then, we define the polar of $\mathbf{v}$,

$$\mathbf{l}_3 = \mathtt{E}\,\mathbf{v}\,, \tag{5.7}$$

which passes through the projected circle center $\mathbf{c}$, and intersects the ellipse in two new points $\mathbf{d}$ and $\mathbf{e}$ because $\mathbf{v}$ and $\mathbf{c}$ are conjugates. Furthermore, by definition, the polars of $\mathbf{d}$ and $\mathbf{e}$ are tangents to the ellipse and intersect at $\mathbf{v}$; they are thus parallel to $\mathbf{l}_1$, and perpendicular to $\mathbf{l}_3$. As the equation of the line $\mathbf{l}_1$ is known in the sports field template, it is direct to derive the 2D position of each derived point.

5.3.2 Case 2: Point correspondence

Given a point correspondence $\mathbf{x}$, we can derive the line that passes through that point and the projected center of the circle:

$$\mathbf{l}_1 = \mathbf{x} \times \mathbf{c}. \tag{5.8}$$

As can be seen in the right part of Fig. 5.4, besides intersecting the ellipse in two points $\mathbf{a}$ and $\mathbf{b}$, the line $\mathbf{l}_1$ is in a pole-polar relationship with a point on the vanishing line of the plane

$$\mathbf{v} = \mathtt{E}\,\mathbf{l}_1. \tag{5.9}$$

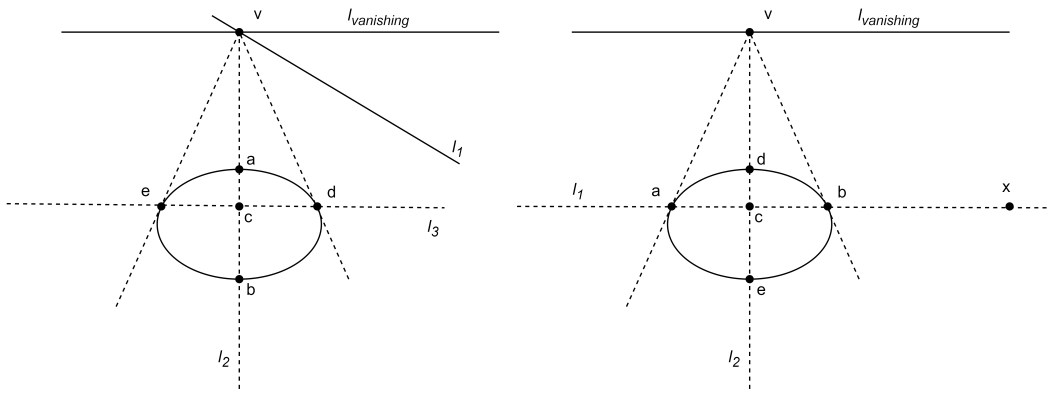


Figure 5.4: Starting either from a line correspondence $\mathbf{l}_1$ (left) or a point correspondence $\mathbf{x}$ (right), we derive the dashed lines and the set of points $\mathbf{a}$, $\mathbf{b}$, $\mathbf{d}$, and $\mathbf{e}$.

As $\mathbf{l}_1$ intersects the ellipse $\mathtt{E}$ in two points, the tangents to these two points $\mathbf{a}$ and $\mathbf{b}$ intersect at $\mathbf{v}$. The point at infinity $\mathbf{v}$ is the vanishing point of the parallels to the tangents of $\mathbf{a}$ and $\mathbf{b}$, and can be used to retrieve another parallel that passes through the projected circle center :

$$\mathbf{l}_2 = \mathbf{v} \times \mathbf{c}\,, \tag{5.10}$$

that intersects the ellipse in two points $\mathbf{d}$ and $\mathbf{e}$ and, as the tangents of $\mathbf{a}$ and $\mathbf{b}$ are parallel to $\mathbf{l}_2$, the line $\mathbf{l}_2$ is a perpendicular to $\mathbf{l}_1$. The exact position of the derived points can be easily retrieved from the known position of $\mathbf{x}$ and the circle in the sports field template.

While this method works for a generic registration case, its application to soccer is very limited, as the only point correspondences available in central views all lie on the middle line. In this instance, the first Equation (5.8) deriving a line through the point correspondence and the circle center retrieves the equation of the middle line, and thus the method falls back on the first case of Section 5.3.1. We will thus not investigate this method further in our paper.

5.3.3 Case 3: Unknown center

In some cases, the detection of the center of the circle might fail, due to occlusion with players, poor markings' definition in the image, or its absence in the sports field markings. While projective transformations deform a circle into an ellipse, the center of the ellipse is not the projection of the center of the original circle. The estimation of the original center of the circle is not straightforward, even if the ellipse equation is known in the image.

To recover the imaged center, we leverage the fact that the field markings have a consistent thickness of about ten centimeters and can be considered defining two concentric circles at their interface with the ground.

Once ellipses are detected in the image, we rely on a result of Huang *et al.* [126] to retrieve the imaged center. Let $\mathtt{C}_1$ and $\mathtt{C}_2$ be the concentric circles, and let $\mathtt{E}_1$ and $\mathtt{E}_2$ be their respective images. These circles share a common pole-polar, as their common center maps the vanishing line of the supporting plane π. In their work, Huang *et al.* showed that the image of the circles' center $\mathbf{c} = \mathtt{H}\,\mathbf{o}$ is an eigenvector of the ${\mathtt{E}_2}^{-1}\mathtt{E}_1$ matrix. By estimating the ellipses' coefficient matrices $\mathtt{E}_1$ and $\mathtt{E}_2$, we first derive $\mathbf{c}$ as it is one of the eigenvectors of ${\mathtt{E}_2}^{-1}\mathtt{E}_1$. Then, $\mathbf{l}_{\text{vanishing}}$ is obtained as the polar of $\mathbf{c}$: $\mathbf{l}_{\text{vanishing}} = \mathtt{E}_1\,\mathbf{c} = \lambda\mathtt{E}_2\,\mathbf{c}$. Subsequently, depending on the support present in the image, we can derive a sufficient number of correspondences by following the methods developed in Sections 5.3.1 and 5.3.2.

Eventually, for all the aforementioned cases, the ambiguity about left-right and up-down correspondences on the circle are solved based on priors on the camera position. Note that the trapezoid formed by the tangents of the points **a**, **b**, **d**, and **e**, has diagonals that intersect the ellipse in the green points from Fig. 5.2. In total, our method gives 8 point correspondences, which can be augmented by 8 lines from their tangents to the ellipse.

5.4 Experiments

In Section 5.4.1, we will first validate our hypothesis that the previously mentioned virtual augmentation of keypoints method fails to generalize to zoomed-in on the center circle, out-of-domain views. In Sections 5.4.2 and 5.4.3, we assess the quality of our method from a circle and a line correspondence. Finally, in Section 5.4.4, we assess the robustness of the variant of our method that can deal with missing circle center detection. As our method consists of a correspondence extractor, we will combine it with the calibration algorithm of PnLCalib [119] to obtain the final calibrations.

5.4.1 PnLCalib keypoint detector

A hypothesis on which this work relies is that virtual keypoint neural networks will fail to generalize to Out-of-Domain (OOD) images. In this experiment, we evaluate the keypoint detector of the best performing camera calibration method on SoccerNet, PnLCalib [119]. For this experiment, we gathered an OOD test set of

300 central views of empty, smaller stadiums, with lower viewpoints, without ground truth. We make comparisons with the SoccerNet dataset (SN), which consists of the images of the test sets of the *sn-calibration* and *sn-gamestate* datasets where i) the central circle is visible, and ii) the number of visible lines is lower than 3; this test set consists of 8,565 images. A way to validate the correctness of predicted keypoints without ground truth is to verify the alignment of points that are supposed to belong to the same line.

There are 8 points predicted by PnLCalib neural network on the center circle, of which we discard the points that are the intersection of the center circle and the middle line (red points in Fig. 5.2). Then, from each pair of opposed predicted points on the circle, we estimate a line. The center of this circle is supposed to belong to this line, and thus we measure the distance between the center and its projection on the line to assess the soundness of the predicted keypoints. The results are depicted in Fig. 5.5. While the error distribution appears bounded on the SoccerNet dataset, the out-of-distribution images are more subject to high error rates. By construction, our geometric method derives pairs of points on the ellipse that are collinear with the circle center.

5.4.2 Qualitative results

Starting from the semantic segmentation detector of TVCalib [69], we estimate the ellipse parameters. The center and the intersection of the center circle with the middle line are estimated thanks to the keypoint detector of PnLCalib [119]. From these detections, we derive point and line correspondences using the method described in Section 5.3.1. A series of results are displayed in Fig. 5.6, while the effect of including these points in a full camera calibration pipeline is depicted in Fig. 5.7. The results shown in Figs. 5.6 and 5.7 indicate that, if the initial detector that provides the circle correspondence and the distinguishable keypoints are performing well, our geometric approach gives promising results, enabling to calibrate cameras in challenging cases, or improving the estimated perspectivity as shown in Fig. 5.7.

5.4.3 Quantitative Results

The difficulty to showing the benefits of our methods using a quantitative analysis is that, as previously mentioned, it tackles mainly out-of-domain images that are by definition not in the annotated datasets. For the sake of completeness, however, we conducted a quantitative analysis on the SoccerNet dataset. For this setup, given the validation of the predicted keypoint correctness on the SoccerNet dataset of

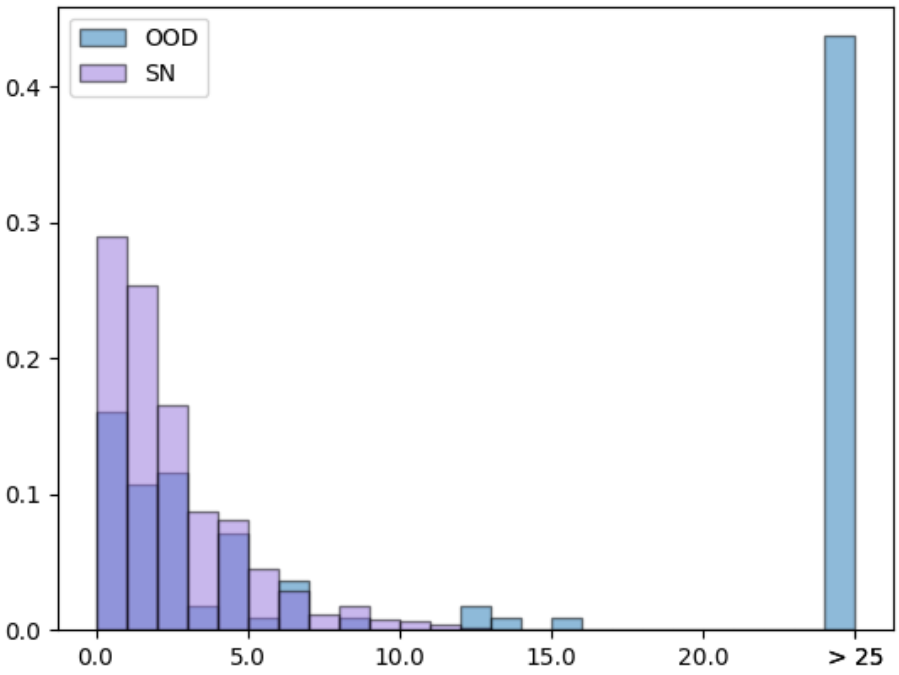


Figure 5.5: Distribution of Euclidean distances, in pixels, between the circle center and each line formed by opposed keypoints pairs. Compared to SoccerNet (SN), the OOD dataset shows a significant number of unbounded errors.



Figure 5.6: Various results of our method extracting lines and points from the center circle and the middle line correspondences. The images come from either our OOD dataset, or SoccerNet. The few failures are due to poor ellipse fitting or to bad center detection.



Figure 5.7: Left-column: vanilla PnLCalib results on OOD frames. Right-column: PnLCalib calibration when our keypoints are included.

	JaC_5 (↑) [%]	MRE (↓) [px]	CR (↑) [%]
TVCalib [69]	10.2	28.8	**100**
PnLCalib [119]	**22.4**	**12.8**	79.5
PnLCalib*	**22.4**	12.9	79.5

Table 5.1: Jaccard Index for camera calibration at 5 pixels (JaC_5) [1] in full-HD frames, Mean Reprojection Error (MRE) in pixels and Completeness Rate (CR) of previous works and PnLCalib* tuned with our geometrically derived keypoints instead of their learned version, evaluated on the central views of SoccerNet test sets.

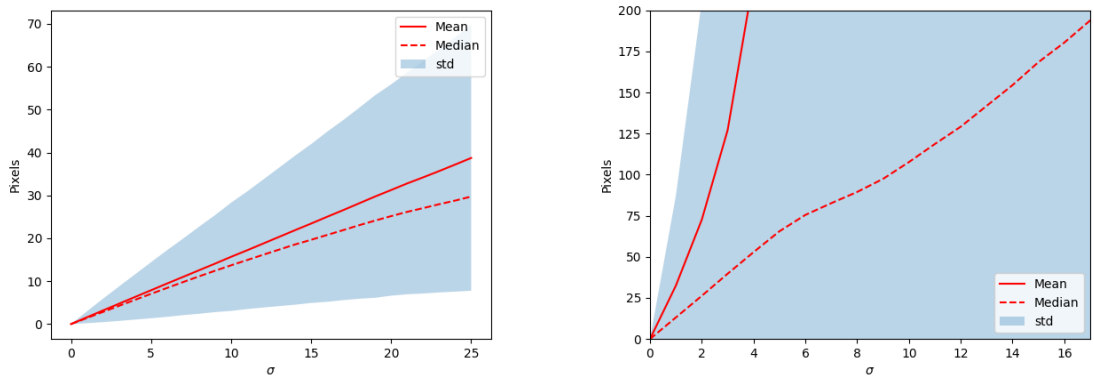
Section 5.4.1, we do not expect our geometric approach to change the results of PnLCalib when we replace their predicted keypoints by our geometric keypoints. The results are reported in Table 5.1, and as expected, no significant change is to be observed with our keypoints.

5.4.4 Unknown center

We will now assess our method that allows calibration even when the circle center is not detected in the image. In this experimental setup, we assess our method independently of sports field markings detectors.

Experimental setup. We randomly generate plausible views of soccer main cameras. To do so, we sample common camera positions, orientations and focal length values, and keep the images displaying the center circle and the halfway line only. The pinhole parameters of these cameras are then converted to homographies mapping the synthetic image to the top-view soccer field template with the playfield located in the plane $Z = 0$. These ground-truth homographies $\mathtt{H}$ are compared to the estimated homographies $\hat{\mathtt{H}}$ through the reprojection error of the visible 3D points belonging to the soccer field in the synthetic image. Specifically, given a set of 3D visible points $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ in a synthetic frame, we define the mean reprojection error in the image as: $\text{MRE} = \frac{1}{N} \sum_i^N \left\| \mathtt{H}\,\mathbf{x}_i - \hat{\mathtt{H}}\mathbf{x}_i \right\|_2$.

Results on synthetic data. In our simulations, assuming a field marking thickness of 8 centimeters, we sample 8 points along the inner and outer ellipses corresponding to the central concentric circles. Those points are projected in the full-HD (1080p) image, and a zero-centered Gaussian noise with increasing standard deviation σ is applied independently to each point. The ellipses are fitted with Fitzgibbon's method [127], then our method is applied to derive the final homography. In our synthetic data experiments, we vary the standard deviation σ between 0 and 25 pixels for 100 randomly generated cameras. As it can be observed in Fig. 5.8a, we find that the level of error seems to linearly follow the noise applied to the ellipse



(a) (b)

Figure 5.8: Reprojection errors when points are modified with a Gaussian noise with zero mean and standard deviation σ. (a) corresponds to an increasing noise applied to ellipse points, and (b) to an increasing noise applied to the circle center estimate.

points. In a second experiment, we measure the sensitivity of our method to the location of the circle center. To do so, we apply the same noise to the 2D position of the center. Figure 5.8b shows that the reprojection error rises rapidly with this type of noise, hence highlighting the need for a precise center estimation.

Results on real data. The existing sports circular markings detectors do not detect both inner and outer edges of the markings. Usually, they perform semantic segmentation, and neglect the field markings thickness. To detect both edges, we first use the semantic segmentation of the circle by using Theiner's line segmentation neural network [69], which determines a rough local area to search for inner and outer ellipses. To estimate those, we apply Canny's edge detector in the segmented area. We roughly estimate the center of the detected area, and starting from this center, we throw rays that should intersect the detected edges of both ellipses only a few pixels apart. The first detected edges along the rays belong to the candidates for the inner circle, and the second ones belong to the candidates for the outer circle. We use RANSAC to derive the final ellipses' equations in the image. Figure 5.9 shows the results of each intermediate step. Finally, examples of the center derived from the two ellipses are displayed in Fig. 5.10. The retrieved centers are too noisy to be used in real-world scenarios. The views, where only a part of the ellipse is visible, excluding the middle line for instance, remain very challenging views to calibrate or register. This is also true for the views where the center keypoint detector fails, for example when the center is occluded or almost invisible due to poor image quality.



(a) (b) (c)

Figure 5.9: Concentric circles detection. (a) displays the semantic segmentation of the "Center circle" class of the soccer field, (b) shows the result of Canny's detector in the segmented area, and (c) shows inliers points for inner (blue) and outer (red) ellipse points.



Figure 5.10: Circle centers (yellow) detected based on the concentric circles' detection. The results are poor, as synthetic experiments suggest.

5.5 Discussion

Our contribution to the quality of the annotations of keypoints on circles is obvious, as our method enables to annotate views for which no homography could be derived from points or lines alone. Furthermore, as shown in Fig. 5.3, it considerably improves the geometric consistency compared to great axis annotation. The benefits of our method for calibrating cameras are more difficult to measure, as it tackles mainly out-of-domain images that are by definition not in the annotated datasets. Still, our qualitative experiments indicate that our geometric method can retrieve points and lines correspondences that enable camera calibration in difficult scenarios by relying on the generalization capacities of detectors for actual recognizable patterns such as circular field markings or line intersections. The main limitations of our method come from the dependence on such robust detectors, which may overfit the few open-source datasets, and from the noise that may arise in ellipse estimation, mainly in the concentric setup where the line thickness is probably insufficient to guarantee a reliable circle center extraction. The cases depicting only a small part of the ellipse require further investigation to enable their calibration.

5.6 Conclusion

In this paper, we have shown how to better annotate central views in sports by using a geometric approach. Furthermore, while we show that learned keypoint detectors such as the one of PnLCalib [119] perform well to detect keypoints on circular patterns in in-domain scenarios, our geometric approach is effective when the test images are out-of-domain. However, some challenging views remain to be solved for single-frame camera calibration when the central circle center is out of the image or cannot be detected, which we will explore in future work.

Chapter 6

Sport-agnostic Metric Rectification of Sports Field (4^{th} article)

This chapter reproduces the following paper, currently under review: Floriane Magera et al. Beyond Field Registration: Sports-Agnostic Metric Rectification from a Single Image. *Submitted to a conference. 2026.*

Prologue

Existing works for camera calibration in sports largely exploit the presence of field markings to get point and line correspondences. Detecting the field markings is inherently sports-specific and requires the acquisition and annotation of sports footage, which might be scarce, difficult to publish because of broadcasting rights, and cumbersome to annotate.

In this final chapter, we investigate to what extent it is possible to recover most of the geometric information from a single image without relying on a dedicated sports field detector. By leveraging advanced neural networks, we seek to exploit generic geometric cues available in the scene.

Indeed, as previously mentioned in **Chapter 2**, geometry foundation models are showing great progress at estimating the geometry of a scene from a single frame. In this work, we studied how to best leverage those, despite their difficulties to generalize to large-scale scenes like sports venues. To support these strong detectors, we leverage a more mature, but still generic family of methods: line detectors, which provide strong projective constraints.

In the end, by combining geometry foundation models, line detection, and

projective constraints, we manage to recover the metric properties of the playing surface. This enables sports-agnostic metric rectification and the insertion of AR graphics on the playing surface.

This work was initiated during a research stay at the Computer Vision and Geometry lab of ETH Zürich from April to August 2026.

Beyond Field Registration: Sports-Agnostic Metric Rectification from a Single Image

Abstract

Augmented reality insertion in sports broadcasting relies on metric knowledge of the scene, yet current pipelines obtain it through field registration, which depends on a sport-specific field model and its detection in the image. Geometric foundation models promise a generic alternative, but broadcast footage, with its wide range of focal lengths and large-scale scenes, falls largely outside their training distribution: while monocular depth and surface normals transfer reasonably well, focal length, a global projective quantity with no local visual signature, is estimated poorly and inconsistently across domains. We propose a field-model-free approach to in-plane augmented reality insertion that combines the strengths of both worlds: plane normals are obtained from monocular geometry foundation models, while focal length is recovered geometrically from pairs of parallel lines detected with a generic line detector, without assuming a Manhattan world. As no single predictor proves reliable across all sports, we further introduce a lightweight and production-available prior in the form of sport- and camera-dependent field-of-view bounds to filter noisy focal estimates. On four datasets spanning football, basketball, and volleyball, our method is the most consistently competitive across domains, achieving the lowest reprojection error on every dataset, whereas each foundation-model baseline excels on some datasets and collapses on others.

6.1 Introduction

Sports are among the most active application domains of computer vision, driven by the economic and editorial stakes of live broadcasting. Vision systems now support a broad range of tasks, from tactical and performance analysis for storytelling, medical assistance, scouting, refereeing, and augmented-reality effects. Classical pipelines for sports applications often rely on three key components: (1) camera calibration, (2) player tracking, and (3) ball tracking.

Among these, camera calibration is the one that allows to bridge from the image world to the actual scene. It is the fundamental block for any spatially coherent reasoning. Camera calibration in standard sports environments is a well-studied problem with known field markings [7, 69, 71, 73, 90, 95, 121, 128], where the

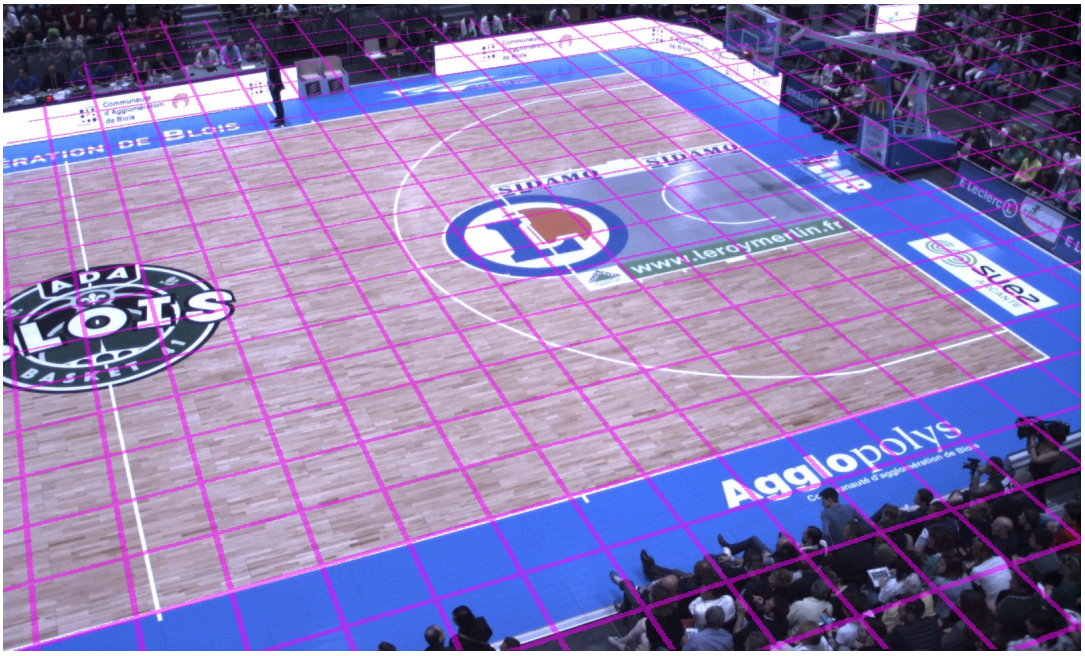


Figure 6.1: **Sports-agnostic metric rectification from a single image.** Given a single broadcast frame, our method segments the dominant planes, estimates their normals from monocular depth priors, and recovers the camera focal length from parallel lines, enabling metric rectification and augmented-reality insertion without any field model.

literature largely focuses on learning field layouts to register the playing surface. A central difficulty is that live broadcast footage is a scarce and expensive resource; open-source datasets are rare, and reliable ground truth is often available only as pseudo-ground truth.

With the rise of foundation models for single-image geometry, we argue that several of these applications can be addressed without any sport-specific field model or marking detection. This is the purpose of our work: we investigate the open problem of in-plane augmented-reality graphics insertion in sports scenes, without relying on any field template or sport-specific training data.

Given the capabilities of today's geometric foundation models [31], one might consider this problem essentially solved off-the-shelf. Sports content, however, is a highly specific domain: stadiums are rarely represented in standard training data, and broadcast cameras, with their wide focal ranges, differ strongly from the consumer or robotics common cameras these models are trained on. As a result, live sports footage often falls out of distribution, and the claims of the latest state of the art do not transfer to this setting. In fact, our experiments indicate that newer is far from better for sports applications: the latest models are not necessarily the most stable across sports datasets (Section 6.4). Our key insight is that foundation

models, despite their unreliability on out-of-distribution broadcast views, still carry information that is exploitable for a geometrically simple task such as in-plane graphics insertion, provided explicit geometric constraints anchor it. In particular, although the focal length has weak local visual evidence and is poorly regressed, it becomes strongly constrained by projective geometry once plane normals and parallel lines are available—quantities that foundation models predict reliably. Combining learned priors with this geometric evidence is therefore sufficient to recover metric rectifications where direct regression fails.

Our contributions are threefold: (1) We propose a **sports-agnostic pipeline for in-plane metric rectification** from a single image, which combines monocular depth foundation models with line and vanishing-point geometry, and requires no sport-specific field model. (2) We introduce a **geometric focal length recovery** from pairs of parallel lines lying on a plane of known normal, yielding a closed-form estimate per line pair, aggregated under a lightweight, production-available prior (camera- and sport-dependent field-of-view bounds). (3) We provide a **comprehensive evaluation across four sports datasets and three sports**, showing that no single monocular predictor generalizes across broadcast domains, while our method is the only one consistently competitive in focal, normal and reprojection accuracy.

6.2 Related Work

6.2.1 Sports camera calibration

The last few years have shown great improvement in the quality of broadcast camera calibration, from sports field registration techniques on small datasets to larger initiatives such as the SoccerNet challenges [8, 56, 57] and the FIFA challenge with the WorldPose dataset [129]. These efforts, combined with finer camera models [2, 128], have pushed camera calibration to high accuracy. However, these techniques rely on accurate and dense (virtual) field-marking detection [50, 74, 75, 90], making them dependent on the availability of training data to tackle new sports.

6.2.2 Single-image camera calibration

Beyond field markings, another line of work estimates camera intrinsic parameters directly from a single image. Recently, a new task has emerged: predicting perspective fields, *i.e.* estimating, densely and per pixel, the local up-vector and latitude relative to the horizon, which can then be converted into camera intrinsics. Perspective Fields [20] and GeoCalib [21] follow this paradigm, regressing calibration parameters

and the gravity vector directly from a single image. Also worth mentioning are the field-of-view fields later proposed by [22] with AnyCalib, which drop the constraint of visible gravity cues to model only the perspective and distortion effects of the camera. These methods enable focal length regression from a single image, but the quality of their predictions can degrade sharply outside the training distribution, as we quantify in Section 6.4.

6.2.3 Monocular depth estimation

MDE has progressed from relative, scale-ambiguous predictions to *zero-shot metric* depth, where models trained on large and diverse datasets predict absolute metric depth on unseen domains without per-scene tuning. Metric3D [24, 130] and Depth Anything [131] recover metric depth up to a known focal length, while Depth Pro [26], MoGe [25, 132] and FoundationGeo [133] additionally regress the focal length, yielding a full metric reconstruction from a single image. Several of these models also predict surface normals, either as an auxiliary output or through a geometry-aware formulation; jointly reasoning about depth and normals has been shown to improve geometric consistency and to yield more reliable local orientation than differentiating a noisy depth map. In our pipeline, we exploit precisely this property: monocular normals transfer well across domains, whereas the regressed focal length remains sensitive, which motivates recovering it geometrically rather than by direct regression.

6.2.4 Line and vanishing point detection

A complementary line of work extracts explicit geometric primitives. Line segment detection has evolved from handcrafted detectors such as LSD [134] to learning-based and hybrid approaches such as SOLD [135], HAWP [136], LETR [137], DeepLSD [138], which combine the robustness of deep networks with sub-pixel geometric precision. From these segments, vanishing points, and hence constraints on the camera intrinsics, are recovered by robust multi-model estimators such as J-Linkage [139], Progressive-X [140], or convex relaxation [141], and can be further constrained by a known gravity direction [142]. These primitives are sports-agnostic and geometrically grounded, but classical pipelines built on them typically assume a Manhattan-world structure and remain sensitive to detection noise on cluttered broadcast imagery.

6.2.5 Rectification

Recovering the metric structure of a plane from a single image is a long-standing problem in geometric vision. Classical approaches rectify a plane up to a similarity transform by exploiting known geometric cues, such as vanishing lines, parallelism, orthogonality, or known length ratios, through the projective and affine properties of the imaged plane [11, 143, 144]. These methods are accurate when reliable primitives are available, but they assume that such cues can be extracted precisely, which rarely holds in cluttered, real-world imagery.

In the sports domain, metric rectification is typically obtained indirectly through field registration, which aligns the image with a known field model [71, 74, 75]. Because the real-world field dimensions are known, this not only yields metric geometry but also resolves the absolute scale ambiguity inherent to single-view rectification. Its price, however, is a strong sport-specific prior: a dedicated field model and marking detector are required for each sport. In contrast, our method recovers metric rectification up to an unknown global scale without any field model by combining monocular normals with line-based focal recovery. This makes it applicable across sports and playing surfaces, at the cost of the absolute scale that a known field would otherwise provide, which is often not necessary for AR insertions.

6.3 Method

6.3.1 Problem statement

From a single image, we aim to recover the homographies that metrically rectify the dominant planes of the scene, up to a similarity transform preserving angles and length ratios. Under the assumption of a simplified pinhole camera model with square pixels and a centered principal point, our method segments the planar regions of the image, estimates their normals, and recovers the camera focal length.

The rectifying homography of the plane of normal $\mathbf{n}$ depends solely on $\mathbf{n}$ and the focal length f:

$$\mathtt{H} = \mathtt{KRK}^{-1} = \mathrm{diag}(f, f, 1)\mathtt{R}_{\mathbf{n}}\mathrm{diag}\left(\frac{1}{f}, \frac{1}{f}, 1\right), \tag{6.1}$$

where $\mathtt{R}_{\mathbf{n}}$ is the rotation of the minimal angle that aligns $\mathbf{n}$ with the optical axis. Recovering each plane reduces to estimating two quantities: the normal $\mathbf{n}$ and the focal length f, which is precisely what our method provides. In the context of this work, we assume that lens distortion is negligible.

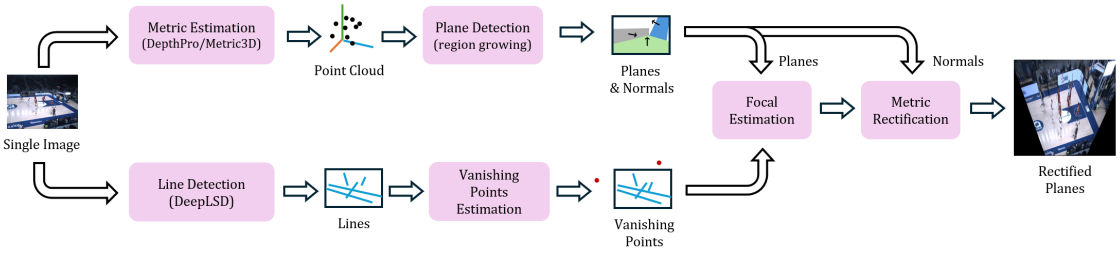


Figure 6.2: Overview of the proposed sport-agnostic pipeline. From a single image, two branches are computed: a depth foundation model yields a metric point cloud from which dominant planes are detected, and a parallel line detection. Planes, lines, and vanishing points jointly feed the focal length estimation, enabling metric rectification of each plane.

6.3.2 Pipeline overview

Figure 6.2 illustrates our complete pipeline. From an input image I, we run two branches: (1) a monocular depth estimation branch, which yields plane segmentations and their normals, and (2) a parallel line detection branch whose outputs jointly feed our focal length estimation. The recovered focal length, the plane normal and its segmentation are finally used to metrically rectify the plane.

6.3.3 Plane segmentation from depth models

We detect planar regions and their normals using monocular metric depth foundation models. Existing methods either jointly estimate depth and focal length or assume a known focal length. For the latter type, we set the field of view of the camera to a plausible value.

The resulting focal length is then used to lift the depth map into a 3D point cloud. For each point, we form a local neighborhood from the points that were lifted from its adjacent pixels, and estimate an initial normal as the eigenvector of the local covariance matrix associated with the smallest eigenvalue (Principal Component Analysis (PCA)).

To favor large, spatially coherent planar regions, we apply a region-growing algorithm to these oriented points. Starting from arbitrary seed points whose normal and position define an initial plane hypothesis, the candidate plane (*i.e.* the region) grows by considering the neighbors of the points already assigned to the region. A neighboring point is added to the current region if it lies within a distance of τ_d of the current plane hypothesis and if the angle formed by its normal and the plane normal is smaller than τ_θ. The region-plane equation is updated every few hundred iterations via PCA. This yields a set of planar regions and their corresponding normals $\{\mathbf{n}_i\}$.

6.3.4 Parallel lines detection

The second type of sport-agnostic detector we use is line detection. We detect line segments in the image with DeepLSD [138], a hybrid detector that leverages the robustness of deep neural networks to produce enhanced image gradients, enabling us to detect low-contrast and partially occluded lines, which are common in broadcast footage. This surrogate image gradient is fed to a classical line detector, LSD [134], keeping the sub-pixel precision of handcrafted methods. This algorithm enables the recovery of mowing patterns and field markings, and yields a set of segments $L = \{\mathbf{l}_i\}^N$.

We then estimate vanishing points from L with Progressive-X [140], a robust multi-model estimator. Each Vanishing Point (VP) groups the segments whose supporting direction is consistent with it, and thus identifies the lines that are parallel in 3D.

A segment is assigned to a candidate VP v through a consistency criterion. Rather than a point-to-line distance, the standard metric [145] measures the deviation between the endpoints of l and the line joining the midpoint of l to v. This criterion is length-dependent: for a fixed orientation, a short segment may be accepted while its longer counterpart is rejected, biasing VP estimation towards short lines. This is undesirable in sports scenes, where long markings (typically touchlines and sidelines) are better localized and more robust to detection noise, hence more informative.

To remove this bias, we normalize all segments to a common length before the inlier decision, so that acceptance depends on orientation alone. Once the inlier set of a VP is fixed, segment length becomes a legitimate confidence cue: we therefore weight each segment's contribution to the VP position by its original length, as in standard VP estimation [140].

6.3.5 Focal length from parallel lines

From each pair of parallel detected lines ($\mathbf{l}_1 = (a_1, b_1, c_1)^\top, \mathbf{l}_2 = (a_2, b_2, c_2)^\top$) inscribed in a plane, we can derive an equation for f. The rectified version of the lines $\mathbf{l}'_i = \mathtt{H}^{-\top}\mathbf{l}_i$ are parallel, meaning their first two coefficients are multiples of each other, or if we rotate by 90 degrees one of their normal, their dot product equals zero: $\mathbf{l}'_1[1,2]^\top \mathtt{R}_{90^\circ}\mathbf{l}'_2[1,2] = 0$. The rectifying homography $\mathtt{H}$ is of the form:

$$\mathtt{H}^{-\top} = \mathtt{K}^{-1}\mathtt{R}_{\mathbf{n}}\mathtt{K} = \begin{bmatrix} r_1 & r_2 & r_3/f \\ r_4 & r_5 & r_6/f \\ r_7 f & r_8 f & r_9 \end{bmatrix} \tag{6.2}$$

and thus the rectified lines can be derived as:

$$\mathbf{l}'_i = \begin{pmatrix} r_1 a_i + r_2 b_i + \dfrac{r_3 c_i}{f} \\ r_4 a_i + r_5 b_i + \dfrac{r_6 c_i}{f} \\ r_7 f a_i + r_8 f b_i + r_9 c_i \end{pmatrix}. \tag{6.3}$$

This gives the final equation to solve for f:

$$\begin{aligned} &(r_4 a_1 + r_5 b_1 + \frac{r_6 c_1}{f})(r_1 a_2 + r_2 b_2 + \frac{r_3 c_2}{f}) \\ -&(r_1 a_1 + r_2 b_1 + \frac{r_3 c_1}{f})(r_4 a_2 + r_5 b_2 + \frac{r_6 c_2}{f}) = 0. \end{aligned} \tag{6.4}$$

Although Eq. (6.4) appears quadratic in $1/f$, the second-order terms cancel identically $r_3 r_6 c_1 c_2 - r_3 r_6 c_1 c_2 = 0$, leaving a linear equation in $1/f$. The focal length therefore admits the unique closed-form solution:

$$\begin{aligned} \text{Num} = \; &- a_1 c_2 r_1 r_6 + a_1 c_2 r_3 r_4 + a_2 c_1 r_1 r_6 - a_2 c_1 r_3 r_4 \\ &- b_1 c_2 r_2 r_6 + b_1 c_2 r_3 r_5 + b_2 c_1 r_2 r_6 - b_2 c_1 r_3 r_5, \end{aligned} \tag{6.5}$$

$$f = \frac{\text{Num}}{(a_1 b_2 - a_2 b_1)(r_1 r_5 - r_2 r_4)}. \tag{6.6}$$

The estimate is valid only when $(a_1 b_2 - a_2 b_1)$ and $(r_1 r_5 - r_2 r_4)$ are bounded away from zero; pairs of near-parallel image lines are discarded, and the sign ambiguity of $\mathbf{n}$ is resolved by enforcing $f > 0$.

6.3.6 Focal length aggregation

Each pair of parallel lines contained in the same plane yields a focal length estimate, such that a few hundred different values $\{\hat{f}_i\}$ are computed per single frame. These values are corrupted by errors in the detected lines and estimated normals, and their distribution exhibits strong outliers. Rather than seeking a single estimator that would be robust across all domains, which we could not find in our experiments, we leverage a weak, readily available prior: both the targeted camera type and targeted sport constrain the plausible field of view, which can be turned into a plausible focal length interval $[f_{min}, f_{max}]$. We discard the estimates falling outside this interval and take the mean. In our experiments, the interval filtering removed the dominant outliers, after which the arithmetic mean provided stable estimates.

	WorldPose[129]				WorldCup[71]				DeepSportRadar[54]				MultiSports[128]			
Method	Mean	Med.	δ_1	δ_2	Mean	Med.	δ_1	δ_2	Mean	Med.	δ_1	δ_2	Mean	Med.	δ_1	δ_2
GeoVP [142]	2.00	0.51	21	48	1.74	0.44	26.6	57.7	0.50	0.14	70.3	89.4	1.49	1.05	10	40
GeoCalib [21]	0.43	0.45	12	64	0.47	0.49	10.1	51.1	0.15	0.08	70.8	100.0	0.42	0.34	30	60
AnyCalib [22]	0.35	0.36	13	95	0.31	0.31	29.4	95.2	0.12	0.10	98.3	100.0	0.38	0.36	20	70
FoundationGeo [133]	0.49	0.49	0	53	0.46	0.47	2.8	61.8	0.20	0.15	57.6	100.0	0.37	0.36	20	70
DepthPro [26]	**0.11**	**0.09**	**92**	**99**	0.26	0.25	49.9	92.9	**0.09**	0.08	**99.2**	100.0	0.38	0.43	20	**90**
MoGe-2 [25]	0.49	0.49	0	52	0.40	0.40	5.1	83.5	0.22	0.19	54.2	100.0	0.37	0.35	20	80
Ours	0.20	0.18	68	97	**0.24**	**0.23**	**58**	**98**	0.10	**0.07**	94	100.0	**0.22**	**0.25**	**50**	**90**

Table 6.1: Focal length estimation across datasets. Relative error is lower-is-better (↓); δ_1 and δ_2 denote the percentage of images whose relative focal error is below 25% and 50% respectively (↑). Best per column in **bold**, second best underlined.

6.4 Experiments

6.4.1 Experimental setup

Datasets

We validate our approach on four sports datasets spanning 3 sports (football, basketball, and volleyball) in both indoor and outdoor setups. The **WorldPose** dataset [129] was released by FIFA, containing football main camera footage with pseudo-ground truth camera parameters provided by Vizrt, a company specializing in AR effects [146]. We sampled the videos of the dataset at 1 fps to obtain a final set of 2,958 HD frames along with their camera parameters. The second dataset is the WorldCup2014 dataset [71]. It contains broadcast main-camera footage of the 2014 World Cup, with 395 720p frames annotated with homographies. The third dataset is the DeepSportRadar dataset [54] which contains fixed cameras around 15 basketball stadiums. It is annotated with camera parameters. Following [128], we set the resolution to 1624×1234 pixels, and keep the frames satisfying our assumption of square pixels, resulting in a set of 236 images. The last dataset is the Multisports-Calib dataset [128] which was annotated with field markings that we converted in homographies. Besides volleyball, it contains ice hockey and handball, which do not contain enough line markings to compute homographies. We only keep the volleyball frames, which make a set of 10 images.

For the datasets annotated with homographies (WorldCup and Multisports), we convert the registration homographies into pinhole camera parameters by exploiting the fact that the sports field is the ground plane $Z = 0$. Under this assumption, $\mathtt{H} = \mathtt{K}[\mathbf{r}_1\mathbf{r}_2\mathbf{t}]$ so that the homography encapsulates the first two columns $\mathbf{r}_1$ and $\mathbf{r}_2$ of the rotation matrix and the translation vector up to scale. Enforcing the

orthonormality of $\mathbf{r}_1$ and $\mathbf{r}_2$, we can recover the rotation matrix $\mathtt{R}$ and the intrinsic matrix $\mathtt{K}$ with Cholesky decomposition, providing a common parametrization across all datasets.

Metrics

We evaluate our pipeline at different levels: plane normal estimation, focal length estimation, and the final rectifying homography.

Normal. In sports broadcasting, the dominant plane is often the sports field plane, which is by convention the ground plane $Z = 0$. Hence the normal of the field plane is obtained from the last column $\mathbf{r}_3$ of the rotation matrix. This evaluation is partial as other planes are not evaluated; we argue, however, that this is the most relevant one for AR graphics in sports images. We report the mean and median angular error of the estimated normal.

Focal length. We report mean and median relative errors, as well as δ_1 and δ_2 rates defined as the percentage of estimated focal lengths whose relative error $|\hat{f} - f_{GT}|/f_{GT}$ is below 0.25 and 0.5 respectively.

Rectifying homography. We measure the reprojection error over a uniform grid of pixels. For a sampled pixel $\mathbf{p}$ in homogenous coordinates $\mathbf{p} = (x, y, 1)^\top$, the error is defined by $E = \|\pi(\mathtt{H}\mathtt{H}_{GT}^{-1}\mathbf{p}) - \pi(\mathbf{p})\|_2$ where $\|\cdot\|_2$ is the Euclidean distance in the image plane, and $\pi((x, y, w)^\top) = (x/w, y/w)^\top$ is the projection function. The reprojection error for an image I is the mean of E over the sampled pixels, and we report mean and median values over all images.

Hyperparameters

Line detection. We run DeepLSD [138] on each image downsized by several factors $[2, 3, 4, 5]$ to ensure a high number of detected lines and further tune the sigma scale of LSD [134] to 0.9. **Normal estimation.** Depth estimators that assume the focal length to be known, like Depth Anything[131] and Metric3Dv2 [130], require a candidate focal length to lift their depth map predictions into point clouds. For the different datasets, we use a fixed vertical field-of-view value of 30 degrees. Concerning our region-growing algorithm, we set τ_d to 80 cm and τ_θ to 37 degrees for all datasets and depth estimators. As our experiments suggest in Table 6.2, we use the plane normals obtained with Metric3Dv2 [130] for football and MoGe-2 [25] for other sports. **Focal length filtering.** The focal length bounds $[f_{min}, f_{max}]$ depend on the sport and camera type. Broadcast cameras typically have a wide range of fields of view. Based on empirical analysis of sports datasets, we set the vertical field of view to

[6, 28] for football main-camera footage, [13, 60] degrees for basketball, and finally [10, 28] degrees for volleyball.

6.4.2 Baselines

We evaluate our pipeline against three families of baselines, chosen to cover the main paradigms for single-image calibration.

Direct calibration. GeoCalib [21] and AnyCalib [22] both regress the camera intrinsics directly. Moreover, GeoCalib predicts the gravity vector, making it the only method that can be compared against our full pipeline, *i.e.* focal length, and ground plane normal. We use the pinhole camera models of both methods, neglecting lens distortion in the context of this work.

Depth and normal foundation models. Metric3Dv2 [130] and Depth Anything [131] provide metric depth from which we derive ground-plane normals, but assume a known focal length. MoGe-2 [25], DepthPro [26] and FoundationGeo [133] additionally regress the focal length, and thus represent the direct focal regression paradigm—the very component our geometric approach aims to replace.

Geometric estimation. We also integrate a classical computer vision approach to estimate the focal length from orthogonal vanishing points. We use the implementation of [142]; it is the closest counterpart to our method and isolates the contribution of combining lines with monocular normals. For conciseness, we denote it "GeoVP" in our result tables.

For all baselines, we use the official checkpoints and the hyperparameters reported by their authors, without any re-tuning, to ensure a fair comparison.

6.4.3 Main results

Table 6.1 provides the performance of deep foundation models against our pipeline for focal length estimation. All baselines rank well on the DeepSportRadar dataset, likely because it exhibits a wider field-of-view range, which is more representative of the distributions commonly encountered in generic computer vision datasets. DepthPro [26] is excellent at predicting focal lengths, and as shown in Fig. 6.3, the predicted values are closer to the ground-truth distribution of field of views present in the sports dataset. Interestingly, the performance of DepthPro is uneven across football datasets while they contain the same type of content: main camera footage. Traditional computer vision approach GeoVP [142] is limited by its reliance on the detection of pairs of orthogonal vanishing points to estimate a focal length. Indeed, line detection often provides enough support to reliably estimate only one of the

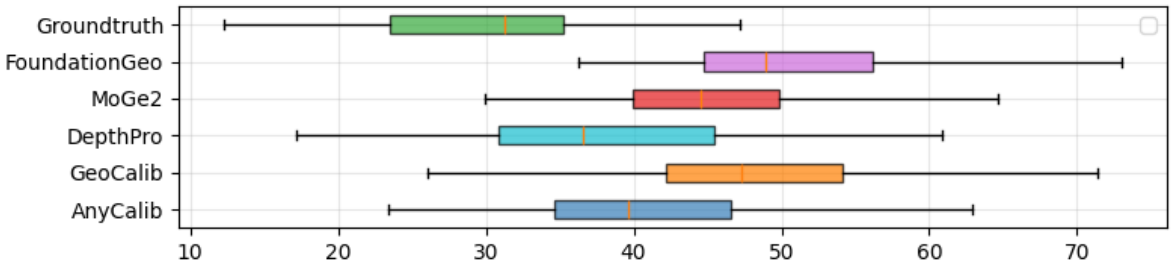


Figure 6.3: Vertical Field of View (°) predictions of different baselines against the field-of-view values present in the sports datasets. Most methods predict vertical fields of view that are significantly wider than the range typically observed for broadcast cameras. This highlights the importance of domain-specific camera priors. Among the evaluated methods, DepthPro exhibits the closest distribution to the ground truth.

two vanishing points associated with rectangular sports field markings. Our method ranks among the top one or two methods, except for δ_1 metric on DeepSportRadar.

Table 6.2 demonstrates the precision of the normal prediction when combining our region-growing algorithm with depth estimators (except GeoCalib [21] which predicts the normal vector directly). Metric3Dv2 [130] is a strong baseline, despite us using a default value for lifting its depth map. GeoCalib [21] ranks second on three of the datasets.

Finally, Table 6.3 shows the reprojection error of the obtained rectifying homographies. Our method consistently obtains the best rectification, even though its focal length prediction is not systematically better than DepthPro's one, showing that accurate rectification depends on the joint consistency of focal length and plane normal rather than on focal length accuracy alone.

These results support our central claim: since no generic predictor is reliable across domains, a lightweight, production-available prior—such as the sport- and camera-dependent field-of-view bounds—is what turns noisy predictions into usable metric rectifications, which is further supported by our qualitative analysis.

6.4.4 Qualitative results

Figure 6.4 compares the metric rectifications obtained with DepthPro [26], GeoCalib [21], MoGe-2 [25] and our method on representative frames from each dataset. For every image, we rectify the ground plane to a fronto-parallel view. A correct rectification maps the field markings onto straight, mutually orthogonal lines with preserved length ratios.

The baselines reflect the quantitative trends of Section 6.4.4. GeoCalib and MoGe-

	WorldPose [129]		WorldCup [71]		DeepSportRadar [54]		MultiSports [128]	
Method	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Ours+DAv3[131]	3.19	3.02	2.53	2.35	11.11	11.14	**6.36**	<u>4.39</u>
Ours+Metric3Dv2[130]	**1.67**	**1.47**	**1.86**	**1.64**	8.81	9.26	14.87	**2.56**
Ours+MoGe-2[25]	5.68	5.65	4.00	3.79	**1.50**	**1.37**	12.19	5.29
Ours+DepthPro[26]	3.68	3.52	<u>2.50</u>	<u>2.32</u>	4.63	4.85	20.77	25.39
GeoCalib[21]	<u>3.17</u>	<u>2.54</u>	3.91	2.88	<u>4.31</u>	<u>3.85</u>	<u>9.79</u>	7.35

Table 6.2: Ground plane normal estimation using our region-growing algorithm on top of depth estimators, angular error in degrees (↓). GeoCalib predicts directly the ground plane normal with its gravity vector. Best per column in **bold**, second best <u>underlined</u>.

	WorldPose [129]		WorldCup [71]		DeepSportRadar [54]		MultiSports [128]	
Method	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GeoCalib[21]	1196	1025	1041	835	210	<u>160</u>	<u>922</u>	<u>857</u>
DepthPro[26]	<u>272</u>	<u>248</u>	<u>385</u>	<u>224</u>	<u>202</u>	206	1849	1911
MoGe-2[25]	1679	1528	793	683	350	254	1150	1171
Ours	**260**	**231**	**272**	**203**	**136**	**85**	**618**	**624**

Table 6.3: Reprojection error (in pixels, ↓) across datasets. For each dataset we report the mean and median error. Best result per column in **bold**, second best <u>underlined</u>.

2 often recover a plausible orientation but an inaccurate focal length, which manifests as residual perspective distortion: parallel markings remain slightly convergent and right angles are not preserved. Despite its excellent performance in focal length estimation, DepthPro's rectifications do not particularly stand out from the rest of the baselines. In contrast, our method yields rectifications whose markings stay straight, parallel, and orthogonal across all datasets.

These qualitative results confirm that accurate metric rectification depends on the joint consistency of focal length and plane normal rather than on focal accuracy alone.

6.5 Discussion

The strength of our approach lies in a division of labor: it builds on increasingly accurate foundation models for single-image geometry (depth and surface normals) while delegating the focal length, a physical quantity that remains difficult to infer

reliably, to explicit geometric reasoning. By exploiting pairs of parallel lines, we recover the focal length robustly even in non-Manhattan scenes, where no assumption of three orthogonal directions can be made.

This design, however, comes with clear limitations. Our focal estimation is entirely line-dependent: when no lines are detected, no constraint on f can be formed and the pipeline provides no estimate. The method further relies on a weak prior on the plausible field of view of the camera to filter out noisy focal candidates. These requirements are mild in broadcast sports, where field markings are ubiquitous and camera setups are known, but they define a limit to the applicability of our approach.

Another limitation is the dependency on the underlying geometric foundation model. Our experiments show that no single depth estimator consistently dominates across sports datasets, suggesting that robustness to domain shifts remains an open challenge. Since our focal recovery relies on the estimated plane normals, errors in the underlying geometric predictions propagate directly to the calibration.

6.6 Conclusion

We presented a sports-agnostic approach to metric rectification for in-plane augmented-reality insertion from a single image without relying on field registration or sport-specific models. Our study highlights both the promise and the current limitations of geometric foundation models for camera calibration. While monocular depth and surface-normal estimation have reached a level of maturity that enables reliable geometric reasoning across domains, focal length estimation remains considerably less stable. According to our results, a hybrid formulation in which learned scene geometry and explicit projective constraints complement each other, remains advantageous. Nevertheless, this conclusion should be viewed in the context of the current generation of foundation models rather than as a fundamental limitation of learning-based approaches. As larger and more diverse geometric foundation models emerge, direct regression of camera intrinsics may eventually become sufficiently accurate and robust to reduce or eliminate the need for explicit geometric recovery. In parallel, extending our framework to jointly estimate focal length and lens distortion remains an important direction for future work.

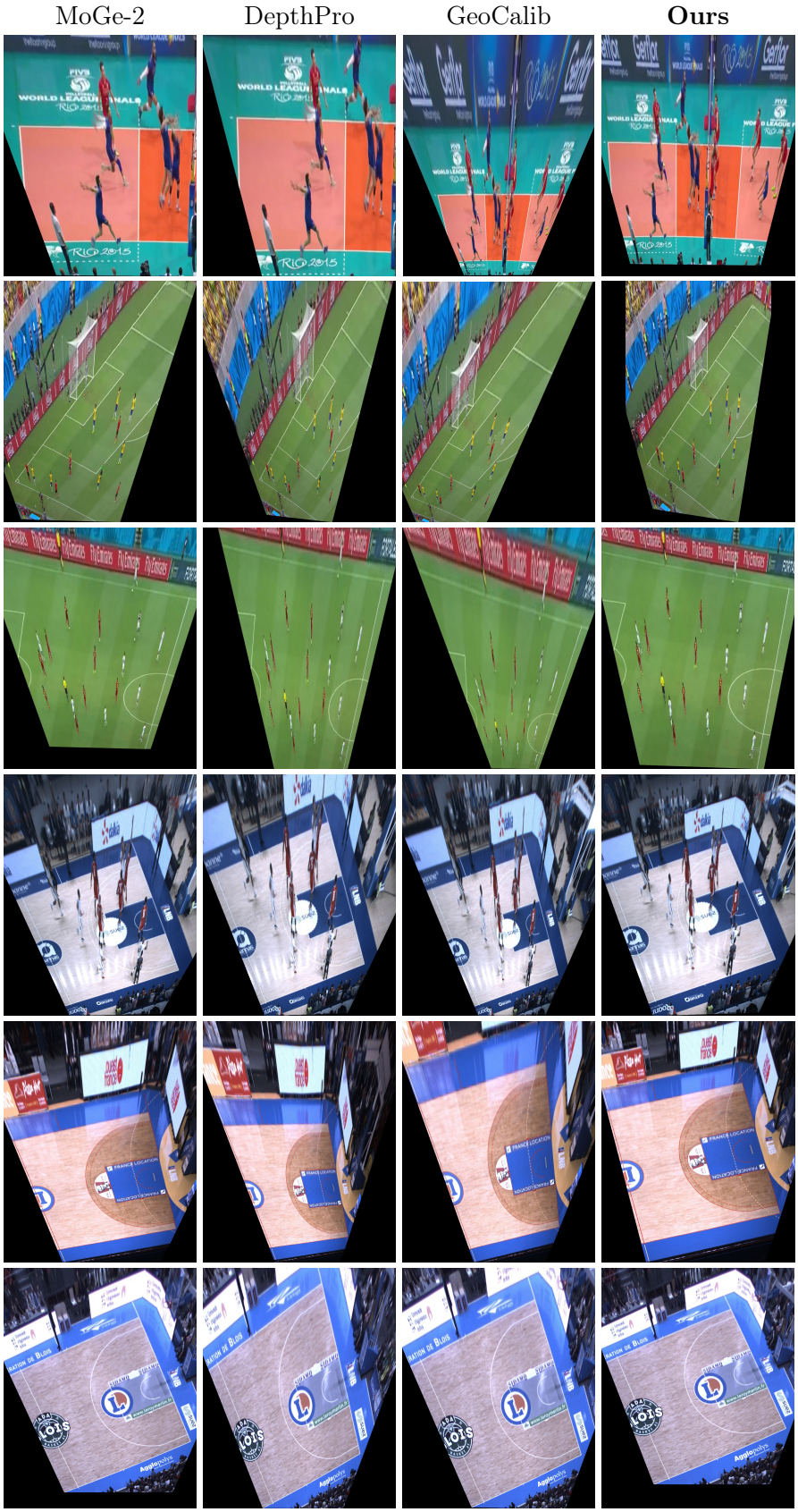


Figure 6.4: Qualitative comparison of metric plane rectification across methods. Each row shows, from left to right, the rectified dominant plane obtained with MoGe-2, DepthPro, GeoCalib, and our method. Rows span football, basketball, and volleyball scenes. Our method produces the most geometrically consistent rectifications.

Chapter 7

Conclusion

7.1 Contributions

In this thesis, we addressed the problem of camera calibration, as it is a key enabler for multiple downstream applications. Our work contributes to the field in multiple ways: how to properly evaluate it, how to make it consistent with the physical world, and finally how to use geometry to avoid sports-specific detectors.

By working on camera calibration benchmarking, we contributed to establishing camera calibration as a scientific problem in its own right, rather than an intermediary step towards other computer vision tasks. Our work emphasized the need to have rigorous evaluation protocols in the absence of ground truth and enabled a new range of methods through our involvement in the organization of challenges. We argue that rigorous assessment requires comparability. By disentangling the evaluation procedure from the camera model, we enabled research on the camera model. With our protocol, we establish a common benchmarking framework that enables fair comparisons across methods.

With BroadTrack, we contributed to the state of the art of camera calibration in sports in terms of quality. More importantly, in this work, we showed the importance of physically plausible camera parameters. Indeed, by enforcing physical properties at the core of our camera model, we achieved consistent camera calibrations rather than just visually satisfying reprojections.

We then showed how to exploit the rich geometric information contained in circular field markings, by deriving additional point and line correspondences through projective geometry. By integrating these new correspondences in existing pipelines, we showed that geometry-based and learning-based methods can be effectively combined to better deal with challenging cases.

Finally, the emergence of geometry foundation models offered a new avenue for

sport-agnostic camera calibration and sports field registration. By combining these foundation models with generic line detection, we again combined strong geometric cues with learning-based approaches to make the most of a single image. This work opened up the possibility of AR graphics insertion for any sport with straight line markings on a planar playing surface.

7.2 Future Work

Validation on additional sports. The experiments of this thesis were mainly conducted on football. Football data is by far the most widely available due to the popularity of the sport, and moreover, it is the most interesting one from a business perspective. Our contributions likely transpose to many different sports, but we will need suitable datasets to establish that.

Temporal benchmarking. Our benchmarking protocol does not account for the temporal consistency of the camera parameters estimated by the evaluated method. One could design a benchmarking protocol that takes into account the temporal consistency of the physical motion of the camera, including some of its internal moving parts such as its lenses.

Exploration of more sophisticated camera models. Finally, our camera model remains a simplification of real-world complex imaging systems, and we believe there is still room for improvement in the disambiguation between camera tripod position, camera focal point position, and radial distortion.

7.3 Perspectives

Between the beginning and the end of this thesis, computer vision evolved at an astounding pace, and any method developed today may become obsolete within a few months. What will remain, however, is our reflection on how we evaluate methods and which physical quantities must be estimated correctly to enable demanding downstream applications.

While increasingly general, highly-performing, and resource-intensive algorithms will likely dominate future benchmarks, we believe that grounding methods in geometric and physical principles remains essential to achieving the level of reliability required by broadcast applications.

Looking forward, we expect geometric foundation models to progressively improve on broadcast scenes that remain largely out-of-distribution today. However, from our industrial experience, there will always be a misalignment between end-to-end

methods and the need for explainability, accuracy, and robustness required for a real product. As these models continue to mature, the focus of the field may shift from geometry estimation itself towards ensuring the quality, robustness and scalability of these large models. Ensuring quality and robustness will require to explore geometry validation, failure detection, and uncertainty quantification, while scalability raises questions regarding their efficient deployment under limited computational resources.

Bibliography

[1] Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. "A Universal Protocol to Benchmark Camera Calibration for Sports". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, Washington, USA: IEEE, June 2024, pp. 3335–3346 (cit. on pp. 3, 4, 15, 38, 39, 46, 68).

[2] Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. "BroadTrack: Broadcast Camera Tracking for Soccer". In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Tucson, Arizona, USA: IEEE, Feb. 2025, pp. 6177–6187 (cit. on pp. 3, 4, 33, 76).

[3] Floriane Magera, Thomas Hoyoux, Martin Castin, Olivier Barnich, Anthony Cioppa, and Marc Van Droogenbroeck. "Can Geometry Save Central Views for Sports Field Registration?" In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, Tennessee, USA: IEEE, June 2025, pp. 6060–6069 (cit. on pp. 3, 53).

[4] Floriane Magera, Olivier Barnich, Marc Van Droogenbroeck, and Daniel Barath. *Beyond Field Registration: Sports-Agnostic Metric Rectification from a Single Image*. Submitted to a conference. 2026 (cit. on pp. 3, 72).

[5] 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 International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), CVsports*. Nashville, Tennessee, USA, June 2021, pp. 4532–4541 (cit. on pp. 3, 4, 21, 22).

[6] EVS Broadcast Equipment. *Multi-camera review system - Xeebra*. https://evs.com/products/video-assistance/xeebra. June 2022 (cit. on pp. 4, 29, 30, 49).

[7] Long Sha, Jennifier Hobbs, Panna Felsen, Xinyu Wei, Patrick Lucey, and Sujoy Ganguly. "End-to-End Camera Calibration for Broadcast Videos". In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, Washington, USA: IEEE, June 2020, pp. 13627–13636 (cit. on pp. 4, 12, 20–23, 37, 38, 56, 74).

[8] Vladimir Somers, Victor Joos, Anthony Cioppa, Silvio Giancola, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir M. Mansourian, Xin Zhou, Shohreh Kasaei, Bernard Ghanem, Alexandre Alahi, Marc Van Droogenbroeck, and Christophe De Vleeschouwer. "SoccerNet Game State Reconstruction: End-to-End Athlete Tracking and Identification on a Minimap". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, Washington, USA: IEEE, June 2024, pp. 3293–3305 (cit. on pp. 4, 19, 33, 36, 54, 56, 76).

[9] Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdulrahman Darwish, Adrien Maglo, Albert Clapés, Andreas Luyts, Andrei Boiarov, Artur Xarles, Astrid Orcesi, Avijit Shah, Baoyu Fan, Bharath Comandur, Chen Chen, Chen Zhang, Chen Zhao, Chengzhi Lin, Cheuk-Yiu Chan, Chun Chuen Hui, Dengjie Li, Fan Yang, Fan Liang, Fang Da, Feng Yan, Fufu Yu, Guanshuo Wang, H. Anthony Chan, He Zhu, Hongwei Kan, Jiaming Chu, Jianming Hu, Jianyang Gu, Jin Chen, João V. B. Soares, Jonas Theiner, Jorge De Corte, José Henrique Brito, Jun Zhang, Junjie Li, Junwei Liang, Leqi Shen, Lin Ma, Lingchi Chen, Miguel Santos Marques, Mike Azatov, Nikita Kasatkin, Ning Wang, Qiong Jia, Quoc Cuong Pham, Ralph Ewerth, Ran Song, Rengang Li, Rikke Gade, Ruben Debien, Runze Zhang, Sangrok Lee, Sergio Escalera, Shan Jiang, Shigeyuki Odashima, Shimin Chen, Shoichi Masui, Shouhong Ding, Sin-wai Chan, Siyu Chen, Tallal El-Shabrawy, Tao He, Thomas B. Moeslund, Wan-Chi Siu, Wei Zhang, Wei Li, Xiangwei Wang, Xiao Tan, Xiaochuan Li, Xiaolin Wei, Xiaoqing Ye, Xing Liu, Xinying Wang, Yandong Guo, Yaqian Zhao, Yi Yu, Yingying Li, Yue He, Yujie Zhong, Zhenhua Guo, and Zhiheng Li. "SoccerNet 2022 Challenges Results". In: *International ACM Workshop on Multimedia Content Analysis in Sports (MMSports)*. Lisbon, Portugal: ACM, Oct. 2022, pp. 75–86 (cit. on pp. 4, 21, 46).

[10] Anthony Cioppa et al. "SoccerNet 2023 challenges results". In: *Sports Engineering* 27.2 (July 2024), pp. 1–18 (cit. on pp. 4, 21, 46).

[11] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. second. Cambridge, UK: Cambridge University Press, 2004 (cit. on pp. 6–8, 39, 40, 60, 78).

[12] Dean Brown. "Decentering distortion of lenses". In: *Photogrammetric Engineering and Remote Sensing* 32.3 (1966), pp. 444–462 (cit. on pp. 9, 40).

[13] Zuzana Kukelova, Jan Heller, and Andrew Fitzgibbon. "Efficient Intersection of Three Quadrics and Applications in Computer Vision". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, Nevada, USA: Institute of Electrical and Electronics Engineers (IEEE), June 2016, pp. 1799–1808 (cit. on pp. 10, 55).

[14] Martin A. Fischler and Robert C. Bolles. *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*. Tech. rep. 213. Menlo Park, CA: SRI International, Mar. 1980 (cit. on p. 11).

[15] Jianhui Chen and James J. Little. "Sports camera calibration via synthetic data". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Long Beach, California, USA: IEEE, June 2019, pp. 2497–2504 (cit. on pp. 12, 20–22, 38, 56).

[16] Alex Kendall, Matthew Grimes, and Roberto Cipolla. "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization". In: *IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, Dec. 2015, pp. 2938–2946 (cit. on pp. 12, 38).

[17] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. "Understanding the Limitations of CNN-Based Absolute Camera Pose Regression". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, California, USA: IEEE, June 2019, pp. 3297–3307 (cit. on pp. 12, 57).

[18] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. "Deep Image Homography Estimation". In: *arXiv* abs/1606.03798 (2016). arXiv: 1606.03798 (cit. on pp. 12, 38, 57).

[19] Thomas Schöps, Viktor Larsson, Marc Pollefeys, and Torsten Sattler. "Why Having 10,000 Parameters in Your Camera Model Is Better Than Twelve". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, Washington, USA: Institute of Electrical and Electronics Engineers (IEEE), June 2020, pp. 2532–2541 (cit. on p. 13).

[20] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F. Fouhey. "Perspective Fields for Single Image Camera Calibration". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, Canada: IEEE, June 2023, pp. 17307–17316 (cit. on pp. 13, 76).

[21] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. "GeoCalib: Learning Single-Image Calibration with Geometric Optimization". In: *European Conference on Computer Vision (ECCV)*. Vol. 15098. Lecture Notes in Computer Science. Springer Nature Switzerland, Nov. 2024, pp. 1–20 (cit. on pp. 13, 76, 82, 84–86).

[22] Javier Tirado-Garín and Javier Civera. "AnyCalib: On-Manifold Learning for Model-Agnostic Single-View Camera Calibration". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. Honolulu, Hawaii, USA: IEEE, Oct. 2025, pp. 8044–8055 (cit. on pp. 13, 77, 82, 84).

[23] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. "ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth". In: *arXiv* abs/2302.12288 (2023). arXiv: 2302.12288 (cit. on p. 13).

[24] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. "Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 9009–9019 (cit. on pp. 13, 77).

[25] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. "MoGe-2: Accurate Monocular Geometry with Metric Scale and Sharp Details". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 39. San Diego, California, USA: Curran Associates, Inc., Dec. 2025, pp. 1–32 (cit. on pp. 13, 77, 82–86).

[26] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. "Depth Pro: Sharp Monocular Metric Depth in Less Than a Second". In: *International Conference on Learning Representations (ICLR)*. Singapore, Apr. 2025, pp. 1–36 (cit. on pp. 13, 77, 82, 84–86).

[27] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. "Accelerated Coordinate Encoding: Learning to Relocalize in Minutes Using RGB and Poses". In: *IEEE/CVF Conference on Computer Vision and Pattern*

Recognition (CVPR). Vancouver, Canada: IEEE, June 2023, pp. 5044–5053 (cit. on p. 13).

[28] Fangjinhua Wang, Xudong Jiang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. "GLACE: Global Local Accelerated Coordinate Encoding". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, Washington, USA: IEEE, June 2024, pp. 5819–5828 (cit. on p. 13).

[29] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. "DUSt3R: Geometric 3D Vision Made Easy". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, Washington, USA: IEEE, June 2024, pp. 20697–20709 (cit. on p. 14).

[30] Vincent Leroy, Yohann Cabon, and Jerome Revaud. "Grounding Image Matching in 3D with MASt3R". In: *European Conference on Computer Vision (ECCV)*. Vol. 15130. Lecture Notes in Computer Science. Springer Nature Switzerland, Nov. 2024, pp. 71–91 (cit. on p. 14).

[31] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. "VGGT: Visual Geometry Grounded Transformer". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, Tennessee, USA: IEEE, June 2025, pp. 5294–5306 (cit. on pp. 14, 75).

[32] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. "π^3: Permutation-Equivariant Visual Geometry Learning". In: *International Conference on Learning Representations (ICLR)*. Rio de Janeiro, Brazil, Apr. 2026, pp. 1–17 (cit. on p. 14).

[33] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manue Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. "MapAnything: Universal Feed-Forward Metric 3D Reconstruction". In: *International Conference on 3D Vision (3DV)*. Vancouver, Canada: IEEE, Mar. 2026, pp. 499–509 (cit. on p. 14).

[34] Wang Lu, Suya You, and Ulrich Neumann. "Single View Camera Calibration for Augmented Virtual Environments". In: *IEEE Virtual Reality Conference*. Charlotte, NC: IEEE, 2007, pp. 255–258 (cit. on p. 17).

[35] Zhenling Ma, Xu Zhong, Hong Xie, Yongjun Zhou, Yuan Chen, and Jiali Wang. "A Combined Physical and Mathematical Calibration Method for Low-Cost Cameras in the Air and Underwater Environment". In: *Sensors* 23.4 (Feb. 2023), pp. 1–18 (cit. on p. 17).

[36] Mark Shortis. "Camera Calibration Techniques for Accurate Measurement Underwater". In: *3D Recording and Interpretation for Maritime Archaeology*. Vol. 31. Coastal Research Library. London, England: Springer International Publishing, 2019, pp. 11–27 (cit. on p. 17).

[37] Susumu Kikkawa, Fumio Okura, Daigo Muramatsu, Yasushi Yagi, and Hideo Saito. "Accuracy Evaluation and Prediction of Single-Image Camera Calibration". In: *IEEE Access* 11 (2023), pp. 19312–19323 (cit. on p. 18).

[38] Hanan Quispe, Jorshinno Sumire, Patricia Condori, Edwin Alvarez, and Harley Vera. "I see you: A Vehicle-Pedestrian Interaction Dataset from Traffic Surveillance Cameras". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35. New Orleans, Louisiana, USA: Curran Associates, Inc., Nov. 2022, pp. 1–6 (cit. on p. 18).

[39] Xinyao Tang, Wei Wang, Huansheng Song, and Ying Li. "Novel Optimization Approach for Camera Calibration in Traffic Scenes". In: *Transportation Research Record: Journal of the Transportation Research Board* 2677.3 (Sept. 2022), pp. 1048–1066 (cit. on p. 18).

[40] Wentao Zhang, Huansheng Song, Lichen Liu, Congliang Li, Bochen Mu, and Qian Gao. "Vehicle localisation and deep model for automatic calibration of monocular camera in expressway scenes". In: *IET Intelligent Transport Systems* 16.4 (Dec. 2021), pp. 459–473 (cit. on p. 18).

[41] Todd N. Schoepflin and Daniel J. Dailey. "Dynamic camera calibration of roadside traffic management cameras for vehicle speed estimation". In: *IEEE Transactions on Intelligent Transportation Systems* 4.2 (June 2003), pp. 90–98 (cit. on p. 18).

[42] Yan Xu, Yu-Jhe Li, Xinshuo Weng, and Kris Kitani. "Wide-Baseline Multi-Camera Calibration using Person Re-Identification". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, Tennessee, USA: IEEE, June 2021, pp. 13129–13138 (cit. on p. 18).

[43] Masanori Kano, Hidehiko Okubo, Masaki Takahashi, Kensuke Ikeya, Kensuke Hisatomi, and Tomoyuki Mishina. "Accurate and Practical Calibration of Multiple Pan-Tilt-Zoom Cameras for Live Broadcasts". In: *IEEE Access* 8 (2020), pp. 153993–154006 (cit. on p. 18).

[44] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, and Steve Seitz. “Soccer on Your Tabletop”. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, Utah, USA, June 2018, pp. 4738–4747 (cit. on p. 18).

[45] Vanyi Chao, Ankhzaya Jamsrandorj, Yin May Oo, Kyung-Ryoul Mun, and Jinwook Kim. “3D Ball Trajectory Reconstruction of a Ballistic Shot from a Monocular Basketball Video”. In: *Annual Conference of the IEEE Industrial Electronics Society (IECON)*. Singapore: IEEE, Oct. 2023, pp. 1–6 (cit. on p. 18).

[46] Wanneng Wu, Min Xu, Qiaokang Liang, Li Mei, and Yu Peng. “Multi-camera 3D ball tracking framework for sports video”. In: *IET Image Processing* 14.15 (Dec. 2020), pp. 3751–3761 (cit. on p. 18).

[47] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. “VARS: Video Assistant Referee System for Automated Soccer Decision Making from Multiple Views”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Vancouver, Canada: IEEE, June 2023, pp. 5086–5097 (cit. on pp. 18, 54).

[48] Jan Held, Hani Itani, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. “X-VARS: Introducing Explainability in Football Refereeing with Multi-Modal Large Language Models”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, Washington, USA: IEEE, June 2024, pp. 3267–3279 (cit. on p. 18).

[49] Piotr Kurowski, K. Szeląg, W. Załuski, and Robert Sitnik. “Accurate ball tracking in volleyball actions to support referees”. In: *Opto-Electronics Review* 26.4 (Dec. 2018), pp. 296–306 (cit. on p. 18).

[50] Adrien Maglo, Astrid Orcesi, Julien Denize, and Quoc Cuong Pham. “Individual Locating of Soccer Players from a Single Moving View”. In: *Sensors* 23.18 (Sept. 2023), pp. 1–28 (cit. on pp. 18, 21, 22, 37, 38, 57, 76).

[51] Gabriel Van Zandycke and Christophe De Vleeschouwer. “3D Ball Localization From A Single Calibrated Image”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New Orleans, Louisiana, USA: Institute of Electrical and Electronics Engineers (IEEE), June 2022, pp. 3471–3479 (cit. on p. 18).

[52] Zhongyu Jiang, Haorui Ji, Samuel Menaker, and Jenq-Neng Hwang. "GolfPose: Golf Swing Analyses with a Monocular Camera Based Human Pose Estimation". In: *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. Taipei City, Taiwan: Institute of Electrical and Electronics Engineers (IEEE), July 2022, pp. 1–6 (cit. on p. 18).

[53] Kanav Vats, Mehrnaz Fani, David A. Clausi, and John Zelek. "Puck localization and multi-task event recognition in broadcast hockey videos". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, Tennessee, USA: Institute of Electrical and Electronics Engineers (IEEE), June 2021, pp. 4562–4570 (cit. on p. 18).

[54] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. "DeepSportradar-v1: Computer Vision Dataset for Sports Understanding with High Quality Annotations". In: *International ACM Workshop on Multimedia Content Analysis in Sports (MMSports)*. Lisbon, Portugal: ACM, Oct. 2022, pp. 1–8 (cit. on pp. 18, 21, 82, 86).

[55] Anthony Cioppa, Silvio Giancola, Adrien Deliège, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. "SoccerNet-Tracking: Multiple Object Tracking Dataset and Benchmark in Soccer Videos". In: *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), CVsports*. New Orleans, Louisiana, USA: IEEE, June 2022, pp. 3490–3501 (cit. on p. 19).

[56] 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/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, Tennessee, USA: IEEE, June 2021, pp. 4503–4514 (cit. on pp. 19, 36, 76).

[57] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. "SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Salt Lake City, Utah, USA: IEEE, June 2018, pp. 1711–1721 (cit. on pp. 19, 36, 76).

[58] Arnaud Leduc, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. "SoccerNet-Depth: a Scalable Dataset for Monocular Depth Estimation in Sports Videos". In: *IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition Workshops (CVPRW). Vol. 12. Seattle, Washington, USA: IEEE, June 2024, pp. 3280–3282 (cit. on p. 19).

[59] Bruno Cabado, Anthony Cioppa, Silvio Giancola, Andrés Villa, Bertha Guijarro-Berdiñas, Emilio J. Padrón, Bernard Ghanem, and Marc Van Droogenbroeck. "Beyond the Premier: Assessing Action Spotting Transfer Capability Across Diverse Domains". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Vol. 27. Seattle, Washington, USA: IEEE, June 2024, pp. 3386–3398 (cit. on p. 19).

[60] IFAB. *Laws of the Game*. Tech. rep. Zurich, Switzerland: The International Football Association Board, 2022 (cit. on pp. 19, 42).

[61] FIBA. *2022 Official Basketball Rules: Basketball Rules & Basketball Equipment*. Tech. rep. Mies, Switzerland: Fédération Internationale de Basketball, 2022 (cit. on pp. 19, 54).

[62] FIVB. *Official volleyball rules*. Tech. rep. Lausanne, Switzerland: Fédération Internationale de Volleyball, 2021 (cit. on p. 19).

[63] IIHF. *IIHF Official Rule Book 2022/23*. Tech. rep. Zurich, Switzerland: International Ice Hockey Federation, July 2022 (cit. on pp. 19, 54).

[64] Richard Szeliski. *Computer Vision: Algorithms and Applications*. second. Springer, 2022 (cit. on p. 19).

[65] Maxime Istasse, Vladimir Somers, Pratheeban Elancheliyan, Jaydeep De, and Davide Zambrano. "DeepSportradar-v2: A Multi-Sport Computer Vision Dataset for Sport Understandings". In: *International ACM Workshop on Multimedia Content Analysis in Sports (MMSports)*. Ottawa, Ontario, Canada: ACM, Oct. 2023, pp. 23–29 (cit. on p. 20).

[66] Buzhen Huang, Tianshu Zhang, and Yangang Wang. "Pose2UV: Single-Shot Multiperson Mesh Recovery With Deep UV Prior". In: *IEEE Transactions on Image Processing* 31 (2022), pp. 4679–4692 (cit. on pp. 20, 21).

[67] Tobias Baumgartner and Stefanie Klatt. "Monocular 3D Human Pose Estimation for Sports Broadcasts using Partial Sports Field Registration". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Vancouver, Canada: IEEE, June 2023, pp. 5109–5118 (cit. on pp. 20, 21).

[68] Leonardo Citraro, Pablo Márquez-Neila, Stefano Savarè, Vivek Jayaram, Charles Dubout, Félix Renaut, Andrés Hasfura, Horesh Ben Shitrit, and Pascal Fua. "Real-time camera pose estimation for sports fields". In: *Machine Vision and Applications* 31.3 (Mar. 2020) (cit. on pp. 20–22, 39).

[69] Jonas Theiner and Ralph Ewerth. "TVCalib: Camera Calibration for Sports Field Registration in Soccer". In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, Hawaii, USA: Institute of Electrical and Electronics Engineers (IEEE), Jan. 2023, pp. 1166–1175 (cit. on pp. 20–22, 29, 39, 40, 42, 46, 47, 57, 64, 68, 69, 74).

[70] Feng Shi, Paul Marchwica, Juan Camilo Gamboa Higuera, Mike Jamieson, Mehrsan Javan, and Parthipan Siva. "Self-Supervised Shape Alignment for Sports Field Registration". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, Hawaii, USA: IEEE, Jan. 2022, pp. 3768–3777 (cit. on pp. 20–22, 38, 56).

[71] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun. "Sports Field Localization via Deep Structured Models". In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, Hawaii, USA: IEEE, July 2017, pp. 4012–4020 (cit. on pp. 21, 22, 30, 37, 39, 74, 78, 82, 86).

[72] Rahul Anand Sharma, Bharath Bhat, Vineet Gandhi, and C. V. Jawahar. "Automated Top View Registration of Broadcast Football Videos". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Tahoe, Nevada, USA: IEEE, Mar. 2018, pp. 305–313 (cit. on pp. 21, 22, 38, 56).

[73] 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: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Snowmass, Colorado, USA: IEEE, Mar. 2020, pp. 201–210 (cit. on pp. 21, 22, 38, 56, 74).

[74] Xiaohan Nie, Shixing Chen, and Raffay Hamid. "A Robust and Efficient Framework for Sports-Field Registration". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, Hawaii, USA: IEEE, Jan. 2021, pp. 1935–1943 (cit. on pp. 21–23, 37, 57, 76, 78).

[75] 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: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New

Orleans, Louisiana, USA: IEEE, June 2022, pp. 3522–3529 (cit. on pp. 21–23, 29, 37, 39, 57, 76, 78).

[76] Neng Zhang and Ebroul Izquierdo. "A Fast and Effective Framework for Camera Calibration in Sport Videos". In: *IEEE International Conference on Visual Communications and Image Processing (VCIP)*. Suzhou, China: Institute of Electrical and Electronics Engineers (IEEE), Dec. 2022, pp. 1–5 (cit. on pp. 21, 22).

[77] Pengjie Li, Jianwei Li, Shouxin Zong, and Kaiyu Zhang. "Soccer Field Registration Based on Geometric Constraint and Deep Learning Method". In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Vol. 13020. Lecture Notes in Computer Science. Beijing, China: Springer International Publishing, 2021, pp. 287–298 (cit. on pp. 21, 22).

[78] Hiroki Tsurusaki, Keisuke Nonaka, Ryosuke Watanabe, Tomoaki Konno, and Sei Naito. "Sports Camera Calibration using Flexible Intersection Selection and Refinement". In: *ITE Transactions on Media Technology and Applications* 9.1 (2021), pp. 95–104 (cit. on pp. 21, 22, 37).

[79] Paul Johannes Claasen and Johan Pieter de Villiers. "Video-based sequential Bayesian homography estimation for soccer field registration". In: *Expert Systems with Applications* 252 (Oct. 2024), pp. 1–15 (cit. on pp. 21, 23, 29, 30, 37–39, 57).

[80] Yuan Shu and Yangang Wang. "A camera calibration method based on scene and human semantics". In: *Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. Beijing, China: IEEE, Nov. 2022, pp. 216–221 (cit. on p. 21).

[81] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. "Scaling up SoccerNet with multi-view spatial localization and re-identification". In: *Scientific Data* 9.1 (June 2022), pp. 1–9 (cit. on pp. 25, 28, 36, 56).

[82] FIFA. *Handbook of test methods for virtual offside line assessment*. Tech. rep. Zurich, Switzerland: Fédération Internationale de Football Association, June 2019 (cit. on pp. 29, 54).

[83] Johannes L. Schönberger and Jan-Michael Frahm. "Structure-from-Motion Revisited". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, Nevada, USA: Institute of Electrical and Electronics Engineers (IEEE), June 2016, pp. 4104–4113 (cit. on p. 37).

[84] Raul Mur-Artal, Jose M. M. Montiel, and Juan D. Tardos. "ORB-SLAM: A Versatile and Accurate Monocular SLAM System". In: *IEEE Transactions on Robotics* 31.5 (Oct. 2015), pp. 1147–1163 (cit. on p. 37).

[85] Elan Dubrofsky and Robert J. Woodham. "Combining Line and Point Correspondences for Homography Estimation". In: *Advances in Visual Computing*. Vol. 5339. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, pp. 202–213 (cit. on p. 37).

[86] Miguel Alemán-Flores, Luis Alvarez, Luis Gómez, Pedro Henriquez, and Luis Mazorra. "Camera calibration in sport event scenarios". In: *Pattern Recognition* 47.1 (Jan. 2014), pp. 89–95 (cit. on pp. 37, 39).

[87] Dirk Farin, Susanne Krabbe, Peter H. N. de With, and Wolfgang Effelsberg. "Robust Camera Calibration for Sport Videos using Court Models". In: *Storage and Retrieval Methods and Applications for Multimedia*. Vol. 5307. Proceedings of SPIE. San Jose, California, USA: SPIE, 2004, pp. 1–12 (cit. on pp. 37, 38).

[88] Jean-Bernard Hayet, Justus Piater, and Jacques Verly. "Fast 2D model-to-image registration using vanishing points for sports video analysis". In: *IEEE International Conference on Image Processing (ICIP)*. Genova, Italy: IEEE, Sept. 2005, pp. 1–4 (cit. on p. 37).

[89] Shuhei Tarashima. "Sports Field Recognition Using Deep Multi-task Learning". In: *Journal of Information Processing* 29.0 (2021), pp. 328–335 (cit. on pp. 37, 38).

[90] Marc Gutiérrez-Pérez and Antonio Agudo. "No Bells, Just Whistles: Sports Field Registration by Leveraging Geometric Properties". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, Washington, USA: IEEE, June 2024, pp. 3325–3334 (cit. on pp. 37, 39, 42, 46, 47, 57, 74, 76).

[91] Carlos Cuevas, Daniel Quilón, and Narciso García. "Automatic soccer field of play registration". In: *Pattern Recognition* 103 (July 2020) (cit. on pp. 37, 58).

[92] Nikolay Falaleev. *Top-1 solution of SoccerNet Camera Calibration Challenge 2023*. https://github.com/NikolasEnt/soccernet-calibration-sportlight. June 2023 (cit. on pp. 37, 57).

[93] Peter Andrews, Njål Borch, and Morten Fjeld. "FootyVision: Multi-Object Tracking, Localisation, and Augmentation of Players and Ball in Football Video". In: *International Conference on Multimedia and Image Processing*. Vol. 15. Osaka, Japan: ACM, Apr. 2024, pp. 15–25 (cit. on pp. 37, 58).

[94] Carlos Cuevas, Daniel Berjón, and Narciso García. "Grass band detection in soccer images for improved image registration". In: *Signal Processing: Image Communication* 109 (Nov. 2022), pp. 1–16 (cit. on p. 37).

[95] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. "KaliCalib: A Framework for Basketball Court Registration". In: *International ACM Workshop on Multimedia Content Analysis in Sports (MMSports)*. Dublin, Ireland: ACM, Oct. 2022, pp. 111–116 (cit. on pp. 37, 74).

[96] Neng Zhang and Ebroul Izquierdo. "A High Accuracy Camera Calibration Method for Sport Videos". In: *IEEE International Conference on Visual Communications and Image Processing (VCIP)*. Munich, Germany: IEEE, Dec. 2021, pp. 1–5 (cit. on pp. 38, 56).

[97] Giacomo D'Amicantonio, Egor Bondarev, and Peter H. N. de With. "Automated Camera Calibration via Homography Estimation with GNNs". In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, Hawaii, USA: IEEE, Jan. 2024, pp. 5864–5871 (cit. on pp. 38, 56).

[98] Mehrnaz Fani, Pascale Berunelle Walters, David A. Clausi, John Zelek, and Alexander Wong. "Localization of Ice-Rink for Broadcast Hockey Videos". In: *arXiv* abs/2104.10847 (2021). arXiv: 2104.10847 (cit. on pp. 38, 56).

[99] Jean-Bernard Hayet, Justus Piater, and Jacques Verly. "Incremental rectification of sports fields in video streams with application to soccer". In: *Advanced Concepts for Intelligent Vision Systems (ACIVS)*. Brussels, Belgium, Aug. 2004, pp. 1–8 (cit. on p. 38).

[100] Ankur Gupta, James J. Little, and Robert J. Woodham. "Using line and ellipse features for rectification of broadcast hockey video". In: *Canadian Conference on Computer and Robot Vision*. St. Johns, Canada, May 2011, pp. 32–39 (cit. on pp. 38, 58).

[101] Jianhui Chen and Peter Carr. "Mimicking Human Camera Operators". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, Hawaii, USA: IEEE, Jan. 2015, pp. 215–222 (cit. on pp. 38–40).

[102] Michael Beetz, Suat Gedikli, Jan Bandouch, Bernhard Kirchlechner, Nico Hoyningen-Huene, and Alexander Perzylo. "Visually Tracking Football Games Based on TV Broadcasts". In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Hyderabad, India, Jan. 2007, pp. 2066–2071 (cit. on p. 39).

[103] Jikai Lu, Jianhui Chen, and James J. Little. "Pan-tilt-zoom SLAM for Sports Videos". In: *British Machine Vision Conference (BMVC)*. Cardiff, Wales: British Machine Vision Association (BMVA), Sept. 2019, pp. 1–14 (cit. on pp. 39, 46).

[104] Peter Carr, Yaser Sheikh, and Iain Matthews. "Point-less calibration: Camera parameters from gradient-based alignment to edge images". In: *IEEE Workshop on the Applications of Computer Vision (WACV)*. Breckenridge, CO, USA: IEEE, Jan. 2012, pp. 377–384 (cit. on pp. 39, 40).

[105] Jianhui Chen, Fangrui Zhu, and James J. Little. "A Two-Point Method for PTZ Camera Calibration in Sports". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Tahoe, Nevada, USA: IEEE, Mar. 2018, pp. 287–295 (cit. on pp. 39, 45).

[106] ITU. *Parameter values for the HDTV standards for production and international programme exchange*. Recommendation ITU-R BT.709-6. 2015 (cit. on p. 39).

[107] ITU. *Parameter values for ultra-high definition television systems for production and international programme exchange*. Recommendation ITU-R BT.2020-2. 2015 (cit. on p. 39).

[108] Zhengyou Zhang. "A flexible new technique for camera calibration". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.11 (2000), pp. 1330–1334 (cit. on p. 40).

[109] Fujifilm. *Television lenses & cine lenses*. Technical note. 2024 (cit. on p. 40).

[110] Sachtler. *Video 25 Plus and Video 25 Plus FB: Manual*. Product description. 2023 (cit. on p. 41).

[111] Vinten. *Vision 250 Pan & Tilt Head*. Specifications. 2023 (cit. on p. 41).

[112] Dorin Comaniciu and Peter Meer. "Mean shift: a robust approach toward feature space analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5 (May 2002), pp. 603–619 (cit. on p. 42).

[113] Jean-Yves Bouguet. *Pyramidal Implementation of the Lucas Kanade Feature Tracker – Description of the Algorithm*. OpenCV documentation. 2001 (cit. on p. 43).

[114] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. "RTMDet: An Empirical Study of Designing Real-Time Object Detectors". In: *arXiv* abs/2212.07784 (2022). arXiv: 2212.07784 (cit. on p. 43).

[115] Suat Gedikli. "Continual and Robust Estimation of Camera Parameters in Broadcasted Sports Games". PhD thesis. Technische Universität München, Apr. 2008 (cit. on p. 45).

[116] Floriane Magera. *SoccerNet Camera Calibration Challenge*. https://github.com/SoccerNet/sn-calibration. June 2022 (cit. on p. 47).

[117] Changjia Tian, Varuna De Silva, Michael Caine, and Steve Swanson. "Use of Machine Learning to Automate the Identification of Basketball Strategies Using Whole Team Player Tracking Data". In: *Applied Sciences* 10.1 (Dec. 2019), pp. 1–17 (cit. on p. 54).

[118] Keisuke Fujii. *Machine Learning in Sports: Open Approach for Next Play Analytics*. SpringerBriefs in Computer Science. Springer Nat. Singap., 2025 (cit. on p. 54).

[119] Marc Gutiérrez-Pérez and Antonio Agudo. "PnLCalib: Sports field registration via points and lines optimization". In: *Computer Vision and Image Understanding* 267 (Apr. 2026), p. 104712 (cit. on pp. 55, 57, 63, 64, 68, 71).

[120] Piotr Skalski. *Camera Calibration in Sports with Keypoints*. Blog. Aug. 2024 (cit. on pp. 56, 58).

[121] Nikolay S. Falaleev and Ruilong Chen. "Enhancing Soccer Camera Calibration Through Keypoint Exploitation". In: *International ACM Workshop on Multimedia Content Analysis in Sports (MMSports)*. Vol. 6. Melbourne, Victoria, Australia: ACM, Oct. 2024, pp. 65–73 (cit. on pp. 57, 74).

[122] Luis Alvarez and Vicent Caselles. "Homography estimation using one ellipse correspondence and minimal additional information". In: *IEEE International Conference on Image Processing (ICIP)*. Paris, France: IEEE, Oct. 2014, pp. 4842–4846 (cit. on p. 58).

[123] Ondřej Chum and Jiří Matas. "Homography Estimation from Correspondences of Local Elliptical Features". In: *IEEE International Conference on Pattern Recognition (ICPR)*. Tsukuba, Japan: IEEE, Nov. 2012, pp. 3236–3239 (cit. on p. 58).

[124] Qian Chen, Haiyuan Wu, and Toshikazu Wada. "Camera Calibration with Two Arbitrary Coplanar Circles". In: *European Conference on Computer Vision (ECCV)*. Vol. 3023. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004, pp. 521–532 (cit. on p. 58).

[125] Haifei Huang, Hui Zhang, and Yiu-Ming Cheung. "Homography Estimation from the Common Self-Polar Triangle of Separate Ellipses". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, Nevada, USA: IEEE, June 2016, pp. 1737–1744 (cit. on p. 58).

[126] Haifei Huang, Hui Zhang, and Yiu-ming Cheung. "The common self-polar triangle of concentric circles and its application to camera calibration". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, Massachusetts, USA: IEEE, June 2015, pp. 4065–4072 (cit. on pp. 59, 63).

[127] Andrew W. Fitzgibbon and Robert B. Fisher. "A Buyer's Guide to Conic Fitting". In: *British Machine Vision Conference (BMVC)*. Swansea, UK: British Machine Vision Association (BMVA), 1995, pp. 1–10 (cit. on p. 68).

[128] Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. "Unified Sports Field Registration with Lens Distortion Modeling". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Denver, Colorado, USA: IEEE, June 2026, pp. 10067–10076 (cit. on pp. 74, 76, 82, 86).

[129] Tianjian Jiang, Johsan Billingham, Sebastian Müksch, Juan Zarate, Nicolas Evans, Martin R. Oswald, Marc Pollefeys, Otmar Hilliges, Manuel Kaufmann, and Jie Song. "WorldPose: A World Cup Dataset for Global 3D Human Pose Estimation". In: *European Conference on Computer Vision (ECCV)*. Vol. 15077. Lecture Notes in Computer Science. Springer Nature Switzerland, Dec. 2024, pp. 343–362 (cit. on pp. 76, 82, 86).

[130] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. "Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-Shot Metric Depth and Surface Normal Estimation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12 (Dec. 2024), pp. 10579–10596 (cit. on pp. 77, 83–86).

[131] Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. "Depth Anything 3: Recovering the Visual Space from Any Views". In: *International Conference on Learning Representations (ICLR)*. Rio de Janeiro, Brazil, Apr. 2026, pp. 1–25 (cit. on pp. 77, 83, 84, 86).

[132] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. "MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision".

In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, Tennessee, USA: IEEE, June 2025, pp. 5261–5271 (cit. on p. 77).

[133] Muxin Liu, Xiaoyang Lyu, Tianhe Ren, Peng Dai, Xiaoshan Wu, Zhiyue Zhang, Jiaqi Zhang, Jiehong Lin, Shaoshuai Shi, and Xiaojuan Qi. "FoundationGeo: Learning Spatial Pixel-Wise Fields for Monocular Metric Geometry". In: *European Conference on Computer Vision (ECCV)*. Vol. 17060. Lecture Notes in Computer Science. Springer Nature Switzerland, 2026, pp. 353–371 (cit. on pp. 77, 82, 84).

[134] Rafael Grompone von Gioi, Jérémie Jakubowicz, Jean-Michel Morel, and Gregory Randall. "LSD: a Line Segment Detector". In: *Image Processing On Line* 2 (Mar. 2012), pp. 35–55 (cit. on pp. 77, 80, 83).

[135] Remi Pautrat, Juan-Ting Lin, Viktor Larsson, Martin R. Oswald, and Marc Pollefeys. "SOLD2: Self-supervised Occlusion-aware Line Description and Detection". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, Tennessee, USA: IEEE, June 2021, pp. 11363–11373 (cit. on p. 77).

[136] Nan Xue, Tianfu Wu, Song Bai, Fudong Wang, Gui-Song Xia, Liangpei Zhang, and Philip H.S. Torr. "Holistically-Attracted Wireframe Parsing". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, Washington, USA: IEEE, June 2020, pp. 2785–2794 (cit. on p. 77).

[137] Yifan Xu, Weijian Xu, David Cheung, and Zhuowen Tu. "Line Segment Detection Using Transformers without Edges". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, Tennessee, USA: IEEE, June 2021, pp. 4255–4264 (cit. on p. 77).

[138] Rémi Pautrat, Daniel Barath, Viktor Larsson, Martin R. Oswald, and Marc Pollefeys. "DeepLSD: Line Segment Detection and Refinement with Deep Image Gradients". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, Canada: IEEE, June 2023, pp. 17327–17336 (cit. on pp. 77, 80, 83).

[139] Roberto Toldo and Andrea Fusiello. "Robust Multiple Structures Estimation with J-Linkage". In: *European Conference on Computer Vision (ECCV)*. Vol. 5302. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, pp. 537–547 (cit. on p. 77).

[140] Dániel Baráth and Jiří Matas. "Progressive-X: Efficient, Anytime, Multi-Model Fitting Algorithm". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, South Korea: IEEE, Oct. 2019, pp. 3779–3787 (cit. on pp. 77, 80).

[141] Bangyan Liao, Zhenjun Zhao, Haoang Li, Yi Zhou, Yingping Zeng, Hao Li, and Peidong Liu. "Convex Relaxation for Robust Vanishing Point Estimation in Manhattan World". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, Tennessee, USA: IEEE, June 2025, pp. 15823–15832 (cit. on p. 77).

[142] Rémi Pautrat, Shaohui Liu, Petr Hruby, Marc Pollefeys, and Dániel Baráth. "Vanishing Point Estimation in Uncalibrated Images with Prior Gravity Direction". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, Oct. 2023, pp. 14072–14081 (cit. on pp. 77, 82, 84).

[143] David Liebowitz and Andrew Zisserman. "Metric rectification for perspective images of planes". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Santa Barbara, California, USA: IEEE, June 1998, pp. 482–488 (cit. on p. 78).

[144] Antonio Criminisi, Ian Reid, and Andrew Zisserman. "Single View Metrology". In: *International Journal of Computer Vision* 40.2 (Nov. 2000), pp. 123–148 (cit. on p. 78).

[145] Jean-Philippe Tardif. "Non-iterative approach for fast and accurate vanishing point detection". In: *IEEE International Conference on Computer Vision (ICCV)*. Kyoto, Japan: IEEE, Sept. 2009, pp. 1250–1257 (cit. on p. 80).

[146] Vizrt. *Viz Arena: The one-box AR and Virtual Advertising suite for live events*. Web site: https://routing.vizrt.com/products/viz-arena/, accessed: 2026-07-01. 2026 (cit. on p. 82).