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

Sport Field Calibration with NeRF-guided Camera Optimization from a Single Image

Liang Fan *
Loughborough University
L.Fan@lboro.ac.uk

Xianqian Liu*
Beijing Jiaotong University
xiaoqianliu@bjtu.edu.cn

Malcolm Roberts
Irish Football Association
malcolm.roberts@irishfa.com

Abstract

Existing camera calibration methods struggle in dynamic sports environments due to transient scenes, poor lighting, and occlusions. Many state-of-the-art approaches depend on large-scale synthetic or multi-view datasets, which limit their real-world applicability. To overcome these challenges, we propose a novel monocular camera calibration framework for single-frame football field image, leveraging semantic attention mechanisms and neural radiance fields (NeRFs) to enhance camera pose estimation. Our approach extracts key structural features from the sports field using a DeepLabv3+ model to calculate initial object coordinates and camera calibration parameters. We refine spatial feature representation by integrating scene semantics into Pixel-NeRF via an attention mechanism for improved positional encoding. Finally, we globally optimize camera calibration parameters using semantic and geometric features to achieve defined 3D reconstruction and camera pose estimation. Our method achieves state-of-the-art results on the SoccerNetV3 and World Cup 2014 datasets, reducing the reprojection error by 0.5% and reaching 92.1% IoU.

1. Introduction

In the field of sports video analysis, monocular cameras [1, 13, 26] offer advantages such as low cost, ease of installation, and rapid data processing, making them suitable for a variety of sports, including basketball and tennis, and are particularly widely used in soccer scenarios. However, factors such as lens distortion, perspective changes, and lighting variations, necessitate regular calibration of monocular cameras to ensure data accuracy and system stability.

The calibration of monocular cameras in sports field analysis can be broadly classified into three methodological paradigms: feature-based, retrieval-based, and learning-based approaches. The feature-based methods [4, 16, 33] rely on extracting geometric features from the sports field, such as field lines, goalposts, and boundary markers, which are then aligned with a predefined sports field model to estimate camera parameters through robust feature detection. Although effective, these methods often struggle with occlusions and varying lighting conditions. The retrieval-based methods [5, 19] present camera calibration as a template-matching problem [25], utilizing pre-computed edge maps or homography field templates to optimize the alignment of an input image. These methods incorporate strong geometric constraints, their performance is inherently limited by template variability and restricted camera viewpoints. In contrast, learning-based methods [9, 36] introduce specialized loss functions to optimize camera parameters directly using end-to-end deep neural networks. By incorporating scene structure information derived from segmentation, these methods refine camera poses through gradient-based optimization, demonstrating improved robustness in complex and dynamic environments.

In summary, existing methods offer viable solutions for specific camera calibration tasks, their effectiveness is significantly constrained in dynamic sports environments due to the highly transient nature of sports scenes, suboptimal lighting conditions, and frequent occlusions. Furthermore, a notable limitation of many state-of-the-art approaches, including pseudo-stereo 3D framework and DUSt3R [7, 39], is their reliance on large-scale synthetic datasets and multi-view datasets for training. Acquiring such datasets in real-world scenarios is often impractical, limiting the generalizability and scalability of these methods in practical applications. These shortcomings underscore the need for more robust and data-efficient calibration frameworks tailored to the complexities of sports video analysis. Therefore, we propose a novel framework for monocular camera calibration in the context of single-frame football field image, integrating semantic perceptual attention mechanisms and enhancing camera pose estimation through neural radiance fields (NeRF). The key contributions of this work are summarized as follows:

- **A single-frame monocular camera calibration method.** We present a calibration scheme that does

*Equal contribution.

not rely on temporal frame relationships. Our approach calibrates the camera using a single image to enable stability even when the viewpoint changes frequently.

- **Fusion of depth estimation and scene geometry priors.** We utilize geometric priors that are from monocular depth estimation in a sports field to improve the accuracy of camera calibration estimation. This fusion compensates for the lack of depth data that is typically encountered in monocular camera setups.
- **Semantic-aware attention positional encoding.** The attention mechanism dynamically adjusts the weights of geometric information in the sports field. This mechanism enhances the model's ability to perceive details in key areas for improving the accuracy of geometric alignment.
- **Joint constraint strategy to optimize camera poses using scene structure and geometric information.** By aligning rendered images with real-world images, the joint constraint strategy refines camera parameters through geometric and photometric consistency, improving calibration accuracy, particularly for wide-angle views.
- **Reprojection and Homography Optimization for SoccerNetV3 and World Cup 2014.** We evaluate the reprojection and homography performance on the SoccerNetV3 and World Cup 2014 datasets. The reprojection error on the SoccerNetV3 dataset is reduced by 0.5% compared to the state-of-the-art method. Furthermore, the Intersection over Union (IoU) on the World Cup 2014 dataset increases to 92.1%, achieving the state-of-the-art performance.

2. Related work

2.1. Single-frame monocular camera calibration

Single-image camera calibration has emerged as a critical area of research in computer vision, particularly in scenarios where access to multiple views or calibration patterns is limited. Traditional approaches rely predominantly on geometric cues, such as vanishing points, parallel lines, and symmetry, to infer camera parameters. For instance, the single-view auto-calibration method [18, 27] introduces a solver that combines imaged translational symmetries and parallel scene lines, allowing for simultaneous estimation of lens distortion and camera pose. While such geometry-based methods are effective in estimating camera angles within the Manhattan world framework, their performance is often compromised by insufficient geometric constraints and images with sparse arc structures, limiting their applicability in complex environments. To address these limitations, recent advancements have incorporated deep learning techniques. CTRL-C [23] feeds multimodal data (for example, image feature and line segments of the image) into a transformer architecture, by combining the learned features with detected vanishing points, to improve performance under the Manhattan world assumption. Perspective Fields [21], for example, proposed a deep network-based approach to recover intrinsic parameters by detecting and leveraging vanishing points from intersecting lines. In contrast, MSCC [34] introduced a transformer-based model that employs an attention mechanism to capture long-range dependencies with input images. However, these methods often struggle with inference time and require extensive labelled data. In the domain of sports field calibration, challenges are exacerbated by the dynamic and unstructured nature of sports scenes, including frequent occlusions, varying lighting conditions, and stringent requirements for real-time processing. A common approach to sports field calibration involves estimating homography matrices from sports videos using Direct Linear Transformation (DLT), such as a RVR model [32]. State-of-the-art methods [18, 38] often integrate geometric constraints with machine learning techniques to enhance the accuracy and the robustness. Chen et al. [4] proposed a framework that initializes camera poses and trains a Siamese Generative Adversarial Network (GAN) using synthetic data with known 2D–3D correspondences. This network extracts field edges by comparing the differences between synthetic scenes and actual images, enabling precise edge alignment. Chu et al. [9] introduced an instance segmentation approach to isolate individual field areas, while a Variational Autoencoder (VAE) built on a U-Net architecture was employed to learn the geometry of the scene and refine the precision point features with additional contextual information. Tvcalib [36] established an optimization-based framework that a keypoint-less semantic segmentation model computes camera parameters by minimizing image registration and segment projection errors. NBJW [16] takes advantage of the geometric characteristics of the field to extract a predefined keypoint grid and employs an encoder-decoder network to predict keypoint locations. Jiang et al. [20] proposed a two-step deep learning framework to optimize homography estimation through iterative refinement. However, current calibration methods rely on large amounts of data to optimize the reprojection error. Additionally, few multiple perspective data make the reprojection error difficult to refine.

2.2. NeRF and poses joint optimization

Neural Radiance Fields (NeRF) have revolutionized 3D scene reconstruction and novel view synthesis by implicitly modeling scenes as continuous volumetric functions. A critical challenge in NeRF-based methods lies in the joint optimization of scene geometry and camera poses, particularly when initial pose estimates are inaccurate. Recent research has introduced innovative techniques to address this challenge, significantly improving both rendering quality and pose accuracy.

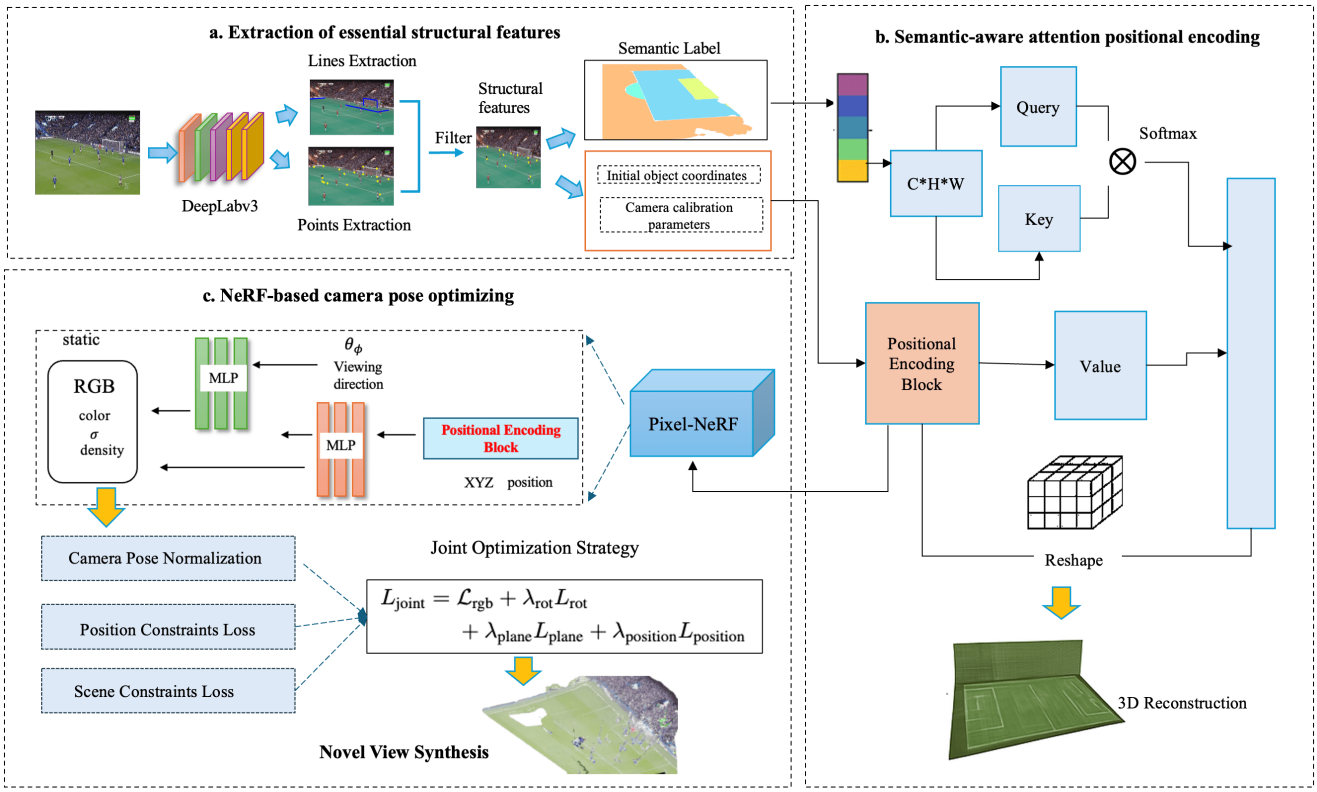


Figure 1. The monocular camera calibration framework of a single-frame football field image. The framework consists of: (a) key structural feature extraction using DeepLabv3+ for initial object coordinates and camera calibration; (b) semantic-aware attention-based positional encoding to enhance spatial representation in Pixel-NeRF; and (c) NeRF-based camera pose optimization using semantic and geometric features for accurate 3D reconstruction.

ColNeRF [30] improves scene geometry convergence by incorporating collaborative compensation and constraints into the NeRF framework, leading to more accurate camera pose optimization. Similarly, PixelNeRF [42] conditions NeRF on local features, projecting input images directly into the latent space of a pre-trained 3D-aware generative network. This approach enables the prediction of continuous neural representations conditioned on the input image, rather than relying solely on input view supervision, thereby improving generalization. iNeRF [41] explores pose refinement within a pre-trained NeRF model, leveraging the implicit scene representation to iteratively optimize camera poses. NeRFDeformer [35] introduces 3D scene flow on surfaces to map the original configuration to a transformed one, enabling successful scene transformation using only a single RGBD observation of the altered scene. NoPe-NeRF [2] proposes a pose-free NeRF optimization method that constrains camera pose by correcting undistorted monocular depth priors. CamP [31] further advances joint optimization by introducing a preconditioner to reduce correlations between camera parameters, enhancing the simultaneous optimization of camera poses and NeRF parameters.

3. Methodology

As shown in Fig. 1, we use a DeepLabv3+ model [6] to extract key structural features from the sports field. Combined with the extracted structural features, the initialized coordinates of each object and a set of camera calibration parameters in the scene are further calculated. To enhance the network's ability to represent spatial details, we introduce the attention mechanism that integrates scene semantic labels into the Pixel-NeRF network to achieve enhanced position encoding of spatial features. Finally, the calibration parameters network is globally optimized based on the scene's semantic information and geometric structure features to obtain more accurate 3D reconstruction and camera parameter estimation using a PixelNeRF architecture.

3.1. Extraction of essential structural features

DeepLabv3+ adopts an encoder-decoder architecture to extract high-precision semantic labels. The encoder utilizes ResNet as the backbone network for feature extraction to capture semantic information at different scales through multiple convolutional layers and residual connections from

the image *I*. After the image *I* passes the backbone network, the Atrous Spatial Pyramid Pooling (ASPP) module, which consists of a $1\times$ 1 convolution layer and three $3\times$ 3 convolution layers with different dilation rates, efficiently capture global contextual information and a wide range of receptive fields Y_{ASPP}. After four upsampling layers from the encoder, the decoder analyzes the semantic scene labels Y_{decoder} with spatial structural features.

3.2. Semantic-aware attention positional encoding

Traditional Perspective-n-Point (PnP) methods [24] rely on known 3D scene points and their 2D projections to estimate camera parameters. However, in single-image scenarios, the absence of depth information introduces ambiguity, as a single 2D projection can correspond to multiple possible 3D positions. This ambiguity makes it challenging to accurately determine the camera translation vector $\mathbf{t}$ and the rotation matrix $\mathbf{R}$. While NeRF can infer camera parameters by learning implicit representations f_θ, this method struggles with limited perspective constraints using a single-view image. The PixelNeRF addresses this challenge by combining ray encoding and positional encoding to model the relationship between image pixels and scene depth, enabling 3D structure and camera parameter estimation from a single image. Nevertheless, PixelNeRF's lack of spatial detail enhancement limits its ability to accurately capture the geometry and positional relationships of specific regions in dynamic scenes, even when positional encoding aligns image plane pixels with 3D rays.

We propose an attention-based positional encoding network augmented with semantic labels to overcome these limitations. This approach integrates semantic information (e.g., semantic scene labels) and attention mechanisms into the positional encoding process, enhancing the network's ability to represent spatial details and dynamically weight different scene regions. As illustrated in Fig. 2, for a pixel (u_i, v_i), the position of the feature $\gamma(u_i, v_i)$ is derived from the following positional encoding formulation:

$$\begin{aligned}\gamma(u_i, v_i) = \Big(& \sin(2^0\pi(u_i, v_i)), \cos(2^0\pi(u_i, v_i)), \\ & \cdots, \\ & \sin(2^L\pi(u_i, v_i)), \cos(2^L\pi(u_i, v_i))\Big)\end{aligned} \tag{1}$$

where L denotes the encoded frequency level. A larger value of L incorporates more high-frequency features, enhancing the model's ability to capture fine-grained details. To further leverage the effectiveness of semantic information, an embedding layer is applied to the semantic scene labels generated in Section 3.1, encoding them as $E(\gamma(x))$. This embedding layer maps discrete semantic labels into a continuous, dense representation, enriching the encoded

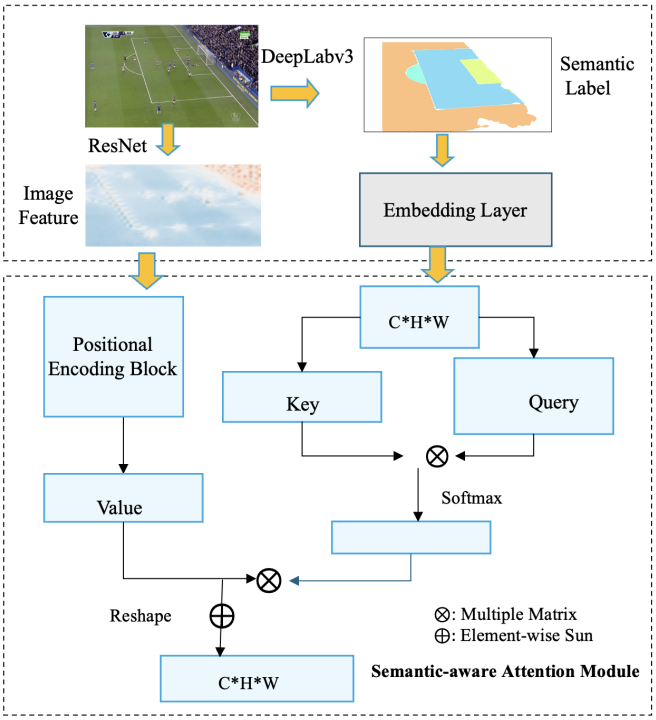


Figure 2. The structure of an attention-based positional encoding network with the semantic label.

features with contextual semantic information. Subsequently, a scaled dot-product mechanism is employed to quantify the similarity between the query Q and the key K, enabling the model to dynamically focus on the most relevant spatial and semantic features.

$$Similarity(Q, K) = \sum_i Q \cdot k \tag{2}$$

Then, the similarity score for all pixels $Similarity(Q, K)$ is normalized using the SoftMax function to compute an attention weight, which dynamically assigns importance to different spatial regions. Finally, the attention-based positional encoding is derived as follows:

$$\gamma_{\text{attention}}(u_i, v_i) = Similarity(Q, K) \times V \tag{3}$$

where V represents the high-dimensional feature F extracted by the image encoder in the PixelNeRF model. This attention-based positional encoding mechanism leverages semantic scene and geometric features to enhance the precision of 3D scene reconstruction and camera parameter estimation. By incorporating semantic scene context and adaptive feature weighting, the method significantly improves the model's ability to capture complex spatial relationships and refine geometric details, leading to more accurate and robust results in challenging scenarios.

3.3. NeRF-based camera pose optimizing

After transferring the original feature positions into high-dimensional feature representations through attention encoding, a multi-layer perceptron (MLP) network is used to refine the camera pose. In the multi-layer perceptron (MLP), the camera poses' parameters and geometric information related to the 3D structure are gradually extracted from spatial information and perspective-dependent features based on the optimization of NeRF's parameters Θ^*.

In the context of pose-unknown Neural Radiance Fields (NeRF), our objective is to jointly optimize the scene model M_θ by refining both the NeRF parameters θ and the camera parameters ϕ to accurately reconstruct the observed image. Given an input image I and its corresponding camera pose π, the NeRF framework minimizes the photometric error $\mathcal{L}_{rgb}$ through a volumetric rendering process. The core challenge lies in regulating the camera rays $\mathbf{r}$ after joint optimization of the parameters Θ^* and π^*, which can be formally expressed as:

$$\Theta^*, \Pi^* = \arg\min_{\Theta, I} \mathcal{L}_{rgb}(\hat{I}, \hat{\Pi}|I) \tag{4}$$

where $\hat{\Pi}$ represents the updated camera parameters during the optimization process. Typically, camera parameters Π encompass intrinsic parameters (e.g., focal length, principal point), extrinsic parameters (e.g., rotation and translation), and lens distortion coefficients. In this work, we focus exclusively on estimating camera poses using NeRF, without explicitly modelling intrinsic parameters or lens distortions.

To address the challenges of joint optimization, we introduce a joint optimization strategy that penalizes incorrect camera poses, thereby regularizing the optimization process and ensuring smoother convergence. Unlike Unposed-NeRF methods [8], which optimize each camera pose independently and often suffer from overfitting to incorrect poses for target images, our approach incorporates three novel loss functions: pose normalization, position constraints, and planarity constraints. These losses collectively constrain camera poses, promoting more accurate and stable optimization. Pose normalization ensures that the estimated poses remain within a physically plausible range, position constraints enforce consistency in the camera's spatial location, and planarity constraints leverage geometric priors to refine the alignment of the reconstructed scene.

3.3.1. Camera pose normalization

To ensure the stability and generalization capability of the model, the rotation matrix R is constrained to remain an orthogonal matrix throughout the optimization process, preserving the rigid transformation properties of the camera pose. This is particularly critical when the initial camera parameters are close to the true values, as it prevents deviations from physically valid rotations. To enforce this constraint, we introduce a distance metric based on a Lie algebra, defined as the loss function $L_{rot} = \|\log(R)\|^2$. Here, $\log(\cdot)$ represents the logarithmic mapping in a Lie algebra, which quantifies the difference between rotation matrices in a continuous and differentiable manner.

3.3.2. Planarity constraints

The planarity constraints loss leverages salient planes in the football field, such as penalty areas and 10-meter zones, to ensure that the 3D structure generated by our Pixel-NeRF network aligns with the real-world scene after optimizing the camera parameters. Each image in the dataset contains these salient planes, which form a set of geometric constraints. Given the flat and horizontal nature of football fields, all 3D points are assumed to lie on the ground plane. Thus, the position loss is defined as:

$$L_{plane} = \frac{1}{N}\sum_{i=1}^{N} |Ax' + By' + Cz' + D| \tag{5}$$

where N represents the number of 3D points, and (A, B, C) is the normal vector of the plane. Assuming that the football pitch is perfectly horizontal, the normal vector simplifies to $(0, 0, 1)$, and the depth D of the plane is set to 0. Consequently, the position loss further simplifies to:

$$L_{plane} = \frac{1}{N}\sum_{i=1}^{N} |z'| \tag{6}$$

3.3.3. Position constraint

To reduce the depth distortion caused by NeRF, depth information provides a priori for ray sampling, assisting in sampling by assuming the distance and relative position of the players. We introduce a position constraint based on a Gaussian kernel function [22], utilizing a monocular depth map to model spatial relationships between players. Specifically, players in closer proximity are assigned higher weights, while those farther apart receive lower weights, emphasizing the spatial dependencies between players. The scene constraint is formulated as:

$$L_{position} = \sum_{(m,n)} D_{mn} \tag{7}$$

$$D_{mn} = w_{mn} \cdot \|m - n\| \tag{8}$$

where D_{mn} is the distance between the depth of player m and player n. We introduce DETR (Detection Transformer) [3] to detect m and n locations with the use of the depth priors which are calculated by the Depth Anything2 model [40]. ω_{mn} is the weight between player m and player n.

$$\omega_{mn} = exp(-\frac{\|m - n\|^2}{2\sigma_{pc}^2}) \tag{9}$$

The weight values ω_{mn} are computed based on the proximity of players using a Gaussian kernel applied to the distance matrix. $\|\mathbf{m} - \mathbf{n}\|^2$ represents the squared Euclidean distance between the player m and the player n. σ_{pc} denotes the standard deviation of the Gaussian kernel, controlling the influence of proximity on the weight assignment.

The overall joint loss function L_{joint} integrates three components: camera pose normalization loss, planarity constraints, and position constraints. This unified loss function not only optimizes the absolute 3D positions of players but also ensures that the relative distances between players align with the real-world scene.

$$L_{\text{joint}} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{rot}} L_{\text{rot}} + \lambda_{\text{plane}} L_{\text{plane}} + \lambda_{\text{position}} L_{\text{position}} \tag{10}$$

where λ_{rot}, λ_{plane}, and $\lambda_{position}$ are the weighting coefficients for the rotation loss, planarity loss, and position loss, respectively. $\mathcal{L}_{\text{rgb}}$ is the rendering loss function of the pixelNeRF. In our implementation, we set $\lambda_{rot} = 0.3$, $\lambda_{plane} = 1$, and $\lambda_{position} = 1$ to balance the contributions of each term. The joint optimization strategy aims to minimize L_{joint}, which is expressed as:

$$\Theta^*, \Pi^* = \arg\min_{\Theta, I} \mathcal{L}_{joint}(\hat{I}, \hat{\Pi}, \hat{D} | I, D) \tag{11}$$

Through this optimization, our NeRF framework jointly refines the camera pose Π^* and the scene representation parameters Θ^*, ensuring accurate 3D reconstruction and pose estimation in complex scenes.

4. Experiment

4.1. Hyperparameters & metrics

Our model architecture builds upon the PixelNeRF, leveraging its framework to train camera calibration parameters. We employ the Adam optimizer with distinct learning rates for different components: the initial learning rate for NeRF is set to 1×10^{-4}, while the learning rates for pose and distortion parameters are set to 5×10^{-4}. A weight decay of 1×10^{-6} is applied to regularize the optimization process. During training, we first optimize the model using all loss functions, including photometric, geometric, and surface-based constraints. Once the surface-based constraints loss converges, we transition to refining the model exclusively with the photometric reconstruction loss $\mathcal{L}_{rgb}$.

IoU: We evaluate calibration accuracy using the Intersection-over-Union (IoU) score, which measures the alignment between projected field models from ground truth and estimated camera parameters in a top-down view. To address challenges, such as occlusion, viewing angles, and cropping, we use partial IoU for visible regions and whole IoU for overall viewpoint accuracy, ensuring robust evaluation under real-world conditions.

Reprojection error for camera calibration: To assess the accuracy of mapping 3D world points to 2D image points, we employ the reprojection error as a key metric. For each feature point in the 3D scene x_w, the smaller Euclidean distance between the projected point x_{proj} and the corresponding point x_{actual} in the actual image reflects the high accuracy of the calibration.

$$E_{reproj} = \sqrt{\begin{matrix}(x_{proj,x} - x_{actual,x})^2 \\ + (x_{proj,y} - x_{actual,y})^2\end{matrix}} \tag{12}$$

To provide a comprehensive evaluation, we calculate the mean reprojection error by averaging the errors across all points, offering an overall assessment of the calibration model. In addition, we compute the median reprojection error to mitigate the influence of outliers and extreme errors, ensuring a robust measure of calibration performance. These metrics collectively provide a thorough analysis of the precision and reliability of our calibration framework.

4.2. Dataset

To validate the performance of our camera calibration system, we use the SoccerNetV3-Calibration dataset [11] [15] for training. The SoccerNetV3 test data and World Cup 2014 (WC14) datasets [19] are testing data.

The SoccerNetV3 dataset The SoccerNetV3-calibration dataset includes 22,819 images, which collected from 500 matches (2014–2017) from six European leagues. All images in the SoccerNetV3-calibration dataset are split into 16,464 training, 3,213 validation, and 3,142 test images. All images were captured from five camera viewpoints (master, off-side, and high behind). Each image has detailed field line annotations, including curved elements like the center circle and penalty arcs, represented as point lists. This diversity enhances our model's robustness.

The World Cup 2014 dataset The WC14 dataset [12, 28] contains 395 broadcast images from 20 matches. This dataset provides homography matrices, though annotations contain errors, affecting Intersection-over-Union (IoU) evaluation. The WC14 dataset offers realistic scenarios with varying illumination, angles, and occlusion, making it a valuable benchmark for real-world calibration.

4.3. The initialized camera parameters

The joint optimization strategy Pixel-Nerf received camera parameters as initialization. We selected the pinhole camera model [17]:

$$S \bullet \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \bullet [R|t] \bullet \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{13}$$

to project pixel coordinates from camera coordination. The intrinsic matrix $K \in \mathbb{R}^{3\times3}$, including the principal point

(C_x, C_y) and the focal length (f_x, f_y), is an intrinsic matrix to control the internal geometry of the projection. We assume that the principal point is in the center of the image, thus $C_x = \frac{Width}{2}$,$C_y = \frac{Height}{2}$. The focal length $f_x = f_y = 0.8 \times \max(W, H)$. For the extrinsic matrix $[R|t]$, the direct linear transformation (DLT) method is utilized to build a linear projection $(u_i, v_i) = K([R|t] \cdot (x_i, y_i, z_i)$ between the 3D point $(x_i, y_i,\ z_i)$ and its corresponding image point(u_i, v_i). To deal with lens distortion, common models [14, 37] adopted polynomials to transform (u_i, v_i) into $D((u_i, v_i)\ ,\ K)$, where K are the D distortion parameters. The undistortion function D^{-1} can be computed using a radial distortion estimation [38].

4.4. Evaluation for calibration accuracy and projection optimization

To evaluate the effect of our proposed camera calibration method, we compared our method with several state-of-the-art calibration methods on the World Cup 2014 dataset, including TVcalib [36], CCBV-SN [10], NBJW [16] and [4] and Optimizing for accurate sports field registration [20]. Tab. 1 presents the comparison results using the IoU metric. The IoU (whole) score of our method achieves 92.1%, which supports accurate geometric alignment across regions less visible in the sports field. The IoU (whole) score provides that our model infers the positions of field markings beyond the visible region, effectively generalizing from incomplete visual information. In IoU (part), our method achieves a score comparable to TVCalib. Even under challenges, such as player occlusions and partial field visibility, our method consistently maintains precise alignment in terms of IoU (part) performance, as illustrated in Tab. 1. Compared to baseline methods, our model exhibits significantly reduced projection errors along both straight lines (e.g., sidelines and penalty box boundaries) and curved lines (e.g., center circle and penalty arcs), particularly in complex areas like the penalty box and center circle. By improving the IoU (part) and IoU (whole) evaluations, our method not only enhances global geometric accuracy across the entire field, but also achieves high precision in calibrating visible and unoccluded regions.

To validate the performance of the reprojection, we selected TVCalib [36], BroadTrack [29], and NBJW [16] as baseline methods using the SoccerNetV3 calibration dataset. Tab. 2 summarizes the performance under standard reprojection error below 5 pixels and below 10 pixels, and calibration recall. In detail, our approach achieves a JaC10 score of 86.9, significantly higher than TVCalib, BroadTrack, and NBJW in different camera perspectives and field conditions. Even with the tolerable error reduced to 5 pixels, the JaC5 score using our approach is 75.8, which is significantly higher than TVCalib, BroadTrack, and NBJW. Furthermore, Tab. 2 illustrates qualitative comparisons of

Approaches	IoU (Whole) Mean	IoU (Whole) Med.	IoU (Part) Mean	IoU (Part) Med.
CCBV-SN	79.8	81.7	88.5	92.3
[4]	89.4	93.8	94.5	96.1
TVcalib [36]	-	95.3	96.6	-
NBJW	89.4	93.0	94.4	96.9
[14]	89.8	92.9	95.1	96.7
Ours	**92.1**	**94.6**	**96.3**	**97.8**

Table 1. The homograph estimation on World Cup 2014 test dataset using IoU. The number of World Cup 2014 test datasets is 186 images. "–" indicates that the model does not report results for this metric.

reprojection precision, where our method exhibits minimal projection drift along both straight and curved field lines of the field. This improvement is particularly evident in different critical regions, such as the penalty box, center circle, and goal areas, where precise calibration is essential.

Approaches	JaC5 (↑)	JaC10 (↑)	CR (↑)
BroadTrack	75.3	86.8	69.8
TVCalib	52.9	73.4	66.5
NBJW	73.7	86.7	77.5
Ours	**75.8**	**86.9**	**78.3**

Table 2. Evaluating the reprojection error induced by camera parameters ϕ on SoccerNetV3-Calibration test dataset.

4.5. Optimized camera pose estimation

We estimate the camera pose using our method on the World Cup 2014 dataset and SoccerNetV3-calibration test dataset. Fig. 3 illustrates the optimized camera poses with point-cloud data in scene visualization. As shown in Fig. 3, the estimated camera positions are distributed around the field, with orientations properly aligned to key areas such as the goal and players. The point-cloud visualization further verifies the accuracy of the estimated poses, demonstrating that our method effectively recovers the spatial placement of the cameras. Our approach successfully optimizes camera poses to align with the real-world scene structure while enhancing the representation of scene geometry.

4.6. Ablation study

In the ablation experiment, we systematically evaluate the impact of different components in our camera calibration framework. To test the impact of our modifications, we progressively integrated key components into the baseline model. Using the PixelNeRF as a baseline, we incorporated our attention module to verify its effect on enhancing alignment performance. Otherwise, we observe that

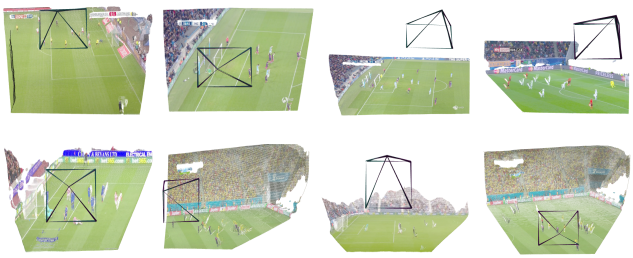


Figure 3. Optimized camera pose estimation with visualization.

optimizing camera parameters solely with the original loss function results in poor convergence due to insufficient supervision and lack of geometric constraints. We established the Pixel-NeRF-based network architecture as the baseline to evaluate the effectiveness of our proposed camera calibration method. To test the impact of our modifications, we progressively integrated key components into the baseline model. First, we incorporated our attention module into the PixelNeRF to verify its effect on enhancing alignment performance. Additionally, we replaced the original PixelNeRF loss function with our joint optimization strategy to analyze its contribution to improve the calibration accuracy. To improve the calibration accuracy, we introduce pose normalization loss, which enforces a structured mapping between 2D image features and 3D camera parameters, leading to a remarkable improvement in reprojection accuracy. Furthermore, we adopt a joint constraint strategy to resolve depth and perspective ambiguities, further enhancing the effect. These results in the Tab. 3 show that the importance of additional constraints in achieving more accurate and stable camera calibration. The results highlight the necessity of structured supervision and geometric constraints in enhancing the camera pose. Our proposed modifications significantly improve convergence and accuracy, demonstrating the effectiveness of our approach in tackling the challenges of camera calibration for the sports field.

Approaches	**JaC5**	**JaC10**	**CR**
PixelNeRF	60.2	66.7	64.5
Semantic-aware attention	68.2	75.2	71.9
PixelNeRF + joint constraint strategy	71.2	82.7	73.0
Semantic-aware attention + pose normalization loss	75.2	86.7	78.0
Semantic-aware attention + joint constraint strategy	**75.8**	**86.9**	**78.3**

Table 3. Ablation study by camera parameters ϕ on SoccerNetV3-Calibration test dataset.

Fig. 4 illustrates the qualitative results using our method. The reprojection error in our method is derived from key-

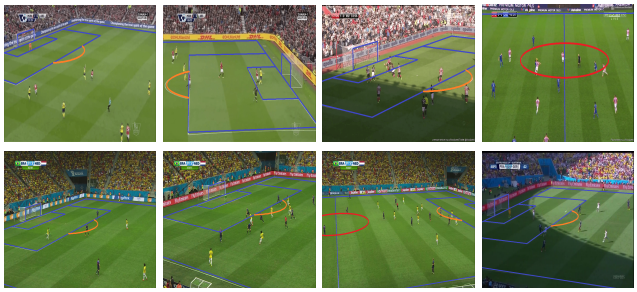


Figure 4. Evaluating the reprojection error induced by camera parameters ϕ on SoccerNetV3-Calibration test dataset and World Cup 2014 test dataset.

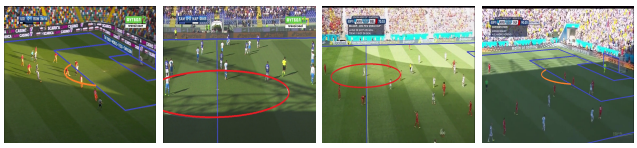


Figure 5. The performance of reprojection error with inconsistent brightness.

point discrepancies, indicating that, while line-based refinement might lead to a higher keypoint reprojection error, it contributes to improved overall calibration. Fig. 5 highlights the performance of reprojection with inconsistent brightness. Our method maintains robust reprojection performance under varying illumination conditions, especially in cases where lighting changes and occlusion cause inconsistencies in the brightness of the same image.

5. Conclusion

In this paper, we introduce a semantic-aware attention mechanism into positional encoding to enhance spatial detail representation in NeRF-based camera calibration. By integrating scene semantics, our method dynamically refines spatial representations, improving alignment between image pixels and the 3D scene structure while effectively handling occlusions and frequent viewpoint changes in sports environments. Additionally, we propose a joint optimization strategy that combines pose normalization, position constraints, and planarity constraints, significantly enhancing the robustness of camera parameter estimation. This approach reduces reliance on frame-to-frame motion consistency and leverages depth information and neural rendering to improve performance under occlusions. Notably, our method enables NeRF-based camera calibration using only image observations, without requiring ground-truth 3D player coordinates. This capability makes it highly suitable for real-world sports broadcasts that lack 3D supervision. For future work, we aim to further exploit correlations between the same features across multiple frames to calibrate challenging camera shots with few or no visible landmarks.

References

[1] 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 (CVPR)*, pages 5109–5118, 2023. 1

[2] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: optimising neural radiance field with no pose prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4160–4169, 2023. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229, 2020. 5

[4] Jianhui Chen and James J. Little. Sports camera calibration via synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–8, 2019. 1, 2, 7

[5] Jianhui Chen, Fangrui Zhu, and James J Little. A two-point method for ptz camera calibration in sports. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 287–295, 2018. 1

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 3

[7] Yi-Nan Chen, Hang Dai, and Yong Ding. Pseudo-stereo for monocular 3d object detection in autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 887–897, 2022. 1

[8] Zezhou Cheng, Carlos Esteves, Varun Jampani, Abhishek Kar, Subhransu Maji, and Ameesh Makadia. Lu-nerf: scene and pose estimation by synchronizing local unposed nerfs. In *IEEE International Conference on Computer Vision (ICCV)*, pages 18312–18321, 2023. 5

[9] 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)*, pages 3523–3530, 2022. 1, 2

[10] Anthony Cioppa, Adrien Deliege, Floriane Magera, Silvio Giancola, 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/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4537–4546, 2021. 7

[11] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up soccernet with multi-view spatial localization and re-identification. *Scientific Data*, 9(1):355, 2022. 6

[12] Paul Johannes Claasen and Johan Pieter de Villiers. Video-based sequential bayesian homography estimation for soccer field registration. *Expert Systems with Applications*, 252: 124156, 2024. 6

[13] Matteo Dunnhofer, Luca Sordi, and Christian Micheloni. Visualizing skiers' trajectories in monocular videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5188–5198, 2023. 1

[14] Olivier D Faugeras, Q T Luong, and Stephen J Maybank. Camera self-calibration: theory and experiments. In *European Conference on Computer Vision (ECCV)*, pages 321–334, 1992. 7

[15] 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)*, pages 1711–1721, 2018. 6

[16] 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)*, pages 3325–3334, 2024. 1, 2, 7

[17] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 6

[18] Xiankang He, Guangkai Xu, Bo Zhang, Hao Chen, Ying Cui, and Dongyan Guo. Diffcalib: reformulating monocular camera calibration as diffusion-based dense incident map generation. *arXiv preprint arXiv:2405.15619*, 2024. 2

[19] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun. Sports field localization via deep structured models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5212–5220, 2017. 1, 6

[20] 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/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 201–210, 2020. 2, 7

[21] 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)*, pages 17307–17316, 2023. 2

[22] S Sathiya Keerthi and Chih-Jen Lin. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation*, 15(7):1667–1689, 2003. 5

[23] Jinwoo Lee, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, Minhyuk Sung, and Junho Kim. Ctrl-c: camera calibration transformer with line-classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 16228–16237, 2021. 2

[24] Shiqi Li, Chi Xu, and Ming Xie. A robust o (n) solution to the perspective-n-point problem. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1444–1450, 2012. 4

[25] Jiawei Liu, Changkun Ye, Ruikai Cui, and Nick Barnes. Self-calibrating vicinal risk minimisation for model calibration.

In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3335–3345, 2024. 1

[26] Paul Liu and Jui-Hsien Wang. Monotrack: shuttle trajectory reconstruction from monocular badminton video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5188–5198, 2022. 1

[27] Yaroslava Lochman, Oles Dobosevych, Rostyslav Hryniv, and James Pritts. Minimal solvers for single-view lens-distorted camera auto-calibration. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2887–2896, 2021. 2

[28] 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 (CVPR)*, pages 3335–3346, 2024. 6

[29] Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. Broadtrack: broadcast camera tracking for soccer. *arXiv preprint arXiv:2412.01721*, 2024. 7

[30] Zhangkai Ni, Peiqi Yang, Wenhan Yang, Hanli Wang, Lin Ma, and Sam Kwong. Colnerf: collaboration for generalizable sparse input neural radiance field. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 4325–4333, 2024. 3

[31] Keunhong Park, Philipp Henzler, Ben Mildenhall, Jonathan T Barron, and Ricardo Martin-Brualla. Camp: camera preconditioning for neural radiance fields. In *Acm Transactions on Graphics (TOG)*, pages 1–11, 2023. 3

[32] Jens Puwein, Remo Ziegler, Julia Vogel, and Marc Pollefeys. Robust multi-view camera calibration for wide-baseline camera networks. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 321–328, 2022. 2

[33] Feng Shi, Paul Marchwica, Juan Camilo Gamboa Higuera, Michael Jamieson, Mehrsan Javan, and Parthipan Siva. Self-supervised shape alignment for sports field registration. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 287–296, 2022. 1

[34] Xu Song, Hao Kang, Atsunori Moteki, Genta Suzuki, Yoshie Kobayashi, and Zhiming Tan. Mscc: multi-scale transformers for camera calibration. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3262–3271, 2024. 2

[35] Zhenggang Tang, Zhongzheng Ren, Xiaoming Zhao, Bowen Wen, Jonathan Tremblay, Stan Birchfield, and Alexander Schwing. Nerfdeformer: nerf transformation from a single view via 3d scene flows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10293–10303, 2024. 3

[36] 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)*, pages 1166–1175, 2023. 1, 2, 7

[37] Steffen Urban, Jens Leitloff, and Stefan Hinz. Improved wide-angle, fisheye and omnidirectional camera calibration. *ISPRS Journal of Photogrammetry and Remote Sensing*, 108:72–79, 2015. 7

[38] 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)*, pages 1–20, 2024. 2, 7

[39] Shuzhe Wang, Vincent Leroya, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: geometric 3d vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 1

[40] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21875–21911, 2025. 5

[41] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. Inerf: inverting neural radiance fields for pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330, 2021. 3

[42] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. Pixelnerf: neural radiance fields from one or few images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 3