# UNIPITCH: PARAMETRIC RECONSTRUCTION OF SOCCER MATCHES FROM BROADCAST VIDEOS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reconstructing soccer matches from broadcast videos requires identifying players across changing views and recovering their positions on the pitch. Existing pipelines estimate player tracks and camera geometry separately, then combine their outputs through projection. This separation leaves reconstructed player positions as downstream outputs, limiting their use as supervision for camera estimation. We present **UniPitch**, a framework for parametric reconstruction of soccer matches that represents persistent player identities, evolving pitch positions, and camera states within a common formulation. A differentiable geometric mapping allows two-dimensional player position errors to propagate back to camera parameters, enabling reconstruction supervision to guide the geometry itself. UniPitch learns persistent identity associations through appearance memory that outlives short-term tracks, models camera evolution using visual and temporal evidence, and refines player positions with identity-linked histories and geometric context. On SoccerNet-GSR, UniPitch achieves 73.30% P-HOTA, improving over the best decoupled baseline by 4.80%, with gains in both pitch-space matching and identity association. Ablation studies support the contributions of persistent identity memory, temporal camera modeling, and differentiable trajectory refinement.

## 1 INTRODUCTION

Analyzing a soccer match requires following how players move and how their spatial arrangement changes over time. Recovering this information from broadcast videos would provide a foundation for tactical analysis (Rao et al., 2026), but the view of the match changes as the camera pans and zooms, and players repeatedly enter and leave the frame. To describe the match consistently across these changing views, we need to determine which player each observation belongs to and where that player is on the pitch. This is a central goal of Game State Reconstruction (GSR) (Somers et al., 2024), which recovers structured match states from video. We focus on reconstructing persistent player identities and their evolving spatial positions, together with the camera geometry that relates the pitch to the observed images.

Existing GSR methods (Somers et al., 2024; Oo et al., 2026; Golovkin et al., 2025) typically divide this goal into player tracking and camera calibration. A tracker detects and associates players in image space, while a calibration module estimates the mapping from the image to the pitch. Projecting the tracked observations through this mapping produces pitch-space trajectories, which can be further refined using roles, jersey numbers, team affiliation, or motion constraints. When tracking and calibration are optimized separately, however, the reconstructed trajectories remain downstream outputs: errors in player positions on the pitch cannot directly supervise the camera estimator through the projection step. This separation is consequential because even well-tracked image observations can produce inaccurate pitch positions under an erroneous camera estimate. Meanwhile, interruptions in identity association fragment the observations needed to describe the same player over time. Reconstructing the match therefore calls for both persistent identity association and a geometric mapping that can learn from the player positions it reconstructs.

We revisit this problem as *parametric reconstruction of soccer matches*. A match is represented by persistent player identities and their evolving two-dimensional positions on a canonical pitch, together with camera parameters that connect these positions to image observations. Learning identity association connects observations of the same player across time, while learning the camera ge-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ometry places those observations in a common spatial reference. Crucially, this explicit geometric relationship also makes the recovered player positions a source of supervision for the camera model. Through a differentiable mapping, two-dimensional position errors propagate back to camera parameter estimation: camera geometry determines the reconstructed trajectories, and trajectory errors in turn guide the estimation of that geometry. This formulation brings identity, position, and camera estimation into an end-to-end learning framework for recovering the match state.

Based on this formulation, we propose **UniPitch**, an end-to-end framework for parametric reconstruction of soccer matches from broadcast videos. UniPitch learns identity association, player localization, and camera estimation with temporal representations at both the player and camera levels. **Persistent Slot Identity Memory (PSIM)** learns to associate current observations with historical appearance prototypes retained beyond short-term track lifetimes, allowing a player's identity to persist across departures from and returns to the camera view. These identities organize observations into player histories. **Temporal Camera State Modeling (TCSM)** learns camera evolution from visual observations and historical states, with player position supervision reaching the camera estimator through differentiable projection. **Shared-Feature Trajectory Refinement (SFTR)** combines the identity-linked histories with shared visual features and geometric context to refine pitch positions. Following staged training, the framework is trained to reconstruct player identities and spatial states while learning the camera geometry that connects them to the video. Experiments on the SoccerNet-GSR (Somers et al., 2024) dataset show that UniPitch consistently improves over decoupled tracking-and-calibration baselines in pitch-space trajectory reconstruction, with gains in both localization and temporal association. Despite not relying on semantic identity cues for refinement, UniPitch remains competitive with rule-augmented methods. Extensive ablation studies further validate the complementary contributions of each component and player position supervision.

Our main contributions are summarized as follows:

- We formulate soccer match reconstruction through a parametric representation of persistent player identities, evolving pitch positions, and camera states. A differentiable geometric mapping enables two-dimensional player position supervision to guide camera parameter estimation.
- We develop **UniPitch**, an end-to-end framework that learns identity association and spatial reconstruction through persistent identity memory, temporal camera modeling, and shared-feature trajectory refinement.
- Extensive experiments demonstrate the effectiveness of UniPitch for temporally consistent pitch-space reconstruction and validate the advantage of treating soccer match reconstruction as a unified parametric learning problem rather than a sequence of decoupled tasks.

## 2 RELATED WORK

### 2.1 MULTI-OBJECT TRACKING

Multi-object tracking methods can be broadly divided into tracking-by-detection approaches and end-to-end trackers. Tracking-by-detection methods first obtain frame-wise detections and then associate them across time using motion and appearance cues. SORT (Bewley et al., 2016) and OC-SORT (Cao et al., 2023) primarily rely on motion modeling, while DeepSORT (Wojke et al., 2017) and QDTrack (Pang et al., 2021) further incorporate learned appearance features for association. ByteTrack (Zhang et al., 2022) improves trajectory completeness by retaining low-confidence detections during matching. Several methods further adapt this paradigm to sports scenarios. SoccerNet-Tracking (Cioppa et al., 2022) and SoccerTrack (Scott et al., 2022) provide soccer-specific tracking benchmarks. MixSort (Cui et al., 2023) and Deep-EIoU (Huang et al., 2024) combine appearance and motion cues to improve the association between visually similar athletes. GTA (Sun et al., 2024) and GTATrack (Jian et al., 2025) further perform global tracklet association to reconnect fragmented trajectories. Meanwhile, SportsSUSHI (Koshkina & Elder, 2025) incorporates jersey numbers, team identities, and field coordinates into hierarchical association. Although effective, these approaches still separate detection from association and rely on explicitly designed matching or tracklet management strategies.

Recent end-to-end trackers instead learn detection and association together, with many methods building upon DETR (Carion et al., 2020) and its query-based object representation. TrackFormer (Meinhardt et al., 2022) and MOTR (Zeng et al., 2022) propagate track queries across frames to maintain object identities, while MOTRv2 (Zhang et al., 2023) incorporates pretrained detection priors. MeMOT (Cai et al., 2022) and MeMOTR (Gao & Wang, 2023) augment track representations with memory to improve temporal association. SambaMOTR (Segu et al., 2025) employs synchronized state-space memories to model long-range dependencies within tracklets as well as interactions across different tracklets. MOTIP (Gao et al., 2025) reformulates association as ID prediction by inferring the identities of current detections from historical trajectories, and FDTA (Shao et al., 2026) further enhances the discriminability of DETR object embeddings for association. Despite these advances, identity association is typically conditioned on finite track or trajectory histories, making long-term re-entry difficult once earlier observations are no longer retained. This limitation is particularly relevant for broadcast soccer, where players frequently leave the camera view for extended periods. In UniPitch, PSIM maintains slot-indexed appearance memory independently of short-term track states, allowing persistent identities to organize player observations into temporally consistent pitch-space trajectories.

### 2.2 Sports Field Registration

The sports field registration task aims to estimate the geometric mapping between the broadcast view and the pitch view. To achieve this, existing methods explore synthetic-pose retrieval (Chen & Little, 2019), learned registration errors (Jiang et al., 2020), dense or keypoint correspondences (Nie et al., 2021; Chu et al., 2022), and self-supervised shape alignment (Shi et al., 2022). TVCalib (Theiner & Ewerth, 2023) estimates camera parameters by minimizing reprojection errors of field markings under a 3D camera model. NBJW (Gutiérrez-Pérez & Agudo, 2024) constructs geometry-derived field keypoints and applies a lightweight DLT-based calibration procedure. PnLCalib (Gutiérrez-Pérez & Agudo, 2026) further combines point and line constraints to improve calibration robustness when field observations are incomplete. BHITK (Claasen & De Villiers, 2024) models temporal homography uncertainty through Bayesian filtering. Most existing approaches primarily estimate calibration results from field evidence, while BroadTrack (Magera et al., 2025) additionally models camera motion over time. UniPitch instead couples temporal camera modeling with player state reconstruction. Through a differentiable image-to-pitch mapping, player pitch-position errors directly supervise TCSM, aligning camera estimation with the trajectory reconstruction objective.

### 2.3 Game State Reconstruction

Game State Reconstruction integrates tracking, field registration, and athlete identification to recover structured match states in the pitch coordinate system. Prior to the formalization of GSR, PRTReID (Mansourian et al., 2023) explored joint athlete re-identification, team affiliation, and role classification for sports visual tracking. SoccerNet-GSR (Somers et al., 2024) subsequently formalizes this task and introduces a modular baseline. Golovkin et al. (2025) and Broadcast2Pitch (Oo et al., 2026) improve modular reconstruction through stronger vision components and identity-aware refinement. SoccerMaster (Yang et al., 2026) takes a different direction by learning a soccer-specific shared visual backbone and adapting the learned representation to multiple downstream tasks through task-specific prediction heads. In contrast to modular GSR pipelines and general-purpose multi-task representation learning, UniPitch formulates persistent player identities, pitch positions, and camera states within a common parametric reconstruction framework.

## 3 Method

### 3.1 Overview

Given a broadcast video $\mathcal{I} = \{I_t\}_{t=1}^{T}$, UniPitch reconstructs the visible players in a canonical pitch coordinate system while preserving their temporal identities. The output at frame $t$ is

$$\mathcal{Y}_t = \{(s_{t,i}, \mathbf{p}_{t,i})\}_{i=1}^{N_t}, \tag{1}$$

where $s_{t,i}$ denotes the predicted identity slot and $\mathbf{p}_{t,i} \in \mathbb{R}^2$ is the corresponding pitch position.

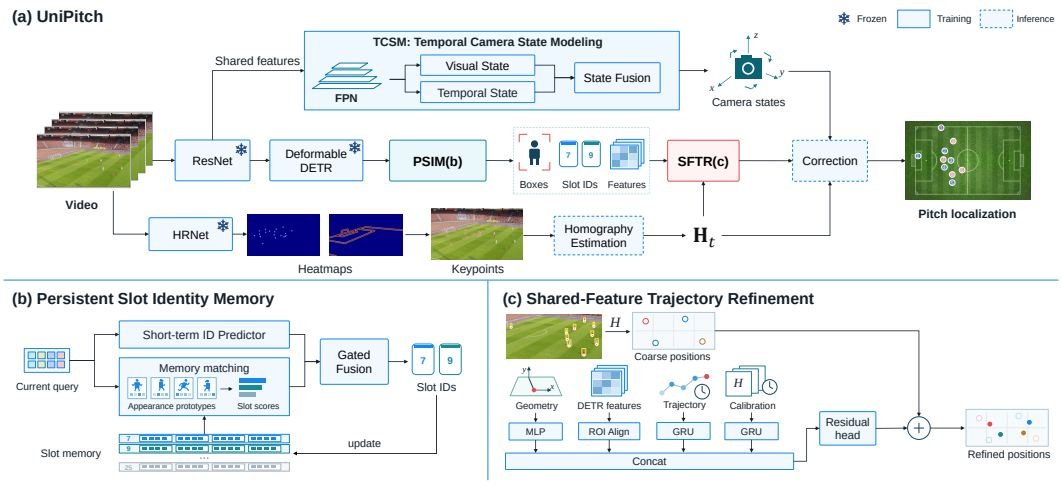


Figure 1: Overview of **UniPitch**, which integrates persistent identity association, temporal camera modeling, and trajectory refinement for pitch-space player trajectory reconstruction.

As illustrated in Figure 1, the input frames are encoded by a shared ResNet (He et al., 2016) backbone, whose features are used by Deformable DETR (Zhu et al., 2021) for player detection and by TCSM for temporal camera modeling. The resulting player queries are passed to PSIM, which augments short-term identity prediction with persistent appearance memory to produce temporally consistent slot identities. In parallel, the HRNet-based predictor adopted from PnLCalib (Gutiérrez-Pérez & Agudo, 2026) estimates pitch keypoints and field lines for frame-wise homography estimation. The homographies project detected players onto the pitch to obtain coarse positions, which are refined by SFTR with identity-linked histories and geometric context. During inference, TCSM models camera evolution from shared visual features and historical states, and is used to support calibration when PnLCalib becomes unreliable.

### 3.2 Persistent Slot Identity Memory

Existing ID-prediction trackers primarily associate current observations using a finite trajectory context. When a player remains outside the camera view for a long period, earlier observations may no longer be available for identity prediction. As shown in Figure 1(b), we introduce **Persistent Slot Identity Memory (PSIM)** to preserve identity information beyond short-term track states. PSIM maintains a persistent memory bank $\mathbf{M} \in \mathbb{R}^{S\times K\times D}$, where $S$ is the number of identity slots, $K$ is the number of appearance prototypes maintained for each slot, and $D$ is the feature dimension. Multiple prototypes are retained to account for appearance variations caused by viewpoint, pose, and occlusion rather than compressing all historical observations into a single representation.

Given a current player query $\mathbf{q}_i$, its affinity to slot $s$ is determined by the most similar valid prototype,

$$m_{i,s} = \max_k \cos\left(\phi(\mathbf{q}_i), \psi(\mathbf{M}_{s,k})\right), \tag{2}$$

where $\phi(\cdot)$ and $\psi(\cdot)$ denote two learnable MLPs that project the current query and memory prototypes into a common matching space, respectively, and $\cos(\cdot,\cdot)$ denotes cosine similarity.

PSIM augments the short-term identity prediction of MOTIP (Gao et al., 2025) with persistent appearance evidence. An adaptive gate $g_{i,s}$ is predicted from the base identity probability, memory affinity, and normalized time elapsed since slot $s$ was last observed. It adaptively controls the contribution of persistent memory according to the reliability of short-term association and historical appearance evidence. The final identity logit is obtained through residual fusion,

$$\ell_{i,s} = \ell^b_{i,s} + \lambda g_{i,s} m_{i,s}, \tag{3}$$

where $\ell^b_{i,s}$ denotes the base MOTIP logit and $\lambda > 0$ is a learnable scaling factor. This formulation preserves the short-term association prior while adaptively incorporating long-term appearance evidence.

After slot assignment, accepted observations update the corresponding appearance prototypes using exponential moving average,

$$\mathbf{M}_{s,k} \leftarrow \mu \mathbf{M}_{s,k} + (1-\mu)\mathbf{q}_i, \tag{4}$$

where $\mu \in [0, 1)$ denotes the EMA momentum and $k$ indexes the selected prototype of slot $s$. The persistent slot memory is maintained independently of short-term track states, allowing historical identity information to remain available after a track becomes inactive and to support association when the player re-enters the view.

### 3.3 Differentiable Pitch-Space Localization

Frame-wise calibration estimates camera geometry from the visual evidence available in each individual frame, and can therefore become unreliable when field markings are sparse, occluded, or affected by rapid camera motion. Moreover, even when a valid homography is obtained, errors in player detection or calibration can be amplified after projection onto the pitch. We address these two sources of localization error at complementary levels. **Temporal Camera State Modeling (TCSM)** models the temporal evolution of the broadcast camera to improve camera estimation beyond isolated frames, while **Shared-Feature Trajectory Refinement (SFTR)** corrects the resulting player locations using multimodal context. Both modules are optimized toward pitch-space player localization, with their training objectives detailed in Section 3.4.

**Temporal Camera State Modeling.** At the camera level, TCSM augments frame-wise calibration with a temporally evolving camera representation. We represent the camera state as

$$\mathbf{c}_t = [\theta_{\text{pan}}, \theta_{\text{tilt}}, \theta_{\text{roll}}, \eta_f, C_x, C_y, C_z] \in \mathbb{R}^7, \tag{5}$$

where the first three dimensions describe camera orientation, $\eta_f$ denotes the log focal-length ratio, and $(C_x, C_y, C_z)$ represents the camera center.

As illustrated in Figure 1(a), multi-scale ResNet features are aggregated through an FPN and spatial attention pooling to obtain a compact frame representation $\mathbf{f}_t$. From the current frame, TCSM predicts an instantaneous camera estimate $\mathbf{c}_t^v$, referred to as the *visual state*. In parallel, a GRU models camera evolution over a causal temporal window of $L$ frames. Based on the preceding $L-1$ frames and the current visual representation, it predicts the camera-state velocity $\mathbf{v}_t$. The resulting *temporal state* is given by

$$\mathbf{c}_t^p = \mathbf{c}_{t-1} + \mathbf{v}_t \Delta\tau_t, \tag{6}$$

where $\Delta\tau_t = \tau_t - \tau_{t-1}$ denotes the time interval between two consecutive frames. The visual state and temporal state are then adaptively fused as

$$\mathbf{c}_t = \mathbf{c}_t^p + \boldsymbol{\alpha}_t \odot (\mathbf{c}_t^v - \mathbf{c}_t^p), \tag{7}$$

where $\boldsymbol{\alpha}_t \in [0, 1]^7$ is a parameter-wise fusion gate. The fused camera state provides temporal camera information to improve player localization in intervals when frame-wise calibration is unreliable, as detailed in Section 3.4.

**Shared-Feature Trajectory Refinement.** At the player level, SFTR uses the frame-wise homography as a geometric initialization and learns to refine the projected player locations. Given the player bounding-box bottom center $\mathbf{u}_{t,i}$ predicted by Deformable DETR and the estimated homography $\mathbf{H}_t$, the coarse pitch position is obtained as $\mathbf{p}_{t,i}^c = \pi(\mathbf{H}_t \tilde{\mathbf{u}}_{t,i})$, where $\tilde{\mathbf{u}}_{t,i}$ denotes the homogeneous coordinate of $\mathbf{u}_{t,i}$ and $\pi(\cdot)$ denotes the homogeneous projection operation.

Starting from the coarse pitch position $\mathbf{p}_{t,i}^c$, **SFTR** predicts a residual correction by combining *current* and *historical* information. For the *current* frame, the geometry branch encodes the player detection and its image-to-pitch projection geometry using an MLP, while the visual branch extracts player-specific features from the shared multi-scale Deformable DETR encoder maps using ROI Align (He et al., 2017), followed by spatial pooling and a learnable projection. For *historical* context, the trajectory branch encodes recent coarse positions associated with the same PSIM slot ID using a GRU, while the calibration branch employs another GRU to model the recent evolution of frame-wise homographies. The four branch features are concatenated into $\mathbf{z}_{t,i}$ and fed into an MLP-based residual prediction head. The refined pitch position $\mathbf{p}_{t,i}^r$ is given by

$$\Delta\mathbf{p}_{t,i} = r_{\max} \tanh(\text{MLP}_{\text{ref}}(\mathbf{z}_{t,i})), \qquad \mathbf{p}_{t,i}^r = \mathbf{p}_{t,i}^c + \Delta\mathbf{p}_{t,i}, \tag{8}$$

where $\mathbf{z}_{t,i}$ denotes the concatenated representation of the four branches, $\Delta\mathbf{p}_{t,i}$ is the predicted position residual, and $r_{\max}$ bounds the correction along each pitch coordinate.

### 3.4 TRAINING AND INFERENCE

**Multi-stage training.** We train UniPitch in three stages. In Stage I, we initialize Deformable DETR with the weights pretrained on SportsMOT (Cui et al., 2023), and jointly train the detector and short-term slot ID predictor. The detector follows the MOTIP (Gao et al., 2025) detection training objectives, while slot prediction is supervised by a cross-entropy loss

$$\mathcal{L}_{\text{slot}} = \text{CE}(\boldsymbol{\ell}, y), \tag{9}$$

where $\boldsymbol{\ell}$ denotes the predicted slot logits and $y$ is the corresponding slot label.

In Stage II, we freeze Deformable DETR and the short-term ID predictor, and optimize only the remaining PSIM parameters using the same slot classification objective. PSIM is trained on two temporally separated windows from the same video sequence, denoted as *source* and *target* windows. The *source* window establishes player-to-slot assignments and initializes the corresponding appearance memories. For players that reappear in the *target* window, their source-assigned slots are used as supervision, encouraging PSIM to recover the same identities across the temporal gap.

In Stage III, all multi-object tracking parameters are frozen, and TCSM and SFTR are trained for pitch-space localization. Both modules are supervised by the same localization objective

$$\mathcal{L}_{\text{loc}} = \rho\left(\hat{p}^{x}_{t,i} - p^{\star,x}_{t,i}\right) + \rho\left(\hat{p}^{y}_{t,i} - p^{\star,y}_{t,i}\right), \tag{10}$$

where $t$ and $i$ denote the frame and player indices, respectively; $\hat{\mathbf{p}}_{t,i} = (\hat{p}^{x}_{t,i}, \hat{p}^{y}_{t,i})$ and $\mathbf{p}^{\star}_{t,i} = (p^{\star,x}_{t,i}, p^{\star,y}_{t,i})$ denote the predicted and ground-truth pitch positions, respectively. The Smooth L1 penalty $\rho(\cdot)$ is defined as

$$\rho(e) = \begin{cases} \dfrac{e^2}{2\beta}, & |e| < \beta, \\ |e| - \dfrac{\beta}{2}, & \text{otherwise}, \end{cases} \tag{11}$$

where $e$ is the difference between the predicted and ground-truth position along one pitch coordinate axis, and $\beta$ denotes the transition threshold between the quadratic and linear regions. For TCSM, the predicted pitch position is obtained as $\hat{\mathbf{p}}_{t,i} = \pi(\mathbf{G}_t \tilde{\mathbf{u}}^{\star}_{t,i})$, where $\tilde{\mathbf{u}}^{\star}_{t,i}$ denotes the homogeneous coordinate of the ground-truth bounding-box bottom center, and $\mathbf{G}_t$ is the image-to-pitch homography matrix derived from the predicted camera state $\mathbf{c}_t$. For SFTR, $\hat{\mathbf{p}}_{t,i} = \mathbf{p}^{r}_{t,i}$, where $\mathbf{p}^{r}_{t,i}$ is the refined pitch position defined above. Thus, TCSM and SFTR share the same pitch-space localization supervision while obtaining their predictions through different paths.

**Inference.** During inference, PSIM first assigns persistent slot identities to the detected players. Then we utilize PnLCalib to estimate a frame-wise image-to-pitch homography, which is used to project the player bounding-box bottom centers onto the pitch, and SFTR further refines these projected positions. We next evaluate the reliability of the frame-wise calibration. A frame is considered unreliable when PnLCalib fails to produce a valid homography, repeatedly reuses a previous estimate, or yields projected player positions that deviate substantially from the SFTR-refined positions. For such frames, we do not directly use the absolute homography derived from TCSM. Instead, the inter-frame camera changes estimated by TCSM are used to propagate neighboring reliable PnLCalib homographies across the unreliable interval and reconstruct corrected homographies. The corrected homography yields an alternative pitch position for each player, which is fused with the SFTR-refined position using a weight determined from the SFTR refinement residuals of the current frame.

## 4 EXPERIMENTS

### 4.1 DATASET

We conduct experiments on the SoccerNet-GSR dataset (Somers et al., 2024), which provides synchronized annotations for player tracking, camera calibration, and pitch-space localization in broadcast soccer videos. The dataset contains 57 training, 58 validation, and 49 test clips, with each clip consisting of 750 frames. All frames are provided at a resolution of $1920 \times 1080$. Following the available annotations, pitch-space evaluation is performed only on frames with valid player and pitch labels, resulting in 33,995 evaluated frames from the 49 test sequences.

Table 1: Comparison of pitch-space trajectory reconstruction performance on the SoccerNet-GSR test set.

| Method | P-HOTA↑ | P-DetA↑ | P-AssA↑ | P-LocA↑ | P-IDF1↑ |
|---|---|---|---|---|---|
| *Decoupled* | | | | | |
| FDTA + TVCalib | 45.86 | 45.46 | 46.38 | 90.87 | 51.72 |
| MOTIP + TVCalib | 47.75 | 45.79 | 49.91 | 90.95 | 54.39 |
| FDTA + PnLCalib | 65.32 | 76.53 | 55.80 | 92.99 | 70.05 |
| MOTIP + PnLCalib | **68.50** | **77.11** | **60.90** | **93.07** | **73.99** |
| *Rule-Augmented* | | | | | |
| SN-Gamestate (Somers et al., 2024) | 43.65 | 43.42 | 44.00 | 90.70 | 46.85 |
| SoccerMaster (Yang et al., 2026) | 77.47 | 80.04 | **75.02** | 94.39 | **83.92** |
| KISTGSR (Oo et al., 2026) | **78.96** | **85.10** | 73.30 | **94.73** | 82.84 |
| *Parametric Modeling* | | | | | |
| **UniPitch (Ours)** | **73.30** | **82.67** | **65.04** | **93.36** | **79.32** |

### 4.2 Evaluation Metrics

We evaluate player trajectory reconstruction entirely in pitch space. Following the HOTA (Luiten et al., 2021) evaluation framework, we replace the standard bounding-box IoU similarity with a distance-based similarity between predicted and ground-truth player positions on the pitch, and denote the resulting protocol as Pitch-space HOTA (P-HOTA). Temporal identities are retained for association evaluation, while semantic attributes such as player role, team affiliation, and jersey number are excluded. We report P-HOTA together with P-DetA, P-AssA, P-LocA, and P-IDF1. P-DetA measures pitch-space detection and matching quality, P-AssA evaluates temporal identity association, P-LocA measures the localization quality of matched observations, and P-IDF1 provides an additional identity-consistency measure. Detailed definitions are provided in Appendix A.

### 4.3 Implementation Details

For player detection, we adopt a ResNet-50 based Deformable DETR (Zhu et al., 2021) initialized from the MOTIP (Gao et al., 2025) model pretrained on SportsMOT (Cui et al., 2023). The short-term slot ID predictor is trained for 13 epochs with 25 identity slots. To reduce redundant overlap, training windows are constructed with an 8-frame stride between consecutive start positions while preserving the original within-window temporal sampling. PSIM is then trained for 10 epochs with the detector and short-term ID predictor frozen. It maintains $K = 4$ appearance prototypes per slot with an EMA momentum of $\mu = 0.9$, and is trained using pairs of non-overlapping 60-frame source and target windows whose first frames are separated by 61, 120, or 240 frames. TCSM is trained for 20 epochs using 16-frame temporal windows. SFTR is also trained for 20 epochs with the upstream tracking and calibration components fixed, using eight-step trajectory and calibration histories. We set $r_{\max} = 8$ meters and the Smooth L1 transition threshold $\beta = 1$. All trainable stages use AdamW with a learning rate of $10^{-4}$ and weight decay of $5 \times 10^{-4}$ for the base tracker and $10^{-4}$ for the remaining modules. We conduct our experiments on 3 NVIDIA A100 80GB GPUs.

### 4.4 Quantitative Comparison

We compare three categories of methods in Table 1. *Decoupled* baselines independently combine multi-object tracking and camera calibration modules without trajectory-level coupling between them. We select MOTIP (Gao et al., 2025) and FDTA (Shao et al., 2026) as representative SOTA trackers, and pair them with TVCalib (Theiner & Ewerth, 2023) and PnLCalib (Gutiérrez-Pérez & Agudo, 2026) as camera calibration baselines. For a fair adaptation to SoccerNet-GSR, MOTIP and FDTA are initialized from their SportsMOT-pretrained weights and fine-tuned on the SoccerNet-GSR training split using the corresponding SportsMOT training hyperparameters. TVCalib and PnLCalib are used with their released soccer-specific weights without additional fine-tuning. For *Rule-Augmented* methods, we compare with SN-Gamestate (Somers et al., 2024), SoccerMaster (Yang et al., 2026), and KISTGSR (Oo et al., 2026). Specifically, the SoccerMaster result refers to the SoccerFactory data extraction pipeline introduced in the original paper. These meth-

Table 2: Ablation study of the proposed components on the SoccerNet-GSR test set.

| ID | PSIM | SFTR | TCSM | P-HOTA↑ | P-DetA↑ | P-AssA↑ | P-LocA↑ | P-IDF1↑ |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | 66.86 | 77.08 | 58.04 | 92.98 | 73.70 |
| 2 | ✓ | | | 68.99 | 77.34 | 61.57 | 92.96 | 76.05 |
| 3 | | ✓ | | 69.51 | 80.33 | 60.20 | 93.02 | 75.92 |
| 4 | ✓ | ✓ | | 71.61 | 80.51 | 63.73 | 92.99 | 78.33 |
| 5 | ✓ | ✓ | ✓ | **73.30** | **82.67** | **65.04** | **93.36** | **79.32** |

Table 3: Ablation study of pitch-position supervision loss on the SoccerNet-GSR test set.

| ID | Pitch-Position Supervision | | P-HOTA↑ | P-DetA↑ | P-AssA↑ | P-LocA↑ | P-IDF1↑ |
|---|---|---|---|---|---|---|---|
| | TCSM | SFTR | | | | | |
| 1 | | | 70.50 | 79.11 | 62.87 | 93.22 | 77.05 |
| 2 | | ✓ | 72.74 | 82.02 | 64.56 | 93.32 | 78.97 |
| 3 | ✓ | | 70.59 | 79.24 | 62.93 | 93.24 | 77.10 |
| 4 | ✓ | ✓ | **73.30** | **82.67** | **65.04** | **93.36** | **79.32** |

ods incorporate additional information such as player role, jersey number, and team or jersey-color cues to refine the reconstructed game state. In contrast, *Parametric Modeling* denotes our UniPitch framework, which learns persistent identity association and pitch-space localization without using these semantic identity cues for player association.

Among the decoupled baselines, MOTIP + PnLCalib performs best with a P-HOTA of 68.50%. UniPitch improves this result to 73.30%, corresponding to a gain of 4.80%. In particular, P-DetA increases from 77.11% to 82.67%, indicating that more players are correctly localized and matched in pitch space. P-AssA further improves from 60.90% to 65.04%, showing that persistent identity modeling strengthens temporal association, especially when players leave and later re-enter the camera view. The corresponding improvement in P-IDF1 from 73.99% to 79.32% further confirms the gain in identity consistency. Together, these results show that parametric modeling benefits both pitch-space reconstruction and long-term identity association compared with independently optimized tracking and calibration stages. We further provide qualitative comparisons with this decoupled baseline in Appendix B.

Rule-augmented methods achieve higher overall P-HOTA by exploiting additional semantic identity cues and task-specific refinement. Nevertheless, UniPitch attains a P-DetA of 82.67%, surpassing SoccerMaster despite not using role, jersey number, color, or team affiliation information. The remaining performance gap is mainly reflected in association, where such semantic cues provide additional identity evidence.

### 4.5 Ablation Studies

**Component ablation.** Table 2 evaluates the contribution of the three proposed components. The baseline (ID 1) combines multi-object tracking with the short-term slot predictor and pitch-space localization obtained from PnLCalib. Adding PSIM to this baseline improves P-HOTA from 66.86% to 68.99% and P-AssA from 58.04% to 61.57%, with a corresponding P-IDF1 gain of 2.35%, confirming its effectiveness in improving temporal identity association. SFTR mainly benefits pitch-space reconstruction, increasing P-DetA from 77.08% to 80.33% and P-HOTA to 69.51% when used alone. Combining PSIM and SFTR further raises P-HOTA to 71.61%, indicating that persistent identity modeling and trajectory refinement provide complementary gains. Finally, incorporating TCSM improves P-HOTA to 73.30% and P-DetA to 82.67%, yielding an overall improvement of 6.44% P-HOTA over the baseline. These results show that the three components contribute complementary improvements to both identity consistency and pitch-space trajectory reconstruction.

**Pitch-position supervision.** Table 3 studies the effect of pitch-position supervision in TCSM and SFTR. Without pitch-position supervision in either module, the model obtains 70.50% P-HOTA. Applying the supervision to SFTR alone substantially improves P-HOTA to 72.74% and P-DetA from 79.11% to 82.02%, showing that direct pitch-space supervision is important for learning effective

Table 4: Ablation study of the inputs of SFTR on the SoccerNet-GSR test set. Traj., Geom., Calib., and DETR denote the trajectory, geometry, calibration, and DETR-feature branches, respectively.

| ID | Traj. | Geom. | Calib. | DETR | P-HOTA↑ | P-DetA↑ | P-AssA↑ | P-LocA↑ | P-IDF1↑ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | ✓ | ✓ | ✓ | 67.54 | 74.92 | 60.96 | 92.02 | 74.96 |
| 2 | ✓ | | ✓ | ✓ | 72.04 | 80.93 | 64.16 | **93.39** | 78.31 |
| 3 | ✓ | ✓ | | ✓ | 73.18 | 82.51 | 64.94 | 93.21 | 79.26 |
| 4 | ✓ | ✓ | ✓ | | 70.07 | 78.11 | 62.91 | 93.25 | 76.36 |
| 5 | ✓ | ✓ | ✓ | ✓ | **73.30** | **82.67** | **65.04** | 93.36 | **79.32** |

trajectory refinement. In comparison, applying the supervision only to TCSM provides a smaller improvement when SFTR is not supervised. The best performance is achieved when both modules receive pitch-position supervision, reaching 73.30% P-HOTA, 82.67% P-DetA, and 65.04% P-AssA. In particular, adding TCSM supervision on top of the supervised SFTR further improves P-HOTA by 0.56%, indicating that trajectory-level supervision can provide additional guidance to temporal camera modeling.

**SFTR input ablation.** Table 4 examines the contribution of the four input branches of SFTR by removing one branch at a time. The trajectory branch is the most influential, as removing it decreases P-HOTA from 73.30% to 67.54%, P-DetA from 82.67% to 74.92%, and P-IDF1 from 79.32% to 74.96%, highlighting the importance of identity-linked trajectory history for refinement. The DETR feature branch also contributes substantially, with its removal reducing P-HOTA by 3.23% and P-IDF1 by 2.96%, demonstrating the value of shared visual information. Removing the geometry branch causes a more moderate P-HOTA drop of 1.26%, while the calibration history branch produces smaller but consistent gains across the main reconstruction metrics. The complete setting achieves the best P-HOTA, P-DetA, P-AssA, and P-IDF1, demonstrating the benefit of jointly incorporating trajectory, geometric, calibration, and visual information in SFTR.

## 5 Limitations

Our current framework still has two limitations. First, although TCSM is learned during training, it is mainly used at inference time to assist frame-wise keypoint-based calibration when the latter becomes unreliable. This limits the extent to which the two camera estimation paradigms can complement each other. Future work will explore a learnable strategy for jointly optimizing geometric calibration and temporal camera modeling. Second, explicit identity cues such as jersey number, team affiliation, and player role can substantially benefit association, but are not incorporated into the current parametric formulation because they typically rely on task-specific engineering. In the future, we plan to investigate how such cues can be integrated into an end-to-end learning pipeline.

## 6 Conclusion

In this paper, we present UniPitch, a framework for parametric reconstruction of soccer matches from broadcast videos. UniPitch preserves player identities beyond short-term track lifetimes, uses reconstructed player positions to guide temporal camera estimation through differentiable projection, and further refines projected trajectories with visual and historical context. The experimental results indicate that reliable pitch-space reconstruction depends not only on stronger tracking or calibration in isolation, but also on how identity and geometry inform each other over time, highlighting the value of trajectory-centered modeling for soccer match reconstruction.

## AI Use Statement

In this work, we used generative AI tools to aid language polishing, literature retrieval and discovery, and limited research execution. We did not use generative AI tools to draft sections of the paper, generate synthetic datasets, or prove mathematical claims. For writing assistance, AI tools were used to improve the clarity, conciseness, and grammatical quality of the manuscript. For literature retrieval,

AI tools assisted in identifying potentially relevant prior work; the authors manually examined every cited paper and verified the relevance, content, citation, and bibliographic information of all references included in the manuscript. For research execution, AI tools assisted with the implementation of selected parts of the codebase; all AI-assisted code was manually inspected, tested, and validated by the authors before being used in the experiments. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including its methodology, implementation, experiments, analysis, claims, and text.

## REFERENCES

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pp. 3464–3468. Ieee, 2016.

Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8080–8090. IEEE, 2022.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9686–9696. IEEE, 2023.

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*, pp. 213–229. Springer, 2020.

Jianhui Chen and James J Little. Sports camera calibration via synthetic data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2497–2504. IEEE, 2019.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3522–3529. IEEE, 2022.

Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3490–3501. IEEE, 2022.

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.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9887–9897. IEEE, 2023.

Ruopeng Gao and Limin Wang. Memotr: Long-term memory-augmented transformer for multi-object tracking. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9867–9876. IEEE, 2023.

Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as id prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27883–27893. IEEE, 2025.

Vladimir Golovkin, Nikolay Nemtsev, Vasyl Shandyba, Oleg Udin, Nikita Kasatkin, Pavel Kononov, Anton Afanasiev, Sergey Ulasen, and Andrei Boiarov. From broadcast to minimap: Achieving state-of-the-art soccernet game state reconstruction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6010–6028. IEEE, 2025.

Marc Gutiérrez-Pérez and Antonio Agudo. No bells, just whistles: Sports field registration by leveraging geometric properties. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3325–3334. IEEE, 2024.

Marc Gutiérrez-Pérez and Antonio Agudo. Pnlcalib: Sports field registration via points and lines optimization. *Computer Vision and Image Understanding*, pp. 104712, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Hsiang-Wei Huang, Cheng-Yen Yang, Jiacheng Sun, Pyong-Kun Kim, Kwang-Ju Kim, Kyoungoh Lee, Chung-I Huang, and Jenq-Neng Hwang. Iterative scale-up expansioniou and deep features association for multi-object tracking in sports. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 163–172. IEEE, 2024.

Rong-Lin Jian, Ming-Chi Luo, Cheng-Wei Huang, Chia-Ming Lee, Yu-Fan Lin, and Chih-Chung Hsu. Gtatrack: Winner solution to soccertrack 2025 with deep-eiou and global tracklet association. In *Proceedings of the 8th International ACM Workshop on Multimedia Content Analysis in Sports*, pp. 180–188, 2025.

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 *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 201–210. IEEE, 2020.

Maria Koshkina and James H Elder. Towards long-term player tracking with graph hierarchies and domain-specific features. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 1175–1185. IEEE, 2025.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021.

Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. Broadtrack: Broadcast camera tracking for soccer. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6177–6187. IEEE, 2025.

Amir M Mansourian, Vladimir Somers, Christophe De Vleeschouwer, and Shohreh Kasaei. Multitask learning for joint re-identification, team affiliation, and role classification for sports visual tracking. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pp. 103–112, 2023.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8834–8844. IEEE, 2022.

Xiaohan Nie, Shixing Chen, and Raffay Hamid. A robust and efficient framework for sports-field registration. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1935–1943. IEEE, 2021.

Yin May Oo, Yewon Hwang, Muhammad Amrulloh Robbani, Vanyi Chao, Ankhzaya Jamsrandorj, Hoang Quoc Nguyen, Kyung-Ryoul Mun, and Jinwook Kim. Broadcast2pitch: Game state reconstruction from unconstrained soccer videos. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4483–4493. IEEE, 2026.

Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 164–173, 2021.

Jiayuan Rao, Tianlin Gui, Haoning Wu, Yanfeng Wang, and Weidi Xie. Gentac: Generative modeling and forecasting of soccer tactics. *arXiv preprint arXiv:2604.11786*, 2026.

Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. Soccertrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3568–3578. IEEE, 2022.

Mattia Segu, Luigi Piccinelli, Siyuan Li, Yung-Hsu Yang, Luc Van Gool, and Bernt Schiele. Samba: Synchronized set-of-sequences modeling for multiple object tracking. In *International Conference on Learning Representations*, volume 2025, pp. 30057–30070, 2025.

Yuqing Shao, Yuchen Yang, Rui Yu, Weilong Li, Xu Guo, Huaicheng Yan, Wei Wang, and Xiao Sun. From detection to association: Learning discriminative object embeddings for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6878–6888, 2026.

Feng Shi, Paul Marchwica, Juan Camilo Gamboa Higuera, Michael Jamieson, Mehrsan Javan, and Parthipan Siva. Self-supervised shape alignment for sports field registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 287–296, 2022.

Vladimir Somers, Victor Joos, Anthony Cioppa, Silvio Giancola, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir M Mansourian, Xin Zhou, Shohreh Kasaei, et al. Soccernet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3293–3305. IEEE, 2024.

Jiacheng Sun, Hsiang-Wei Huang, Cheng-Yen Yang, Zhongyu Jiang, and Jenq-Neng Hwang. Gta: Global tracklet association for multi-object tracking in sports. In *Asian Conference on Computer Vision*, pp. 74–87. Springer, 2024.

Jonas Theiner and Ralph Ewerth. Tvcalib: Camera calibration for sports field registration in soccer. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1166–1175. IEEE, 2023.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pp. 3645–3649. IEEE, 2017.

Haolin Yang, Jiayuan Rao, Haoning Wu, and Weidi Xie. Soccermaster: A vision foundation model for soccer understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21549–21560, 2026.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European conference on computer vision*, pp. 659–675. Springer, 2022.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22056–22065. IEEE, 2023.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

# APPENDIX

In this appendix, we provide details of the evaluation metrics in Appendix A and qualitative comparisons in Appendix B.

## A PITCH-SPACE EVALUATION PROTOCOL

We evaluate player trajectories using the HOTA protocol with observation similarity defined in the canonical pitch coordinate system. For each frame $t$, let $\mathbf{p}^{\star}_{t,i}$ and $\mathbf{p}_{t,j}$ denote the ground-truth pitch position of player $i$ and the predicted pitch position of player $j$, respectively. Their localization similarity is defined as

$$s_t(i,j) = \exp\left(-\frac{\left\|\mathbf{p}^{\star}_{t,i} - \mathbf{p}_{t,j}\right\|_2^2}{2\sigma^2}\right), \tag{12}$$

where $\sigma$ controls the spatial tolerance of the similarity function. Following the SoccerNet-GSR (Somers et al., 2024) evaluation setting, $\sigma$ is chosen such that a pitch-space distance of $5$ meters corresponds to a similarity of $0.05$. Following HOTA, evaluation is performed over

$$\mathcal{A} = \{0.05, 0.10, \ldots, 0.95\}. \tag{13}$$

For each threshold $\alpha \in \mathcal{A}$, we follow the standard HOTA global-alignment and frame-wise Hungarian matching procedure using $s_t(i,j)$ as the observation similarity. Only matched pairs satisfying $s_t(i,j) \geq \alpha$ are retained. Let $TP_\alpha$, $FP_\alpha$, and $FN_\alpha$ denote the numbers of retained matches, unmatched predictions, and unmatched ground-truth observations, respectively. Pitch-space detection accuracy is computed as

$$\text{P-DetA}_\alpha = \frac{TP_\alpha}{TP_\alpha + FP_\alpha + FN_\alpha}. \tag{14}$$

For a ground-truth trajectory $g$ and a predicted trajectory $p$, let $M^{\alpha}_{gp}$ denote the number of successful matches between them at threshold $\alpha$, and let $n_g$ and $n_p$ denote their numbers of observations. Their association quality is

$$a^{\alpha}_{gp} = \frac{M^{\alpha}_{gp}}{n_g + n_p - M^{\alpha}_{gp}}, \tag{15}$$

from which pitch-space association accuracy is computed as

$$\text{P-AssA}_\alpha = \frac{\sum_{g,p} M^{\alpha}_{gp} a^{\alpha}_{gp}}{TP_\alpha}. \tag{16}$$

The localization quality of successfully matched observations is measured by

$$\text{P-LocA}_\alpha = \frac{1}{TP_\alpha} \sum_{(i,j,t)\in\mathcal{M}_\alpha} s_t(i,j), \tag{17}$$

where $\mathcal{M}_\alpha$ denotes the set of retained matches. P-HOTA jointly summarizes detection and association performance:

$$\text{P-HOTA}_\alpha = \sqrt{\text{P-DetA}_\alpha \, \text{P-AssA}_\alpha}. \tag{18}$$

We additionally report P-IDF1 to evaluate identity consistency under pitch-space matching. P-IDF1 uses the same pitch-space similarity but fixes the similarity matching threshold to $0.5$. A sequence-level one-to-one mapping between ground-truth and predicted trajectories is obtained by minimizing identity errors. Let $IDTP$, $IDFP$, and $IDFN$ denote the resulting identity true positives, false positives, and false negatives, respectively. P-IDF1 is defined as

$$\text{P-IDF1} = \frac{2IDTP}{2IDTP + IDFP + IDFN}. \tag{19}$$

For the HOTA-based metrics, statistics are first aggregated across all evaluation sequences at each threshold, after which the final values are obtained by averaging over the 19 thresholds in $\mathcal{A}$. P-IDF1 is instead computed by accumulating $IDTP$, $IDFP$, and $IDFN$ across sequences before applying the above definition.




Figure 2: Qualitative comparison of pitch-space player reconstruction between the decoupled MOTIP+PnLCalib baseline and UniPitch. Each row shows (a) the input frame, (b) the decoupled method's result, (c) the UniPitch result, and (d) the ground truth.

## B Qualitative Comparison

To further demonstrate the advantages of UniPitch beyond quantitative metrics, we provide qualitative comparisons of pitch-space reconstruction results in Figure 2. From left to right, each row shows (a) the input frame, (b) the player positions predicted by the decoupled MOTIP+PnLCalib baseline, (c) the predicted player positions from UniPitch and (d) the corresponding ground truth.

The values below the predicted pitch minimaps report the mean and median pitch-space localization errors, respectively, in meters. Numerical labels in the minimaps denote arbitrary tracklet identities, corresponding to MOTIP track IDs for the decoupled baseline and slot IDs for UniPitch. For visualization only, predicted players are matched to the ground truth using Hungarian matching to determine their display colors. Red denotes the left team, blue denotes the right team, and yellow denotes referees. These semantic attributes are used only for visualization and are not exploited by UniPitch for parametric reconstruction.

Compared with the decoupled baseline, UniPitch produces player positions that are substantially better aligned with the ground truth across different scenarios. The decoupled results frequently exhibit noticeable global displacement and local positional drift after projecting image-space tracks onto the pitch, indicating that errors from tracking and frame-wise calibration can accumulate through independent processing. In contrast, UniPitch preserves the overall formation more faithfully and places individual players closer to their corresponding pitch locations, including challenging views with limited field context or players concentrated in a small image region. The consistently smaller localization errors shown beneath the examples further agree with this visual trend. These comparisons illustrate the benefit of incorporating temporal camera information and trajectory refinement into pitch-space reconstruction rather than relying solely on independently estimated tracking and calibration outputs.