# MuPeT: Multi-Person Transformer for 3D Reconstruction from Broadcast Sports

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

## Abstract

High-level sporting events are often recorded from multiple viewpoints, providing abundant visual data to train models for 3D reconstruction and human pose estimation. However, these settings remain a significant blind spot for current models: they involve highly dynamic scenes with limited texture and repetitive backgrounds, captured by uncalibrated, long-range cameras. In this paper, we present MuPeT (Multi-Person Transformer), a model to reconstruct player poses and impacts in 3D under severely challenging conditions. Since training a Visual Geometry Grounded Transformer (VGGT) requires precise 3D supervision, absent in real sports data, we train it exclusively with synthetic renderings. The key enabler is the input modality: instead of pure RGB, which would present a large sim-to-real gap, our model consumes human segmentations and lines, which are easily rendered or computed from real data. We thus leverage the strengths of pretrained models while enabling purely-synthetic training, to learn the geometric biases inherent to these environments. We show that MuPeT successfully reconstructs difficult impact scenarios where previous approaches fail.

## 1 Introduction

High-level sporting events, such as the Rugby World Cup, are routinely captured by multiple long-range broadcast cameras, generating hours of footage from every match (World Rugby, 2026). Given the sheer volume of high-quality video recorded in elite sports, we might expect this footage to provide an ideal training ground for computer vision models designed to analyse games and player performance. Instead, it presents *almost exactly* the conditions that current models struggle to handle: the scene is highly dynamic, while the background offers little distinctive texture. Reliable matching is particularly difficult, as pitch markings and players' jerseys create highly repetitive visual patterns: many lines are similar, and half the players wear exactly the same colours. Uncalibrated cameras with widely varying zoom levels compound these challenges: in some views, a player may occupy only a handful of pixels; in others, a single player can fill almost the entire frame.

We build on the findings of RugbyBench (Anonymous, 2026b) to develop an approach specifically targeting these limitations. We introduce MuPeT, a modified version of VGGT (Wang et al., 2025a) trained on synthetic data designed to reproduce the multi-view rugby broadcast setting. To reduce the sim-to-real gap, we train MuPeT on 2D segmentations rather than RGB images. This makes it much easier to generalise to real data, as during testing, segmentations can be reliably extracted from real images with pretrained models. With segmentations as input, MuPeT explicitly reasons about player masks and their relationships across views, using the players themselves as cues for reconstructing 3D geometry and camera poses.

Since MuPeT is trained entirely on synthetic scenes without sports-specific poses, a residual sim-to-real gap remains. We therefore introduce an optional EM-based optimisation procedure that jointly refines cross-view player associations and scene geometry on real data. We compare our approach against several state-of-the-art 3D reconstruction and human pose estimation methods, and show that only MuPeT consistently recovers the geometry of these challenging multi-view scenes (Figure 1).

In summary, our contributions are: (i) a synthetic dataset of broadcast-like multi-view scenes; (ii) a synthetically trained reconstruction pipeline that learns scene geometry directly from player representations and lines; (iii) a post-inference optimisation strategy that jointly refines player associ-



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.

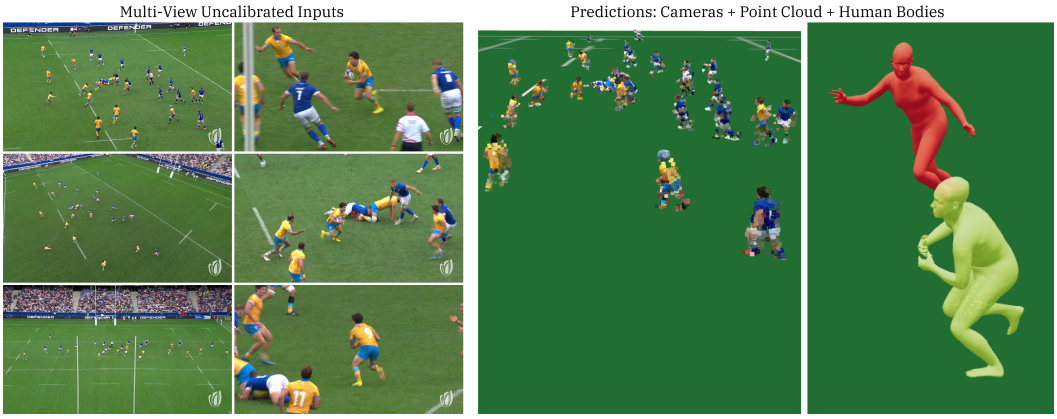


Figure 1: **MuPeT**: Reconstructing sports broadcasts in 3D. From uncalibrated multi-view footage, our method recovers camera geometry, player point clouds, and SMPL-X fits.

ations and scene geometry to mitigate the sim-to-real gap; and (iv) an extensive evaluation against state-of-the-art methods across both general 3D reconstruction and human-centric reconstruction. We will release code, pretrained models and the synthetic data generation pipeline upon publication.

## 2 Related Work

**Human Pose Estimation.** Predicting human pose from monocular images can be framed as regressing 2D keypoints (Jocher et al., 2023; Xu et al., 2022), lifting those detections to 3D (Zhu et al., 2023; Mehraban et al., 2023), or regressing the parameters of a parametric body model such as SMPL (Loper et al., 2023; Kanazawa et al., 2018). Monocular estimation is fundamentally ill-posed, and recent work traces much of the residual error to the approximate camera models used to generate pseudo-ground-truth: TokenHMR reweights 2D and pseudo-GT losses below an empirically derived error threshold and constrains predictions to a tokenized space of valid poses (Dwivedi et al., 2024), while CameraHMR predicts field of view and refits with a full perspective camera to obtain more accurate training annotations (Patel & Black, 2024). Multi-person methods recover all bodies in a single pass: PromptHMR (Wang et al., 2025d) processes the full image under spatial prompts such as boxes or masks to retain scene context, while Multi-HMR 2 (Fiche et al., 2026) predicts a scene-consistent camera jointly with the meshes, enabling metric localization without ground-truth intrinsics. Extending estimation to video (Goel et al., 2023) adds temporal consistency but introduces the problem of disentangling human motion from camera motion, which WHAM (Shin et al., 2023) and GVHMR (Shen et al., 2024) address by grounding motion in a world frame. Additional viewpoints resolve depth ambiguity directly, but existing multi-view methods are typically constrained by requiring known camera calibration (Iskakov et al., 2019; Liao et al., 2023) or by single-person formulations (Matsubara & Nishino, 2024).

**Human and Scene Reconstruction.** Before feed-forward reconstruction, jointly estimating people and cameras relied on the structure-from-motion toolkit (Schönberger & Frahm, 2016; Lindenberger et al., 2023). UncaliPose (Xu & Kitani, 2022) recovers 3D poses from arbitrary uncalibrated camera networks by casting cross-view person matching as a constrained clustering problem, then triangulating and bundle-adjusting the resulting correspondences. STBA (Vo et al., 2020) instead extends bundle adjustment itself: since the triangulation constraint is invalid for points that move between captures, it solves for the sub-frame temporal offsets between unsynchronized cameras alongside the geometry, recovering 3D human trajectories under physics-based motion priors. More recently, dense feed-forward models such as DUSt3R (Wang et al., 2023), MASt3R (Leroy et al., 2024), VGGT (Wang et al., 2025a; 2026) and $\pi^3$ (Wang et al., 2025c) have made scene geometry and camera parameters available without calibration or per-scene optimization, and a line of work has folded humans into that estimation. HSfM (Müller et al., 2024) initializes from off-the-shelf human and scene models and jointly optimizes multiple human meshes, scene point clouds, and cameras from sparse uncalibrated views, using the body model to recover the approximate metric scale that

the underlying reconstruction lacks. JOSH (Liu et al., 2026) applies a comparable joint optimization to monocular video, using human-scene contact as an additional constraint, while HAMSt3R (Rojas et al., 2025) and Human3R (Chen et al., 2026) replace optimization with feed-forward prediction, the latter building on CUT3R (Wang et al., 2025b) to recover multiple SMPL-X bodies, dense scene geometry, and camera trajectories in a single pass. CHROMM (Kim et al., 2026) and AHAP (Qiao et al., 2026) target the same multi-person uncalibrated setting with feed-forward architectures; we do not compare against them or HAMSt3R (Rojas et al., 2025), as there is no public implementation.

# 3 METHOD

Our pipeline consists of two main stages: a synthetic pretraining stage, designed to compensate for the limited geometric understanding of state-of-the-art models in such scenarios, followed by an optimisation procedure on real data to bridge the sim-to-real gap. Details of the synthetic pretraining stage are provided in Section 3.1, while the real-data optimisation procedure is in Section 3.2.

**Notation.** We start by introducing some basic notation. The scene is observed by $V = 6$ uncalibrated broadcast cameras, indexed by $v \in \mathcal{V} = \{1, \dots, V\}$. Each camera is parameterized as $\mathbf{g}_v = [\mathbf{q}_v, \mathbf{t}_v, \phi_v]$, where $\mathbf{q}_v \in \mathbb{R}^4$ is a unit quaternion with rotation $\mathbf{R}_v = \mathbf{R}(\mathbf{q}_v)$, $\mathbf{t}_v \in \mathbb{R}^3$ is the translation, and $\phi_v$ is the field of view. A 3D point $\mathbf{X}$ projects into view $v$ as

$$\pi_v(\mathbf{X}) = \Pi\big(\mathbf{K}_v(\mathbf{R}_v\mathbf{X} + \mathbf{t}_v)\big), \qquad \Pi\big([x, y, z]^\top\big) = [x/z,\ y/z]^\top, \tag{1}$$

where $\mathbf{K}_v = \mathbf{K}(\phi_v)$ is the intrinsic matrix with the principal point at the image centre. The pitch is modelled as a plane $\mathbf{g}_\Pi = [\mathbf{q}_\Pi, \mathbf{o}_\Pi]$ with origin $\mathbf{o}_\Pi \in \mathbb{R}^3$ (generally the pitch centre) and normal $\mathbf{n}_\Pi \in \mathbb{R}^3$. We choose the first view as the global reference frame. We will only use a superscript frame index $(t)$ when relevant.

## 3.1 SYNTHETIC PRETRAINING

**Dataset.** We construct a synthetic dataset designed to reproduce the main geometric and visual challenges encountered in real rugby broadcast footage. To approximate the distribution of broadcast cameras, we estimate the pitch position relative to each RugbyBench camera $\mathbf{g}_v$. These cameras are pseudo-ground truth, obtained by bundle adjustment on annotated 2D keypoints (Anonymous, 2026b), and are therefore subject to the inherent depth/focal-length ambiguity (Hartley et al., 2003). We use them only to fit per-role distributions over camera position and focal length, never individual per-frame values, which suffice to sample realistic broadcast viewpoints.

Given these initial camera and pitch distributions, we render synthetic scenes whose viewpoints and pitch-line projections resemble those observed in the real data. However, broadcast cameras do not independently focus on arbitrary regions of the pitch; instead, they are generally directed towards the region in which the game action is taking place. To reproduce this behaviour, we randomly sample a *Zone of Interest* (ZoI) on the pitch, representing the location of the current gameplay, and orient the sampled cameras towards this region. This will enable us to model high-contact events in detail.

Since realistic volumetric poses of rugby impacts are not available, we simulate a player collision within the Zone of Interest using the procedure introduced by CISyn (Anonymous, 2026a). CISyn uses a physics engine to simulate collisions between SMPL-X rigged humanoids. It reproduces the occlusions and pose entanglement typical of human interaction scenarios and naturally scales to multiple people. We then populate the surrounding region with additional players. Their positions relative to the Zone of Interest are sampled according to a Gaussian fit of player displacements observed in RugbyBench. We pose these surrounding players using human poses sampled from AMASS (Mahmood et al., 2019). In this way, the synthetic scenes reproduce both the dense player interactions around an impact and the broader spatial configuration of players observed in real rugby footage. Section E provides additional details on the synthetic dataset.

**2D Representation.** To reduce the computational cost of synthetic pretraining while leveraging the capabilities of large pretrained models, we avoid rendering photorealistic images. Instead, we render only the minimal set of visual cues required for geometric reconstruction: player representations and pitch-line detections. At inference time, these representations can be extracted directly

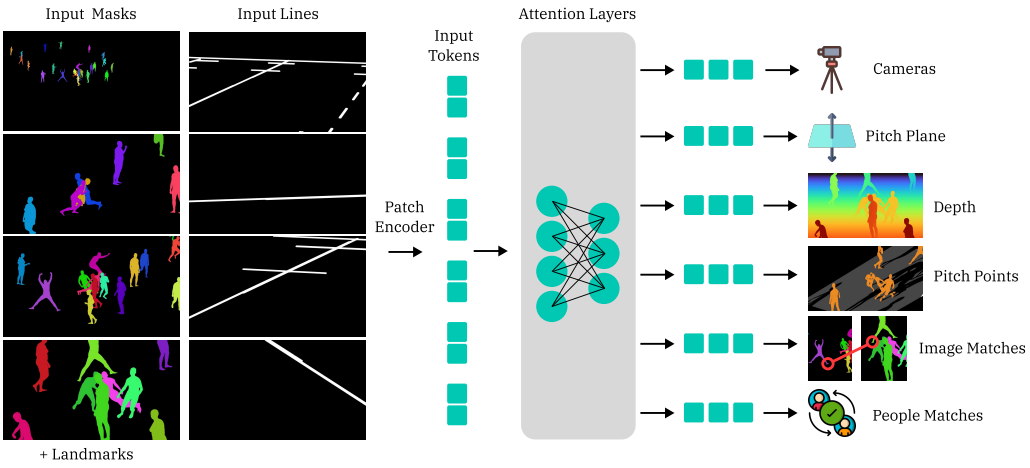


Figure 2: **Synthetic Pretraining Overview.** A VGGT-like backbone takes in instance segmentation masks, dense human landmarks (not shown), and detected lines (pitch lines). It is supervised to output simultaneously all camera poses and intrinsics, the pitch plane (as an origin and normal), a depth map and point map for each view, and cross-view 2D matches and player association affinities.

from real broadcast images using pretrained models, allowing our synthetic pretraining to focus on learning geometry rather than bridging differences in image appearance.

A key consideration is therefore the choice of 2D representation used to encode each player. A common approach in human 3D reconstruction pipelines is to use 2D keypoints (e.g., Uncalipose (Xu & Kitani, 2022)). However, as demonstrated by RugbyBench (Anonymous, 2026b), 2D pose estimation methods, which typically rely on bounding-box detection followed by keypoint regression, struggle under the severe occlusions and frequent player interactions characteristic of rugby footage.

For this reason, we use SAM3 (Carion et al., 2025) masks as our primary 2D player representation. We evaluate SAM3 against ViTPose for player detection on RugbyBench and find that SAM3 achieves an F1 score of 88.6, compared with 82.2 for ViTPose. This motivates our choice to use SAM3 for all further experiments. Further details of this comparison are provided in Section C.

While SAM3 masks provide accurate pixel-level support for distinguishing individual players, they lack the semantic correspondence provided by anatomical keypoints. We therefore augment each mask with dense semantic information derived from MAMMANet (Cuevas Velasquez et al., 2026) landmarks. The resulting representation combines the robustness and spatial coverage of segmentation masks with dense anatomical correspondence, providing the geometric cues the subsequent reconstruction stage requires without photorealistic synthetic rendering.

Formally, the input for view $v$ is $\mathbf{I}_v = (\mathbf{S}_v, \mathbf{\Lambda}_v, \mathbf{L}_v)$, defined over the image domain $\Omega$. $\mathbf{S}_v$ is the instance segmentation, containing a set of masks $\mathcal{M}_v$, where each mask $m$ has pixel support $\Omega_m \subset \Omega$. We write $\mathcal{M} = \bigcup_v \mathcal{M}_v$ for the masks of all views and $v(m)$ for the view containing mask $m$. $\mathbf{\Lambda}_v : \Omega \to \mathcal{K} \cup \{0\}$ assigns each pixel a MAMMA landmark label from $\mathcal{K} = \{1, \dots, K\}$, with 0 denoting no landmark, and $\mathbf{L}_v$ is the pitch-line map. For each mask $m$ we denote by $\mathcal{K}_m \subseteq \mathcal{K}$ its visible landmarks and by $\mathbf{x}_{m,k} \in \mathbb{R}^2$ the 2D position of landmark $k \in \mathcal{K}_m$. Finally, the set of landmark pixels in view $v$ is

$$\mathcal{U}_v = \{\mathbf{u} \in \Omega : \mathbf{\Lambda}_v(\mathbf{u}) \neq 0\}. \tag{2}$$

We associate each landmark pixel $\mathbf{u} \in \mathcal{U}_v$ with a mask $m(\mathbf{u})$ and a landmark ID $k(\mathbf{u})$.

**Architecture.** We base our architecture on VGGT (Wang et al., 2025a), with several modifications to better accommodate the structure and challenges of our setting. First, we replace the RGB inputs with the 2D representations described above. As these inputs differ substantially from natural images, we replace the pretrained DINOv2 (Oquab et al., 2023) image encoder with a ResNet-34 (He et al., 2015) backbone trained from scratch. Since the input representations are simpler, we also reduce the number of attention layers from 24 to 16. Finally, we replace the original VGGT attention blocks with the more efficient register-attention blocks introduced in VGGT-$\Omega$ (Wang et al., 2026).

**Outputs.** Our model jointly predicts the camera intrinsics and extrinsics $\hat{\mathbf{g}}_v$, as well as the pitch plane representation used by the RugbyBench optimisation pipeline (Anonymous, 2026b) ($\hat{\mathbf{g}}_\Pi = [\hat{\mathbf{q}}_\Pi, \hat{\mathbf{o}}_\Pi]$). The model predicts per-view depth maps $\hat{D}_v$ with confidence $\hat{\Sigma}_v$. We also predict point maps $\hat{\mathbf{X}}_v(\mathbf{u}) \in \mathbb{R}^3$, with point-map predictions restricted to landmark pixels $\mathbf{u} \in \mathcal{U}_v$, because we want to focus the model's capacity on the parts of the scene most relevant to human reconstruction, while the pitch prediction already provides a representation of the remaining scene geometry. We also predict cross-view image-level associations $\hat{\mathbf{u}}_{v\to v'}(\mathbf{u})$ (the location in view $v'$ of landmark pixel $\mathbf{u}$ from view $v$), restricted to the same landmarks. Per-mask 3D landmarks are read from the point map at the landmark positions

$$\hat{\mathbf{X}}_{m,k} = \hat{\mathbf{X}}_{v(m)}(\mathbf{x}_{m,k}), \qquad k \in \mathcal{K}_m. \tag{3}$$

**Mask token.** To explicitly reason about player identity across views, we introduce an additional affinity head. Each mask $m \in \mathcal{M}$, with pixel support $\Omega_m$, is represented by a token $\mathbf{z}_m$ combining its appearance and its predicted geometry:

$$\mathbf{z}_m = \big[\, \mathbf{f}_m,\ \mathbf{e}_m \,\big], \qquad \text{with} \qquad \mathbf{f}_m = \frac{1}{|\Omega_m|} \sum_{\mathbf{u}\in\Omega_m} \mathbf{F}_{v(m)}(\mathbf{u}), \qquad \mathbf{e}_m = \psi\big(\,\mathrm{sg}[\hat{\mathcal{X}}_m]\big), \tag{4}$$

where $\mathbf{F}_v$ are the backbone features of view $v$, $\hat{\mathcal{X}}_m = \{\hat{\mathbf{X}}_{m,k}\}_{k\in\mathcal{K}_m}$ are the predicted 3D landmarks of the mask, $\psi$ is an MLP encoder, and $\mathrm{sg}[\cdot]$ denotes stop-gradient, so that the token does not backpropagate the association loss into the 3D prediction head. A 4-layer transformer then processes the tokens of all $V$ views jointly, yielding enriched tokens $\{\bar{\mathbf{z}}_m\}_{m\in\mathcal{M}}$ that aggregate cross-view context.

**Pairwise affinities.** To score whether two masks from different views depict the same player, we compare their enriched tokens: for every pair $m \in \mathcal{M}_v$ and $m' \in \mathcal{M}_{v'}$ with views $v \neq v'$, we create the pairwise representation

$$\mathbf{h}_{mm'} = \big[\, \bar{\mathbf{z}}_m \odot \bar{\mathbf{z}}_{m'},\ |\bar{\mathbf{z}}_m - \bar{\mathbf{z}}_{m'}|,\ \mathbf{c}_{mm'},\ \delta_{mm'} \,\big], \qquad \delta_{mm'} = \underset{k\in\mathcal{K}_m\cap\mathcal{K}_{m'}}{\mathrm{mean}} \big\|\hat{\mathbf{X}}_{m,k} - \hat{\mathbf{X}}_{m',k}\big\|_2, \tag{5}$$

where $\mathbf{c}_{mm'}$ encodes the 2D geometry of both masks, and $\delta_{mm'}$ is the Euclidean distance between their predicted 3D landmarks. An MLP maps $\mathbf{h}_{mm'}$ to an affinity logit $a_{mm'}$ representing the likelihood that the two masks correspond to the same player. Together with a learned "none" logit $a_{m\varnothing} := a_\varnothing$, which allows masks with no corresponding detection in view $v'$ to remain unmatched, the logits are normalised over the masks of that view:

$$\hat{y}_{mm'} = \underset{m'\in\mathcal{M}_{v'}\cup\{\varnothing\}}{\mathrm{softmax}} \big(a_{mm'}\big). \tag{6}$$

At inference time, these predicted probabilities are used to group masks across views, enforcing at most one mask per player in each view. These associations provide the cross-view player correspondences that are required by the subsequent optimisation stage. The complete architecture and prediction pipeline are illustrated in Figure 2.

**Training losses.** The training objectives follow VGGT for the formulation of camera loss: $\mathcal{L}_{\text{camera}} = \sum_{v\in\mathcal{V}} \|\hat{\mathbf{g}}_v - \mathbf{g}_v\|_1$, and the uncertainty-weighted depth loss $\mathcal{L}_{\text{depth}}$. The pitch plane pose head is implemented as a VGGT camera head: its rotation and translation outputs predict $\hat{\mathbf{g}}_\Pi = [\hat{\mathbf{q}}_\Pi, \hat{\mathbf{o}}_\Pi]$, and its field-of-view output is left unsupervised. This yields the loss $\mathcal{L}_{\text{pitch}} = \|\hat{\mathbf{g}}_\Pi - \mathbf{g}_\Pi\|_1$.

In the synthetic data, we write $m \sim m'$ when two masks depict the same player. The ground-truth correspondence of a landmark pixel $\mathbf{u} \in \mathcal{U}_v$ (Eq. 2) in view $v'$ is then $\mathbf{u}_{v\to v'}(\mathbf{u}) = \pi_{v'}\big(\mathbf{X}_v(\mathbf{u})\big)$, the projection of its ground-truth 3D landmark. We denote by $\mathcal{U}_{v\to v'} \subseteq \mathcal{U}_v$ the landmark pixels whose ground-truth 3D landmark lies in front of camera $v'$.

The 3D people point maps ($\mathcal{L}_{\text{3Dpoints}}$) and 2D correspondences ($\mathcal{L}_{\text{2Dpoints}}$) are then supervised using a Huber loss:

$$\begin{aligned} \mathcal{L}_{\text{3Dpoints}} &= \sum_{v\in\mathcal{V}} \sum_{\mathbf{u}\in\mathcal{U}_v} \mathrm{Huber}\,\big(\hat{\mathbf{X}}_v(\mathbf{u}) - \mathbf{X}_v(\mathbf{u})\big), \\ \mathcal{L}_{\text{2Dpoints}} &= \sum_{v\in\mathcal{V}} \sum_{v'\neq v} \sum_{\mathbf{u}\in\mathcal{U}_{v\to v'}} \mathrm{Huber}\,\big(\hat{\mathbf{u}}_{v\to v'}(\mathbf{u}) - \mathbf{u}_{v\to v'}(\mathbf{u})\big), \end{aligned} \tag{7}$$

where $\mathbf{X}_v(\mathbf{u})$ is the ground-truth 3D position at landmark pixel $\mathbf{u}$. The affinity head is supervised with a cross-entropy over the association probabilities $\hat{y}_{mm'}$ (Eq. 6):

$$\mathcal{L}_{\text{people}} = -\sum_{m\in\mathcal{M}} \sum_{v'\neq v(m)} \sum_{m'\in\mathcal{M}_{v'}\cup\{\varnothing\}} y_{mm'} \log \hat{y}_{mm'}, \tag{8}$$

where the ground-truth associations are $y_{mm'} = \mathbb{1}[m \sim m']$, and the special "none" association $y_{m\varnothing} = 1$ if no mask in view $v'$ depicts the same player as $m$.

We additionally use regularisation losses ($\mathcal{L}_{\text{reg}}$), including the VGGT-$\Omega$ point loss restricted to human regions, and a cross-view 3D consistency loss $\mathcal{L}_{\text{cons}}$ on corresponding player and pitch-line pixels. The complete training objective is a linear combination of the terms described above. Further details are provided in Section B.

### 3.2 Real Data Optimisation

Even with these measures to reduce the sim-to-real gap, some residual difference always remains. To refine the model predictions and enforce geometric consistency on real sequences, we therefore introduce an post-prediction optimisation stage. We leverage the standard bundle adjustment (BA) procedure from RugbyBench, described in Section A for completeness. Our main addition is to speed it up by improving the sparsity pattern for analytic gradients (Section A and Fig. 5).

**EM algorithm.** The RugbyBench bundle adjustment assumes ground-truth cross-view correspondences, which are not available in our setting. We therefore jointly optimise player assignments and scene geometry using an expectation-maximisation (EM)-style procedure (Dempster et al., 1977) (Algorithm 1). The M-step performs bundle adjustment by minimising $\mathcal{L}_{\text{BA}}$ (Eq. 12) under the current assignments, while the E-step reassigns masks to players using a score that combines geometric consistency with the affinities predicted by the model. We replace the 2D–3D joint correspondences used in RugbyBench with the denser MAMMA landmarks. To reduce the cost of the first optimisation, we adopt a coarse-to-fine strategy, initially using a reduced set of landmarks before progressively increasing to the full set $\mathcal{K}$ of $K = 512$ landmarks.

**E-step.** At each E-step, the score of assigning mask $m$ to player $p$ depends on both geometric consistency and predicted affinity:

$$s_{mp} = \exp\big(-d_{mp}/\tau\big)\,\alpha_{mp}^{\eta}, \qquad \text{with} \qquad d_{mp} = \underset{k\in\mathcal{K}_m}{\text{median}} \left\|\pi_{v(m)}(\mathbf{X}_{p,k}) - \mathbf{x}_{m,k}\right\|_2, \tag{9}$$

Intuitively, $d_{mp}$ measures how well player $p$'s current reconstruction explains mask $m$, so masks close to a player's projected landmarks score higher; halving $\tau$ at each iteration progressively makes the assignments more one-hot. The learned affinity $\alpha_{mp}$, which is the mean probability $\hat{y}_{mm'}$ between $m$ and the masks already assigned to $p$ (Eq. 6), ensures that appearance is weighted by $\eta$.

**M-step.** The scores are converted into soft assignments $w_{mp}$ using a per-view Sinkhorn normalisation (Knight, 2008) with an explicit *none* option, allowing masks with no cross-view correspondence to remain unassigned. In the subsequent M-step, the terms of each mask $m$ in Eq. 11 are weighted by $w_{mp}$. After each iteration, we update the set of players $\mathcal{P}$: remove players observed in fewer than two views, group geometrically consistent unassigned masks into new players, and merge players whose reconstructed landmarks coincide.

**Application of EM.** We first run EM with a subset $\mathcal{K}^{(64)} \subset \mathcal{K}$ of 64 landmarks. Once the assignments converge, we get discrete correspondences with per-view Hungarian matching (Kuhn, 1955) and keep them fixed while refining the bundle adjustment (BA) with $\mathcal{K}^{(128)}$, and finally $\mathcal{K}$.

**Human shape fitting.** SMPL-X parameters (Pavlakos et al., 2019) can optionally be fitted to each reconstructed player. We initialise the body pose using inverse kinematics on the reconstructed landmarks $\mathbf{X}_{p,k}$ and fix the shape parameters to zero. We then optimise pose and shape for 1500 iterations using Adam, minimising

$$\mathcal{L}_{\text{SMPL-X}} = \sum_{m\in\mathcal{M}_p} \sum_{k\in\mathcal{K}_m} \text{Huber}\left(\pi_{v(m)}\big(\mathcal{S}_k(\boldsymbol{\theta}_p, \boldsymbol{\beta}_p)\big) - \mathbf{x}_{m,k}\right), \tag{10}$$

where $\mathcal{S}_k$ returns SMPL-X landmark $k$ for pose $\boldsymbol{\theta}_p$ (including global pose) and shape $\boldsymbol{\beta}_p$.

Table 1: **3D reconstruction on RugbyBench.** Comparison with general-purpose multi-view reconstruction methods in terms of valid reconstruction coverage, leave-one-out (LOO) reprojection error, and cross-view player matching F1 (precision/recall).

| Method | Coverage ↑ | LOO Reproj. (px) ↓ | People Matches F1 (P/R) ↑ |
|---|---|---|---|
| COLMAP + LightGlue | 27.0% | 764.4 | 8.0 (91.7/4.2) |
| MASt3R | 54.1% | 884.7 | 28.8 (56.2/19.3) |
| VGGT | 84.2% | 496.8 | 11.7 (21.1/8.1) |
| VGGT-Ω | 55.4% | 339.1 | 3.3 (41.8/1.7) |
| $\pi^3$ | 87.8% | 353.9 | 2.5 (33.2/1.3) |
| Ours | **97.5%** | **107.1** | **81.0** (86.8/75.9) |

# 4 RESULTS

We evaluate our method from three complementary perspectives. First, we compare against state-of-the-art feedforward 3D reconstruction methods (Section 4.1). We then evaluate against state-of-the-art single- and multi-view human-centric reconstruction methods (Section 4.2). Finally, we present ablation studies to assess the contribution of each individual component (Section 4.3).

## 4.1 3D RECONSTRUCTION METHODS COMPARISON

We compare our approach against recent feedforward reconstruction methods, including VGGT (Wang et al., 2025a), VGGT-Ω (Wang et al., 2026), MASt3R (Leroy et al., 2024), and $\pi^3$ (Wang et al., 2025c), as well as the SfM pipeline COLMAP, augmented with LightGlue (Lindenberger et al., 2023) correspondences. Results are reported in Table 1.

For each method, a frame is valid only if it poses the cameras and every annotated player, triangulated from the other views, projects in front of the held-out camera; we report the leave-one-out reprojection error in pixels over valid frames and the share of valid frames as coverage. We additionally evaluate cross-view player matching using precision, recall, and F1 score. All methods reconstruct from all six views; the held-out view is excluded only from triangulation, since including it would make its reprojection error near-zero by construction and thus uninformative about cross-view consistency. Since only our method explicitly predicts player correspondences, we derive correspondences from the outputs of the other methods. For methods with a matching head, two players are considered matched if at least half of 32 randomly sampled pixels within one player are matched to the corresponding player in the other view. For methods without an explicit matching head, we instead establish player correspondences from their reconstructed geometry.

Our method outperforms all competing approaches by a substantial margin across the evaluated metrics and is the only method to consistently produce geometrically valid reconstructions in this challenging setting.

## 4.2 HUMAN-CENTRIC METHODS COMPARISON

We compare our approach against human pose estimation and joint human–scene reconstruction methods in Table 2. Since RugbyBench consists of uncalibrated broadcast footage, we restrict the comparison to methods that either do not require camera calibration or operate on individual views. We report world MPJPE (W-MPJPE, in metres) for all methods, computed after a single similarity alignment (rotation, translation and scale) per frame and view, fitted jointly over all players visible in that view; for multi-view methods, including ours, the shared reconstruction is aligned to each view separately. W-MPJPE thus reflects errors in relative player placement and scale, not just per-body pose. For the multi-view methods, HSfM (Müller et al., 2024) and UncaliPose (Xu & Kitani, 2022), we additionally report leave-one-out (LOO) reprojection error.

Both HSfM and UncaliPose fail to recover a consistent scene reconstruction, despite providing HSfM with ground-truth player identities and UncaliPose with ground-truth camera intrinsics. UncaliPose relies on appearance re-identification, which cannot separate players in identical kits, leaving too few cross-view correspondences for its camera estimation, so 79% of frames yield no cam-

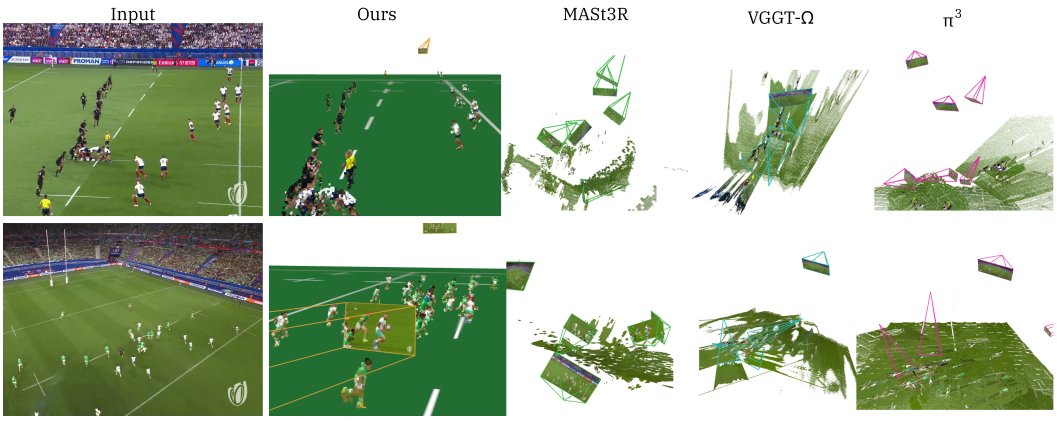


Figure 3: **Qualitative 3D reconstruction comparison.** Reconstructed scene geometry for two RugbyBench examples using MuPeT, MASt3R, VGGT-Ω, and $\pi^3$. MuPeT recovers consistent camera, pitch, and player geometry across the challenging multi-view broadcast scenes.

Table 2: **Human-centric reconstruction on RugbyBench.** Comparison with multi-view and monocular methods using reconstruction coverage, world MPJPE (W-MPJPE) across different view types, and leave-one-out (LOO) reprojection error.

| Method | Coverage↑ | W-MPJPE (m)↓ | | | LOO Reproj. (px)↓ |
|---|---|---|---|---|---|
| | | all | close-up | wide | |
| *Multi-view* | | | | | |
| HSfM (GT identities) | 20.9% | 8.43 | 4.32 | 10.15 | 680 |
| UncaliPose | 1.7% | 7.70 | – | 7.70 | 16758 |
| *Monocular, per view* | | | | | |
| PromptHMR (GT boxes) | 100% | 6.23 | 2.90 | 7.53 | – |
| Multi-HMR 2 | 100% | 3.66 | 2.52 | 4.30 | – |
| Ours | 97.5% | **1.43** | **1.26** | **1.49** | **107.1** |

eras and the rest a collapsed rig. HSfM, designed for a few people filmed by nearby cameras, takes as input pretrained model estimates that are incorrect for distant broadcast views.

For PromptHMR, we use the image-based variant: its video-based method either assumes a static camera, which is incompatible with the pan–tilt–zoom (PTZ) cameras in our setting, or requires an SfM reconstruction for dynamic cameras. We evaluate PromptHMR, Multi-HMR2, and our method in the same single-frame setting. PromptHMR, despite being provided with ground-truth player bounding boxes, and Multi-HMR2 both struggle to recover a coherent 3D configuration in scenes with dense and complex player interactions. In particular, errors in estimated player depth and scale lead to inconsistent relative positions and heights across the reconstructed scene.

### 4.3 ADDITIONAL ANALYSIS

**Ablations** We ablate the design choices leading to our final architecture in Table 4. Since our ultimate objective is human pose reconstruction, we initially explored directly regressing posed SMPL-X meshes from the input representations. We considered two query formulations. First, we used queries anchored to player instances detected in the wide view. This formulation transferred poorly to real data: players are often very small in the wide view and can therefore be missed by the detector, severe occlusions can make them entirely unobservable, and some players may lie outside the camera's field of view. Since a query is created only for players detected in this view, any such missed player cannot be recovered from the remaining views. We also experimented with learned queries, but these transferred even worse, likely due to their lack of an explicit association with any input view. Replacing the SMPL-X output with landmark predictions while retaining the same decoder remained below the wide-view formulation. Instead, removing the decoder in favour

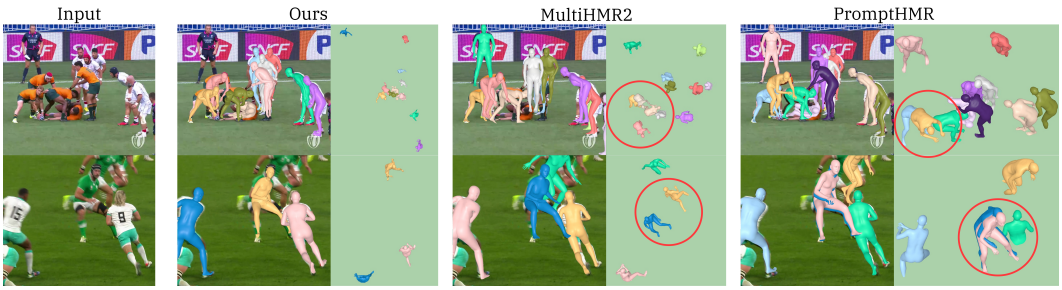


Figure 4: **Human-centric reconstruction comparison.** Qualitative results on two challenging RugbyBench scenes. MuPeT recovers coherent player geometry under dense interactions, while MultiHMR2 and PromptHMR show substantial errors in player depth and relative positioning.

Table 3: **Effect of EM optimisation.** Comparison of the model predictions before and after EM optimisation, evaluated on all frames and the best 50% of initial predictions. EM consistently improves reconstruction geometry and player association.

| Method | Coverage (%)↑ | Reproj. (px)↓ | Assoc. F1 (%)↑ | W-MPJPE (m)↓ |
|---|---|---|---|---|
| *All frames* | | | | |
| model | 96.6 | 185 | 80.3 | 2.71 |
| EM | 97.5 | **107.1** | **81.0** | **1.43** |
| *Best 50%* | | | | |
| model | 98.5 | 114.1 | 88.6 | 1.76 |
| EM | 100.0 | **27.1** | **89.3** | **1.14** |

Table 4: **Architecture ablations.** Ablation of the main design choices leading to our final architecture. Per-mask predictions substantially improve player association, while explicit 3D landmark prediction reduces reprojection error.

| Formulation | Assoc. (F1%)↑ | Reproj. (px)↓ |
|---|---|---|
| SMPL-X dec., wide queries | 48.9 | 781 |
| SMPL-X dec., learned queries | 28.1 | 283 |
| Landmark dec., learned queries | 43.4 | 209 |
| Affinity head, no 3D points | 77.9 | 536 |
| Full model | **80.3** | **185** |

of pixel-aligned, per-mask predictions, linked across views by the affinity head of our final model, raised association on real data from 50% to 80%.

Finally, we evaluate the VGGT-Ω strategy of recovering 3D geometry from the predicted cameras and depth maps alone. While this provides 3D estimates of the landmarks without an explicit 3D prediction head, directly regressing the landmarks in world coordinates halves the reprojection error. We therefore include explicit 3D landmark regression in our final architecture.

**EM contribution** We analyse the effectiveness of our EM procedure on the model predictions in Table 3. The EM optimisation improves both player association and reconstruction geometry, with consistent gains across both the strongest and weakest initial model predictions. The main improvements in association come from players that the model initially fails to associate across views, resulting in single-mask clusters. By combining geometric consistency with the predicted affinities, EM is able to assign many of these masks to their correct player clusters. The EM algorithm takes about three minutes per frame on a 72-core CPU, with no GPU.

## 5 Conclusion

We presented MuPeT, a feed-forward 3D reconstruction network centred on human pose estimation in broadcast sports (namely rugby), which leverages rasterised input modalities to alleviate the sim-to-real gap in synthetic training. We also propose an EM refinement procedure that further boosts performance. Overall, we find that uncalibrated 3D reconstruction of in-the-wild sports footage is possible despite the many challenges. Remaining limitations, including the dependence on upstream segmentation, evaluation against pseudo-ground truth, and the fixed camera rig, are discussed in Appendix F. We hope our work provides a roadmap for scenarios with abundant data but limited 3D ground truth.

### AI USE STATEMENT

Generative AI tools, including coding agents, were used to assist with the implementation of the methods described in this work and with drafting and editing parts of the manuscript. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted text was reviewed and revised for correctness and accuracy. The authors take full responsibility for the final content of the paper, including all code, experimental results, claims, and text produced with the assistance of generative AI tools.

## REFERENCES

Authors Anonymous. Cisyn: Human mesh recovery in close interactions with physics-simulated synthetic data. *Under Review, attached in supplementary*, 2026a.

Authors Anonymous. Tackling dense contact: A multi-view benchmark for 3d human pose in rugby. *Under Review, attached in supplementary*, 2026b.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. In *International Conference on Learning Representations (ICLR)*, 2026.

Hanz Cuevas Velasquez, Anastasios Yiannakidis, Soyong Shin, Giorgio Becherini, Markus Höschle, Joachim Tesch, Taylor Obersat, Tsvetelina Alexiadis, Eni Halilaj, and Michael J. Black. MAMMA: Markerless Accurate Multi-person Motion Acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39 (1):1–22, 1977.

Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1323–1333, 2024.

Guénolé Fiche, Philippe Weinzaepfel, Romain Brégier, and Fabien Baradel. Multi-HMR 2: Multi-person camera-centric human detection, mesh recovery and tracking, 2026. URL https://arxiv.org/abs/2606.14841.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14737–14748, 2023.

Richard Hartley, Andrew Zisserman, et al. *Multiple view geometry in computer vision*, volume 2. Cambridge university press Cambridge, 2003.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7717–7726, 2019.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. URL https://github.com/ultralytics/ultralytics.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

Sangmin Kim, Minhyuk Hwang, Geonho Cha, Dongyoon Wee, and Jaesik Park. Coherent human-scene reconstruction from multi-person multi-view video in a single pass, 2026. URL https://arxiv.org/abs/2603.12789.

Philip A. Knight. The sinkhorn–knopp algorithm: Convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, 2024.

Ziwei Liao, Jialiang Zhu, Chunyu Wang, Han Hu, and Steven L. Waslander. Multiple view geometry transformers for 3d human pose estimation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 708–717, 2023.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023.

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4d human-scene reconstruction in the wild. *The Fourteenth International Conference on Learning Representations*, 2026.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

Yuto Matsubara and Ko Nishino. Heatformer: A neural optimizer for multiview human mesh recovery. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6415–6424, 2024.

Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6905–6915, 2023.

Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21948–21958, 2024.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Priyanka Patel and Michael J. Black. Camerahmr: Aligning people with perspective. *2025 International Conference on 3D Vision (3DV)*, pp. 1562–1571, 2024.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Xiaozhen Qiao, Wenjia Wang, Zhiyuan Zhao, Jiacheng Sun, Ping Luo, Hongyuan Zhang, and Xuelong Li. Ahap: Reconstructing arbitrary humans from arbitrary perspectives with geometric priors, 2026. URL https://arxiv.org/abs/2602.23951.

Sara Rojas, Matthieu Armando, Bernard Ghanem, Philippe Weinzaepfel, Vincent Leroy, and Grégory Rogez. Hamst3r: Human-aware multi-view stereo 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5027–5037, 2025.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. *SIGGRAPH Asia 2024 Conference Papers*, 2024.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2070–2080, 2023.

Minh Vo, Yaser Sheikh, and Srinivasa G. Narasimhan. Spatiotemporal bundle adjustment for dynamic 3d human reconstruction in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:1066–1080, 2020.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. Vggt: Visual geometry grounded transformer. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-$\Omega$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2023.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. $\pi^3$: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025c.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael John Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1148–1159, 2025d.

World Rugby. Fixtures and results: 2026, 2026. URL https://www.world.rugby/tournaments/index.html. World Rugby fixtures and results database.

Yan Xu and Kris Kitani. Multi-view multi-person 3d pose estimation with uncalibrated camera networks. In *British Machine Vision Conference*, 2022.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *ArXiv*, abs/2204.12484, 2022.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15039–15053, 2023. doi: 10.1109/ICCV51070.2023.01385.

## A Bundle Adjustment

We base our optimisation on the bundle adjustment (BA) formulation introduced in RugbyBench (Anonymous, 2026b). Given the observed 2D landmarks $\mathbf{x}_{m,k}$ and an assignment of masks to players, where $\mathcal{M}_p$ denotes the masks of player $p \in \mathcal{P}$, BA jointly refines the cameras $\mathbf{g}_v$ and 3D landmarks $\mathbf{X}_{p,k}$ of each player, by minimising a reprojection error:

$$\mathcal{L}_{\text{reproj}} = \sum_{p\in\mathcal{P}} \sum_{m\in\mathcal{M}_p} \sum_{k\in\mathcal{K}_m} \left\| \pi_{v(m)}(\mathbf{X}_{p,k}) - \mathbf{x}_{m,k} \right\|_2^2, \tag{11}$$

where $\pi_v$ is the projection of Eq. 1, together with geometric constraints on the pitch, feet, and player limb lengths. The full objective is

$$\mathcal{L}_{\text{BA}} = \mathcal{L}_{\text{reproj}} + \mu_{\text{line}}\mathcal{L}_{\text{line}} + \mu_{\text{feet}}\mathcal{L}_{\text{feet}} + \mu_{\text{limb}}\mathcal{L}_{\text{limb}}, \tag{12}$$

where $\mathcal{L}_{\text{line}}$ aligns the projected pitch lines of the plane $\mathbf{g}_\Pi$ with the detected lines $\mathbf{L}_v$, and $\mathcal{L}_{\text{feet}}$ keeps the feet close to the plane, both as in RugbyBench. While this formulation is effective and aligns well with the outputs predicted by our model, its computational cost makes it impractical for our setting: optimising a 15-frame window takes approximately one hour. We introduce two modifications to substantially improve its efficiency. First, we replace the automatically computed finite-difference Jacobian with a closed-form analytical Jacobian. Second, we reparameterise the limb-length constraint to reduce the coupling between the optimised 3D points.

In RugbyBench, the length of bone $b = (k_b, k'_b)$ for player $p$ is $\ell_{p,b} = \left\| \mathbf{X}_{p,k_b} - \mathbf{X}_{p,k'_b} \right\|_2$, and consistency is enforced by penalising deviations from the mean length of that bone across players:

$$\mathcal{L}_{\text{limb}}^{\text{RB}} = \sum_{b\in\mathcal{B}} \sum_{p\in\mathcal{P}_b} \left(\ell_{p,b} - \bar{\ell}_b\right)^2, \qquad \bar{\ell}_b = \frac{1}{|\mathcal{P}_b|} \sum_{p\in\mathcal{P}_b} \ell_{p,b}, \tag{13}$$

where $\mathcal{P}_b$ are the players for which bone $b$ is observed. Since $\bar{\ell}_b$ depends on the 3D landmarks of every player, this formulation couples otherwise independent 3D variables. We instead introduce an independent optimisable limb-length parameter $\ell_b^*$, initialised to $\bar{\ell}_b$, and minimise $\mathcal{L}_{\text{limb}} = \sum_{b\in\mathcal{B}} \sum_{p\in\mathcal{P}_b} \left(\ell_{p,b} - \ell_b^*\right)^2$. This removes the direct coupling between different players' 3D landmarks, making $\mathbf{J}^\top\mathbf{J}$ (restricted to the 3D points) block-diagonal, as illustrated in Figure 5.

Together, the analytical Jacobian and the reparameterised limb-length constraint reduce the computational cost of each bundle-adjustment iteration by approximately two orders of magnitude, making the optimisation practical for use in our reconstruction pipeline.

## B Additional Architecture Details

**Camera views.** The six views come in a fixed order (Wide, High-Behind Left, Cable, High-Behind Right, Tight, Reverse), and all geometry is in metres in the Wide camera frame. Figure 6 provides an

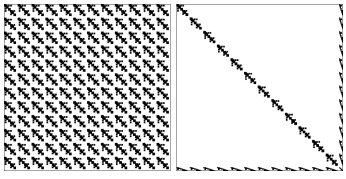

Figure 5: **Sparsity.** Sparsity pattern of $\mathbf{J}^\top\mathbf{J}$ for the coupled formulation (left) and our reduced-coupling reparameterisation (right).

Algorithm 1: EM-style Optimisation

**Require:** predicted cameras $\hat{\mathbf{g}}_v$, probabilities $\hat{y}_{mm'}$, and 3D landmarks $\hat{\mathbf{X}}_{m,k}$ (Eq. 3)
1: Initialise $\mathcal{P}$ from affinity clusters, $\mathbf{X}_{p,k}$ from $\hat{\mathbf{X}}_{m,k}$, and temperature $\tau$
2: **repeat**
3: **M-step:** minimise $\mathcal{L}_{\text{BA}}$ (Eq. 12)
4: **E-step:** compute $s_{mp}$ (Eq. 9) and Sinkhorn assignments $w_{mp}$
5: Update $\mathcal{P}$ (add, remove, merge)
6: $\tau \leftarrow \tau/2$
7: **until** assignments converge
8: Hungarian matching to obtain discrete $\mathcal{M}_p$
9: Refine BA with $\mathcal{K}^{(128)}$, then $\mathcal{K}$

example of what the images from each viewpoint look like. Cameras are encoded as $[\mathbf{t}_v, \mathbf{q}_v, \phi_v^x, \phi_v^y]$ with the principal point at the image centre. Inputs are $1024 \times 576$, split into $32 \times 32$ patches. Because cross-view attention carries no view index and the pitch is symmetric, we add a learned per-view embedding to all tokens of each view, so that, e.g., High-Behind Right is not confused with a mirrored High-Behind Left.

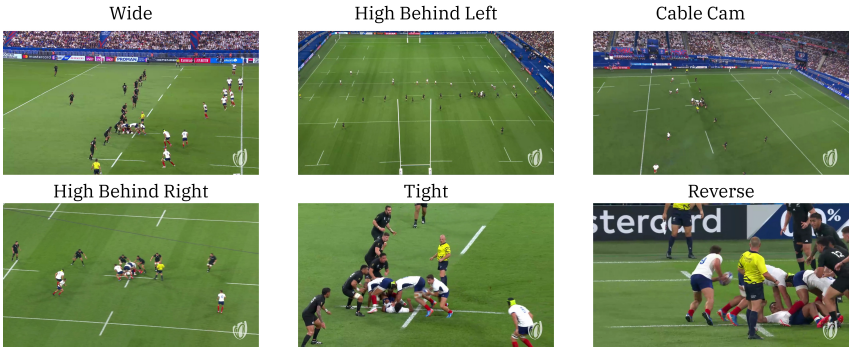


Figure 6: **Camera views in the RugbyBench dataset.** Example images from each of the six broadcast camera viewpoints.

**Regularisation losses.** The regularisation term described in Section 3 is $\mathcal{L}_{\text{reg}} = \mathcal{L}_{\Omega} + \mathcal{L}_{\text{cons}}$. $\mathcal{L}_{\Omega}$ is the point loss of VGGT-Ω (Wang et al., 2026), restricted to the landmark pixels $\mathcal{U}_v$: a Huber loss between $\hat{\mathbf{P}}_v(\mathbf{u})$ and the ground-truth landmark $\mathbf{X}_v(\mathbf{u})$. $\mathcal{L}_{\text{cons}}$ encourages the predicted 3D points of corresponding pixels to agree across views. Using the ground-truth correspondences $\mathbf{u}_{v\to v'}$ of 96 sampled player pixels per view, $\mathcal{Q}_{v\to v'}$, the player term is

$$\mathcal{L}_{\text{cons}}^{\text{player}} = \sum_{v\in\mathcal{V}} \sum_{v'\neq v} \sum_{\mathbf{u}\in\mathcal{Q}_{v\to v'}} \rho\Big(\big\|\hat{\mathbf{P}}_v(\mathbf{u}) - \hat{\mathbf{P}}_{v'}\big(\mathbf{u}_{v\to v'}(\mathbf{u})\big)\big\|_2\Big) + \rho\Big(\big\|\hat{\mathbf{P}}_v(\mathbf{u}) - \mathbf{X}_v(\mathbf{u})\big\|_2\Big), \quad (14)$$

where $\hat{\mathbf{P}}_v(\mathbf{u})$ is pixel $\mathbf{u}$ lifted with the predicted depth $\hat{D}_v$ and camera $\hat{\mathbf{g}}_v$, the second term anchors it to the ground-truth surface, and $\rho(d) = \min(d, 1) + 0.1\max(d-1, 0)$ softly caps residuals at 1 m. The pitch-line term $\mathcal{L}_{\text{cons}}^{\text{line}}$ has the same form over 32 sampled pitch-line pixels per view, and $\mathcal{L}_{\text{cons}} = \mathcal{L}_{\text{cons}}^{\text{player}} + \mathcal{L}_{\text{cons}}^{\text{line}}$.

**Training details.** We train for 100k steps with AdamW (learning rate $3 \times 10^{-4}$, weight decay 0.05, 3k warm-up steps, cosine decay, gradient clipping at 1.0) on 8 NVIDIA GH200 GPUs across two nodes, using 8 scenes per GPU, for a total batch size of 64 scenes (384 images) per step. The transformer has width 512, 8 heads, and 16 frame and 16 global attention blocks. The loss weights are $\lambda_{\text{camera}} = 5$, $\lambda_{\text{pitch}} = 8$, $\lambda_{\text{depth}} = 0.3$, $\lambda_{\text{3Dpoints}} = 20$, $\lambda_{\text{2Dpoints}} = 50$, $\lambda_{\text{people}} = 50$, $\lambda_{\Omega} = 0.5$ and $\lambda_{\text{cons}} = 1$. Training takes approximately 45 hours.

## C Masks vs 2D Keypoints

We compare SAM3 masks and ViTPose keypoints as alternative representations for detecting players in RugbyBench, on 100 randomly sampled annotated frames. ViTPose is run top-down with an RT-DETR person detector. A player counts as detected by ViTPose if at least half of its visible joints lie within 0.2 times the ground-truth box height of their annotated positions, and by SAM3 if a single mask contains more than half of its visible keypoints. Table 5 reports player-level precision, recall and F1 for each camera view. SAM3 achieves a higher F1 in every view, improving the overall F1 from 82.2 to 88.6.

Figure 7 illustrates this difference qualitatively. In dense interactions, keypoint-based detections are more easily disrupted by occlusion and overlapping players, whereas the SAM3 masks provide a more complete separation of the individual player instances. We therefore use masks as the primary player representation and complement them with MAMMA landmarks to recover the semantic information needed for geometric reconstruction.

| | | SAM3 masks | | | ViTPose | | |
|---|---|---|---|---|---|---|---|
| View | GT players | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| Wide | 3 212 | 88.6 | 93.3 | 90.9 | 96.2 | 79.6 | 87.1 |
| HB-L | 3 670 | 87.2 | 83.9 | 85.5 | 93.9 | 65.4 | 77.1 |
| Cable | 2 712 | 85.9 | 89.8 | 87.8 | 89.8 | 75.0 | 81.7 |
| HB-R | 1 183 | 87.6 | 97.7 | 92.4 | 84.6 | 92.3 | 88.3 |
| Tight | 1 187 | 87.7 | 99.4 | 93.2 | 79.7 | 95.1 | 86.7 |
| Rev | 821 | 73.7 | 95.5 | 83.2 | 60.9 | 87.9 | 71.9 |
| All | 12 785 | 86.3 | 91.0 | 88.6 | 87.3 | 77.7 | 82.2 |

Table 5: **SAM3 vs. ViTPose player detection.** Player-level detection performance across the six RugbyBench camera views.




Figure 7: **SAM3 vs. ViTPose.** Qualitative comparison of player detections in challenging scenes with dense interactions and occlusions.

## D Additional results

**Temporal propagation.** The EM algorithm described in Section 3.2 can be applied to propagate matches through time, where the E-step (Eq. 9) uses SAM3 tracks in place of the model affinity: $\alpha_{mp}^{(t)}$ is high when $m$ continues the track of $p$'s mask at $t-1$. Cameras $\mathbf{g}_v^{(t)}$ and 3D landmarks $\mathbf{X}_{p,k}^{(t)}$ are initialised from frame $t-1$, and a final temporal bundle adjustment minimises $\sum_{t\in\mathcal{T}} \mathcal{L}_{\mathrm{BA}}^{(t)}$ (Eq. 12) over the window $\mathcal{T}$, with the pitch plane and static camera centres shared across frames.

**Tracking results.** We propagate the matches of 100 randomly sampled frames over the following 18 frames and measure the reprojection error and the associations retained at the end of the window. The average pairwise association F1 drops from 78.8% to 73.8%, and the reprojection error increases from 133.7 px to 315 px. This increase is mostly due to a single window with an abrupt zoom change: excluding it, F1 drops from 78.8% to 74.2% and the reprojection error decreases slightly, from 133.7 px to 120.6 px. Since matches are currently propagated through tracking alone, such abrupt changes are a limitation; incorporating per-frame model predictions into the propagation is left for future work. The window-level bundle adjustment registers all frames in a common frame but does not change these statistics.

**SMPL-X Fits stats.** On a random 25% of the frames, the SMPL-X fits have a mean reprojection error of 9.5 px on the MAMMA landmarks (7.0 px in the wide views, 15.1 px in the close-ups).

## E DATASET CONSTRUCTION.

Scenes are generated on the fly, one static frame observed by the six broadcast cameras.

**Cameras and pitch.** Camera positions and fields of view are sampled per camera role from Gaussians fitted to the RugbyBench cameras, registered to the pitch with annotated pitch points. These distributions capture the typical broadcast rig, not individual scenes. A Zone of Interest (ZoI) is sampled uniformly between the try lines, and each camera is aimed at it with small random jitter. Pitch dimensions vary between 68–70 m in width and 94–100 m in length.

**Players.** In half of the scenes, the ZoI contains a physics-simulated collision of 2–6 players, taken from a precomputed bank generated with an $N$-person extension of CISyn (Anonymous, 2026a); in the other half, it contains a group of 2–5 AMASS-posed players. The remaining players (up to 40 per scene) are placed around the ZoI following player-layout statistics measured on RugbyBench, and are posed with AMASS (Mahmood et al., 2019) poses and CISyn body shapes. These statistics describe typical play in aggregate rather than individual scenes; no RugbyBench images, poses or correspondences are used for training.

**Inputs.** Each view is rendered at $1024 \times 576$ as a 6-channel image: the instance map (a random colour per player) with pitch lines drawn where unoccluded, and the MAMMA landmark map. To mimic real SAM3 and MAMMA outputs, 70% of scenes are degraded with missed, merged and split masks, contour noise, line dropout, and landmark dropout and jitter; the ground truth is left unchanged.

## F LIMITATIONS

MuPeT's evaluation relies on RugbyBench, whose 3D cameras and poses are pseudo-ground truth obtained by bundle adjustment. They inherit its depth/focal-length ambiguity and are closely related to our own optimisation stage, so absolute metric errors should be read as approximate. Our absolute errors, while substantially lower than those of prior methods, also remain large (1.43 m W-MPJPE, 107 px LOO reprojection), leaving considerable room for improvement.

The model is trained for a fixed rig of six broadcast camera roles with learned per-view embeddings and has only been evaluated on rugby. Transfer to other sports, camera layouts or numbers of views would require new synthetic data and retraining.

Since the synthetic scenes use AMASS and physics-simulated poses rather than sports-specific motion, a residual sim-to-real gap remains, which our EM refinement mitigates but does not remove. The pipeline also depends on upstream SAM3 masks and MAMMA landmarks: merged or missed player masks in dense rucks and scrums propagate to association and reconstruction errors.

Finally, MuPeT operates on single frames. Temporal consistency is only enforced post hoc through EM propagation, which occasionally diverges, and reconstructing long sequences remains computationally costly despite our bundle-adjustment speed-ups.

## G ADDITIONAL VISUALISATIONS

We provide additional visualisations in Figure 9 and Figure 8.



Figure 8: **Landmarks vs. SMPL-X.** Visualisations of landmark clouds and corresponding SMPL-X fits, from multiple viewpoints.

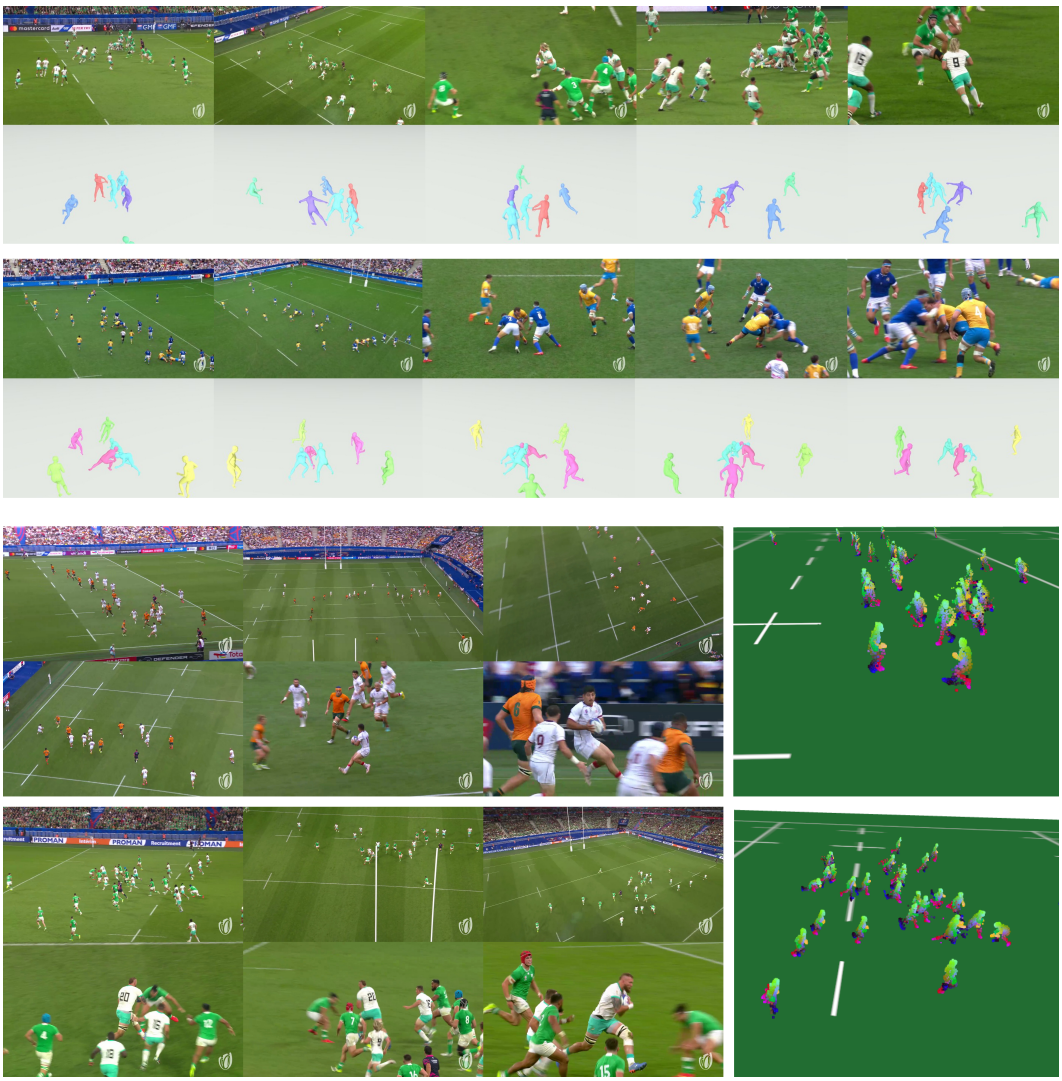

Figure 9: **Additional Visualisations.** SMPL-X fits from multiple views, and point cloud reconstruction.