Robust Camera Pose Estimation and 3D Human Reconstruction for Sports Events

A solution to the Skeletal tracking 'light' challenge

#	△	Team	Members	Score	Entries	Last	Solution
1	—	**Tim**		1.24297	15	1mo	
2	▲ 1	mil		1.31290	37	1mo	
3	▲ 1	arturxarles		1.54698	34	1mo	

Huang Jing **hj00@tju.edu.cn**

June 11, 2025

Formulation of the Challenge

For each Scene:

Input

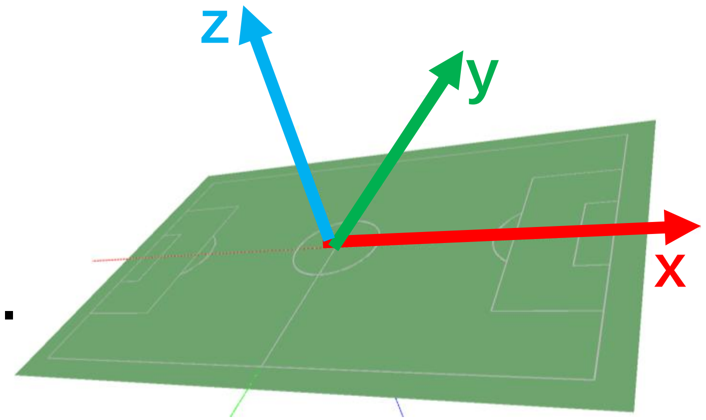
- **Video**
- **Camera Intrinsics of each frames,** $\boldsymbol{K} \in \mathbb{R}^{N\times3\times3}$**,** $\boldsymbol{D} \in \mathbb{R}^{N\times5}$

  N is the number of frames, about 500~3000

  Distortion at OpenCV Format, k_1, k_2, p_1, p_2, k_3
- **Camera extrinsic matrices of the first frame**

$$P_0 = \begin{bmatrix} R_0 & T_0 \\ O & 1 \end{bmatrix} \in \mathbb{R}^{4\times4}$$

- **Tracking bounding boxes**
- **A prior: the court in world space is known.**



Output

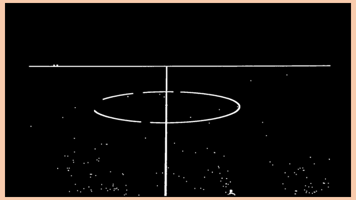
- **3D Joints** corresponding to the given b-boxes **in world space**

Our solution

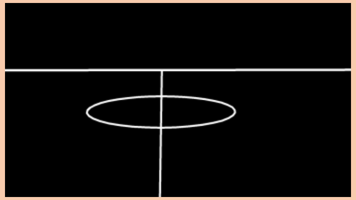
Camera Pose Estimation → **Per-frame SMPL[1] Estimation** → **Smoothing the 3D Trajectories**

- In the spirits of the Baseline*, for each frame： $\underset{P}{\operatorname{argmin}}\, L(I_{\text{detect}}, I_{\text{reproj}})$

Line Detector



I_{detect}

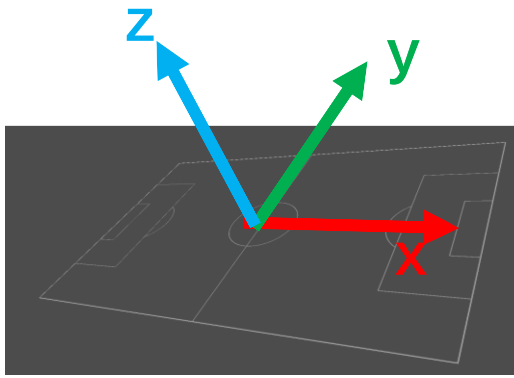


I_{reproj}

To solve P to minimal the difference

Projector

K P



3D court lines in world space

$L(a, b)$ something like $\|a - b\|^2$

$P \in \mathbb{R}^{4\times4}$ camera extrinsic

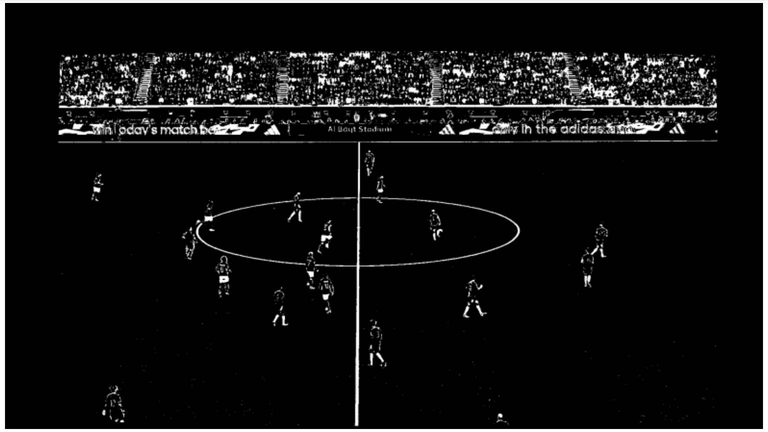
* Baseline: https://github.com/G3P-Workshop/Skeletal-Tracking-Starter-Kit by Tianjian

[1] Matthew, et al. SMPL: A Skinned Multi-Person Linear Model. SIGGRAPH-A. 2015.

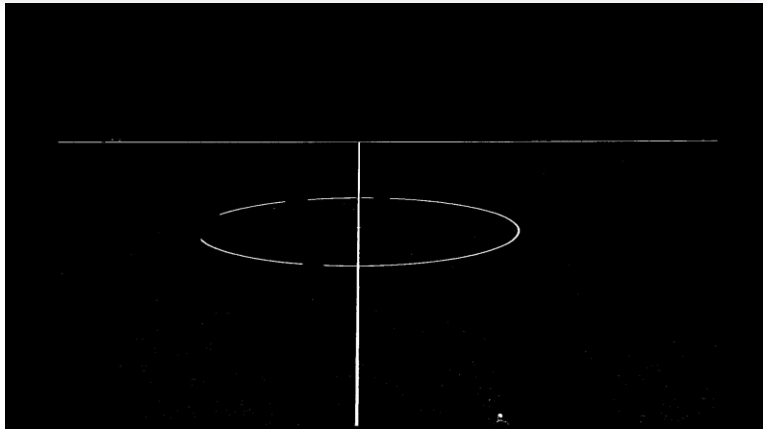
Camera Pose Estimation: **Line Detector**

A set of tradition method.

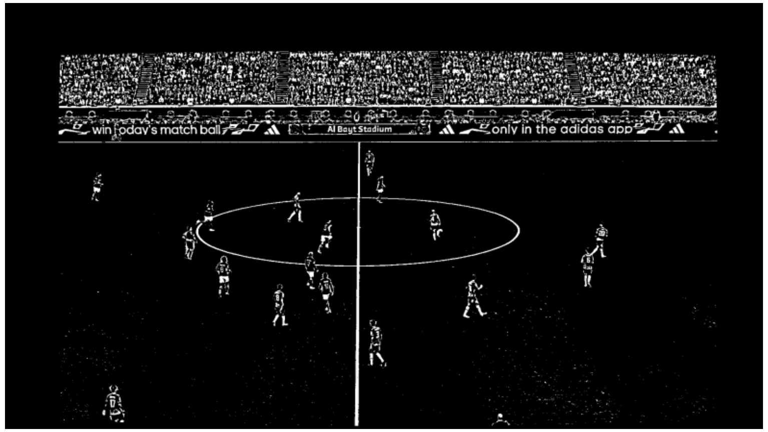
Input Frame



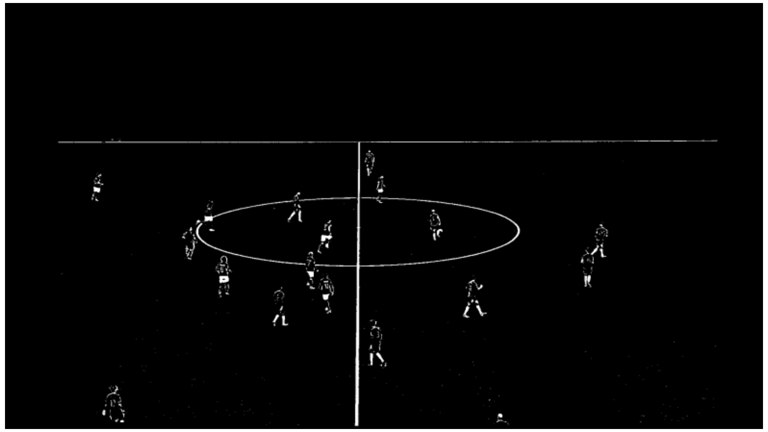
②remove none-white pixels



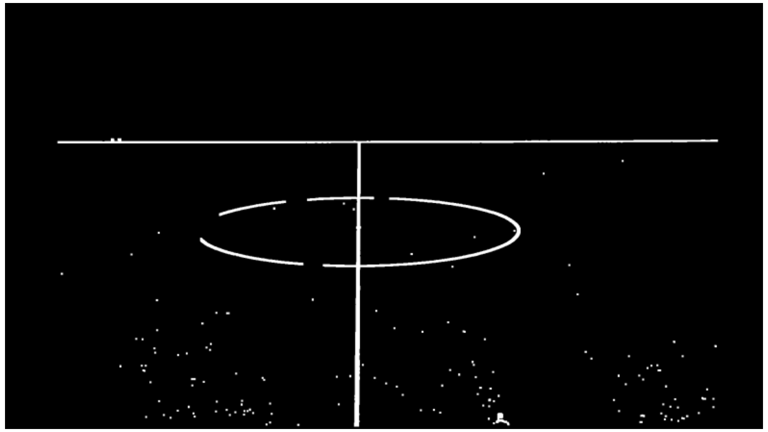
④remove given bounding boxes



①Adaptive Thresholding
(This also used by the baseline)



③remove out-of-court pixels



⑤make lines thinker (dilate)

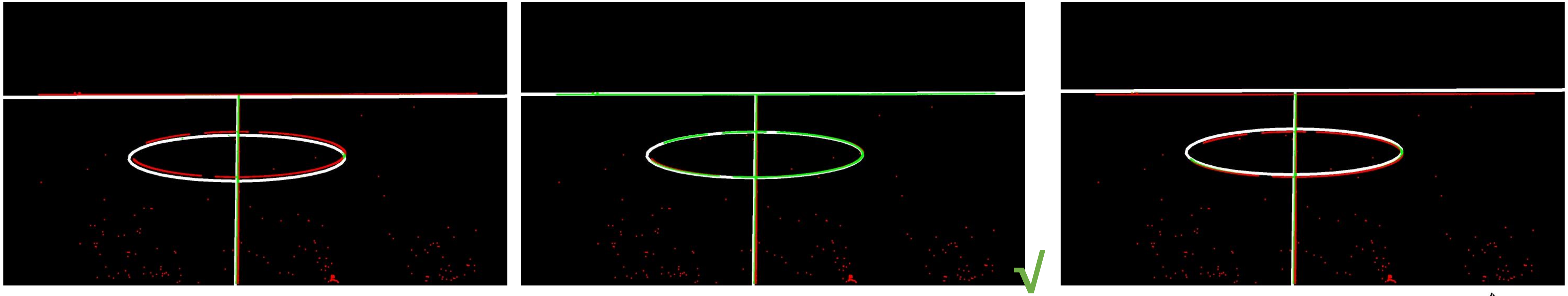
Camera Pose Estimation: **Optimization Algorithm**

We notice that:

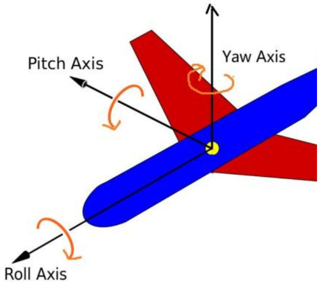
Fixed position, rotation only

- The camera pose can be only 3DoF.
- The Euler angles of the camera pose changes less than 0.5° within a frame.
- 0.05° difference is enough to measure the camera pose accuracy.

We can use a naïve **searching** approach to try these $\mathbf{10^3}$ possible combination.



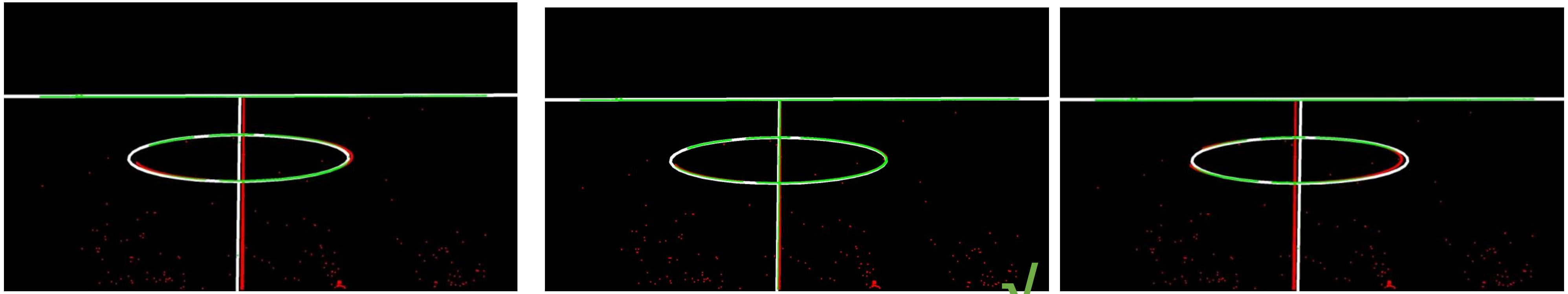
An example of try different pitch angle, it's easy to find a best IoU case.



Red: I_{detect} Green: $I_{detect} \cup I_{reproj}$ White: I_{reproj}

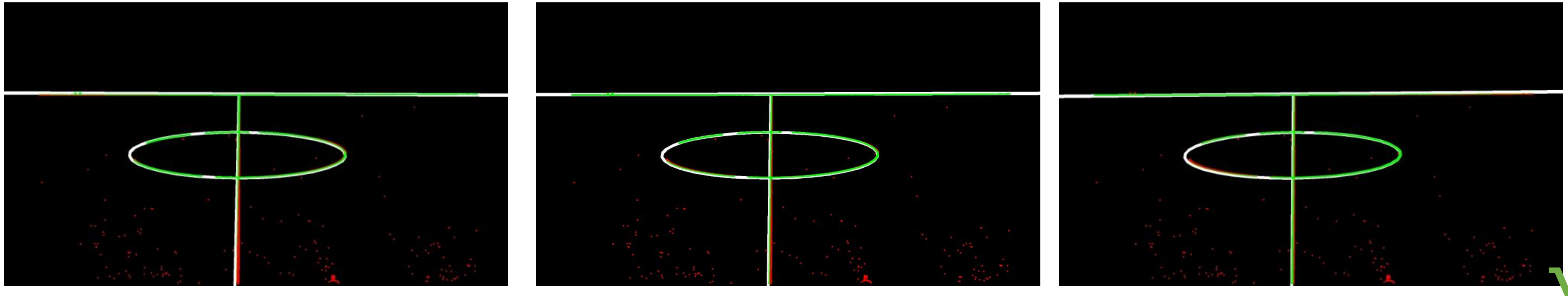
Camera Pose Estimation: **Optimization Algorithm**

We can use a naïve **searching** approach to try these $\mathbf{10^3}$ possible combination.



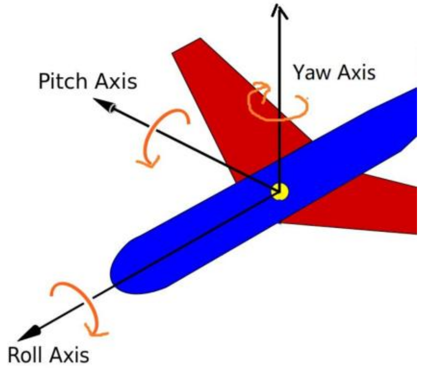
√

An example of try different yaw angles



√

An example of try different roll angles



Red: $\boldsymbol{I_{detect}}$ **Green: $\boldsymbol{I_{detect} \cup I_{reproj}}$** White: I_{reproj}

Camera Pose Estimation: **Optimization Algorithm**

We use a naïve **searching** approach.

- We can **project points and draw lines** rather than rendering to **accelerate** it since the process does not need to be differentiable.
- Generally, this method is **robust** and **relatively fast**. (and easy to write)

Not sensitive to the searching ranges and steps [−0.385°, 0.385°] with 10 steps [-0.5, 0.5] with 7 steps [-0.75, 75] with 10 steps Either of the above searching settings works well on **all** 13 videos.	All 13 video (600-2300 frames) can be done within 2.5 hours on a server with an AMD EPYC 7543 CPU and no GPU. ≈ **5FPS** with multi-processing

Camera Pose Estimation

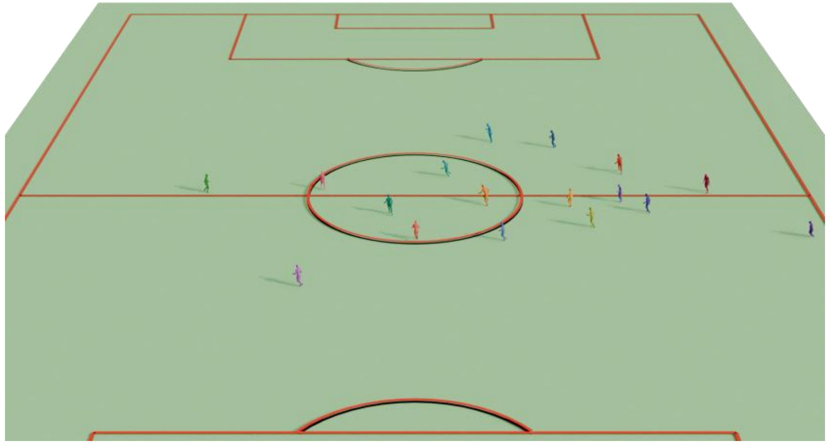
- Reprojection visualization.

- We didn’t have time to test if a differentiable rendering with an optimizer is better.

Our solution

Step1: Camera Pose Estimation

Step2: Per-frame SMPL Estimation: RCR



Frame

Optional **Ground Plane (camera space)**

Optional **Camera Intrinsic**

→ **RCR[1] (Robust Crowd Reconstruction)** →

SMPL[2] Reconstruction in Camera Space

[1] Huang, et al. RCR: Robust crowd reconstruction with up-right space from a single large-scene image. arXiv 2411.06232. 2025.

[2] Matthew, et al. SMPL: A Skinned Multi-Person Linear Model. SIGGRAPH-A. 2015.

Our solution

Step1: Camera Pose Estimation

Step2: Per-frame SMPL Estimation: RCR

RCR: a two-stage method which can also be feed with the ground-truth bounding-boxes given by the Challenge.

Huang, et al. RCR: Robust crowd reconstruction with up-right space from a single large-scene image. arXiv 2411.06232. 2025.

Per-frame SMPL Reconstruction via **RCR**

Key idea 1/2: HVIP concept to estimate the 3D location

To Solve the problem:

- When the 2D body center $\boldsymbol{p_c} \in \mathbb{R}^2$ on the image and the camera intrinsic matrix $\boldsymbol{K} \in \mathbb{R}^{3\times3}$ are known, how to estimate the 3D body center $\boldsymbol{P_c} \in \mathbb{R}^3$?

To estimate the depth?

- Estimate the depth $d \in R$ and use the reverse projection.

$$\boldsymbol{P_c = K^{-1} * F_{homo}(p_c) * d}$$

To estimate the 2D "HVIP"

- Additionally given the ground $\boldsymbol{Ax+By+Cz+D=0}$, $\boldsymbol{A, B, C, D} \in \mathbb{R}$
- HVIP (Human-scene virtual interaction points) is the **3D projection of the body's center onto the ground** along the ground normal direction.

Therefore, we can get P_c by **estimating the 2D HVIP** $p_h \in \mathbb{R}^2$ on the image.

$$\begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} = K \qquad (u_h, v_h) = p_h \qquad (u_c, v_c) = p_c \quad \Rightarrow \quad d = \frac{\left(f_x y_h - z_t(v_c - c_y)\right)}{C(v_c - c_y) - Bf_y}$$

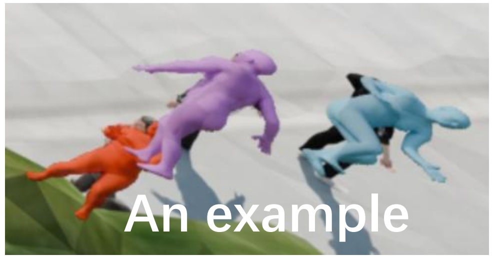
Huang, et al. RCR: Robust crowd reconstruction with up-right space from a single large-scene image. arXiv 2411.06232. 2025.

Per-frame SMPL Reconstruction via **RCR**

Key idea 2/2: Canonical Regression Space

To Solve the problem:

- Simply translate the reconstructed SMPL to the target estimated positions will cause error reprojection and 3D pose inaccuracy.



Torso Center

Human-scene Virtual Interaction Point (HVIP)

angular resolution per pixel

We noticed that:

Perspective distortion varies significantly across the image and with the camera intrinsics, but remains approximately constant within a small local region.

Therefore:

- We define a canonical space **to eliminate these perspective distortion variation.**
- SMPL and 2D HVIP are regressed in the canonical space.

Quick summary of **RCR**

Frame

Optional — Ground Plane

Optional — Camera Intrinsic

→ **RCR (Robust Crowd Reconstruction)** → **SMPL Reconstruction in Camera Space**

Key idea 1/2: HVIP concept to estimate the 3D location

HVIP concept and the **explicit ground plane modeling** provide **spatial consistency** of different bounding boxes and frames.

Key idea 2/2: Canonical Regression Space

2D HVIP and SMPL are estimated in a canonical regression space so that we can ensure the **reprojection accuracy**, further slightly improve the **3D accuracy**.

Other Features

- Support single frame input (estimate the camera and ground parameters automatically).
- Support any FoV (Field of View) **without** any test-time optimization.

Huang, et al. RCR: Robust crowd reconstruction with up-right space from a single large-scene image. arXiv 2411.06232. 2025.

Post-process

- Recover the estimated SMPL to the world space by the camera pose.
- Smooth the 3D positions sequence of each person by removing outliers, interpolating the outliers, and filtering.

Experiments

Visualization of the reconstructed SMPLs.

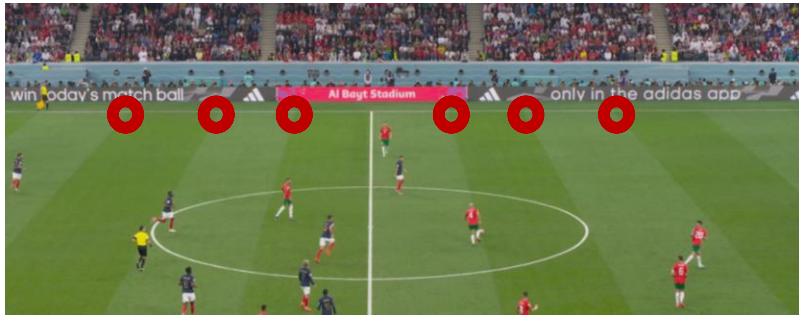
Reprojection

World Space

Discussion

For Camera Pose Estimation:

- The simple and robust camera pose estimation may be helpful for the next year's challenge. (new baseline?)
- These points with ID are hard to recognize so that COLMAP[1] is not easy to work nicely.



- End-to-end camera pose (like VGGT[2]) estimation methods are hard to add constraints (e.g., a fixed camera position).

For G3P (Global 3D Human Pose):

- An explicit scene modeling may be good idea.
- Currently, it seems that some methods are balancing the reprojection (our RCR) and the reasonable 3D pose sequence (GVHMR, PHC). But I believe it could be finally solved by achieving the accurate world space poses. At that time, we don't need to strike this balance.

[1] COMAP, https://github.com/colmap/colmap
[2] Wang et al. VGGT: Visual Geometry Grounded Transformer. CVPR 2025.
[3] Shen et al. GVHMR: World-Grounded Human Motion Recovery via Gravity-View Coordinates. SIGGRAPH-A 2024.
[4] Luo et al. Perpetual Humanoid Control for Real-time Simulated Avatars. ICCV 2023.

Thank my teammates, my supervisor and the organizers.

Jing Huang[1]
hj00@tju.edu.cn

Hanrong Zhuang[1]
zhr_2021@tju.edu.cn

Lin Zhang[1]
Zhanglin_just@tju.edu.cn

Yuxiang Liu[1]
lyx1021@tju.edu.cn

Prof. Kun Li[1,*]
lik@tju.edu.cn

Lab's home page: https://cic.tju.edu.cn/faculty/likun/index.html

1: Tianjin University, Tianjin China.
*: Supervisor