PACE: Human and Camera Motion Estimation from in-the-wild Videos

Muhammed Kocabas[1,2,3], Ye Yuan[1], Pavlo Molchanov[1], Yunrong Guo[1], Michael J. Black[2], Otmar Hilliges[3], Jan Kautz[1], Umar Iqbal[1]

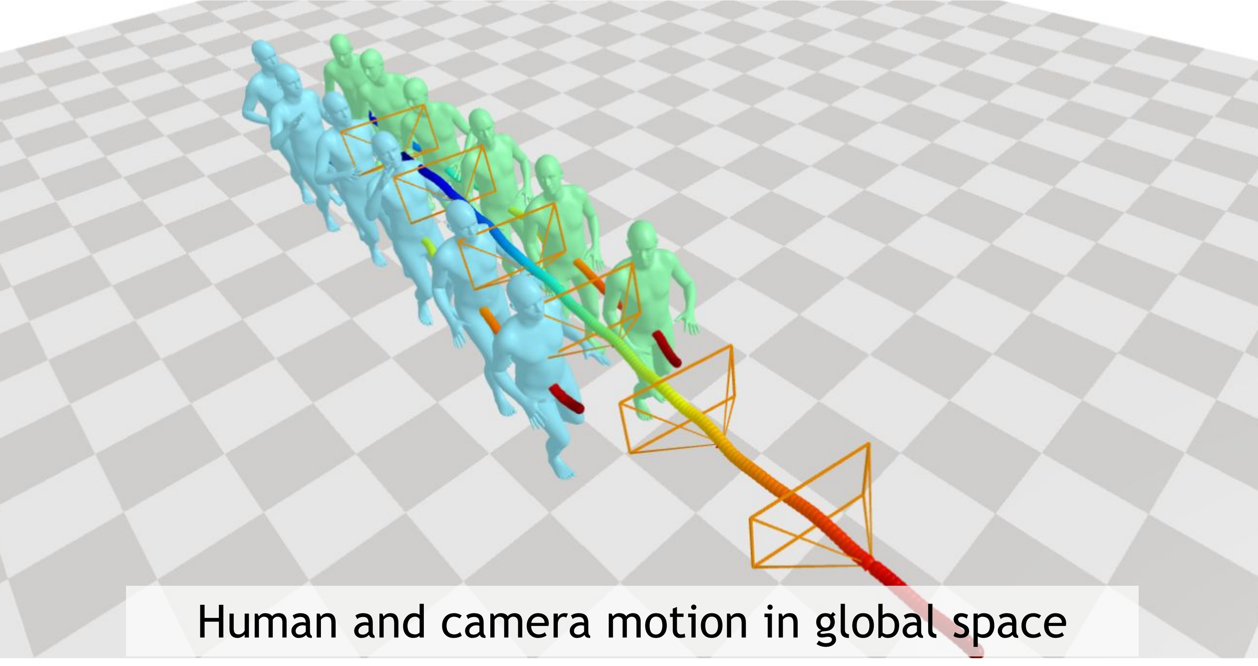
1 NVIDIA 2 MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS 3 ETH zürich

nvlabs.github.io/PACE/

3DV 2024

Goal

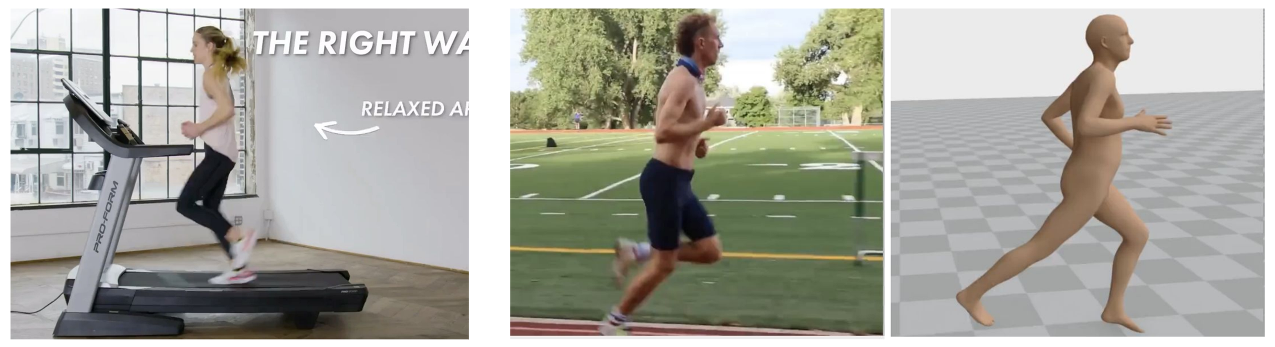

Input video


Human and camera motion in global space

- **Input:** RGB video of multiple people captured with moving camera
- **Output:** Motion of humans and camera in the same global coordinate frame along with a ground plane.

Problem

- Existing works, e.g. GLAMR [1], only rely on local body motion to estimate global motion. However, there are several issues with mapping from local motion to global trajectories:
  - mapping is ambiguous especially under root rotations,
  - local body motion estimation can be erroneous,
  - people can also move in-place.

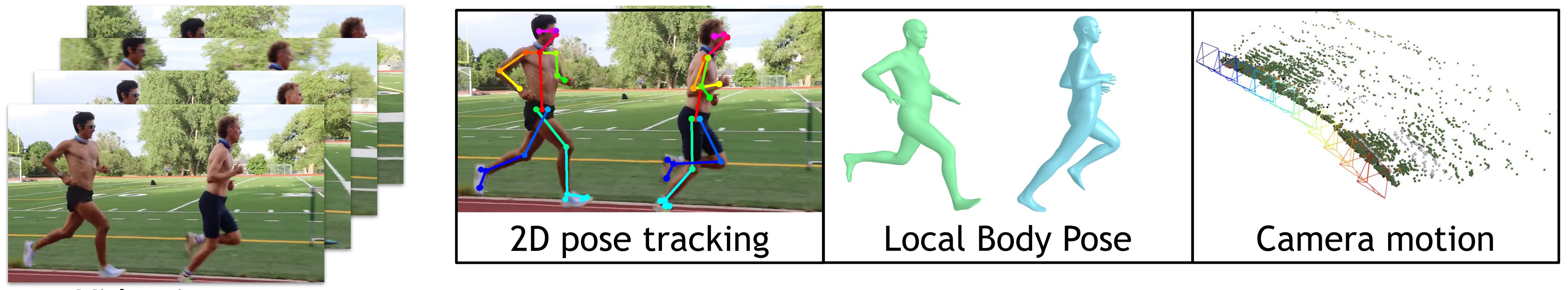


In-place movement Erroneous local body pose estimation

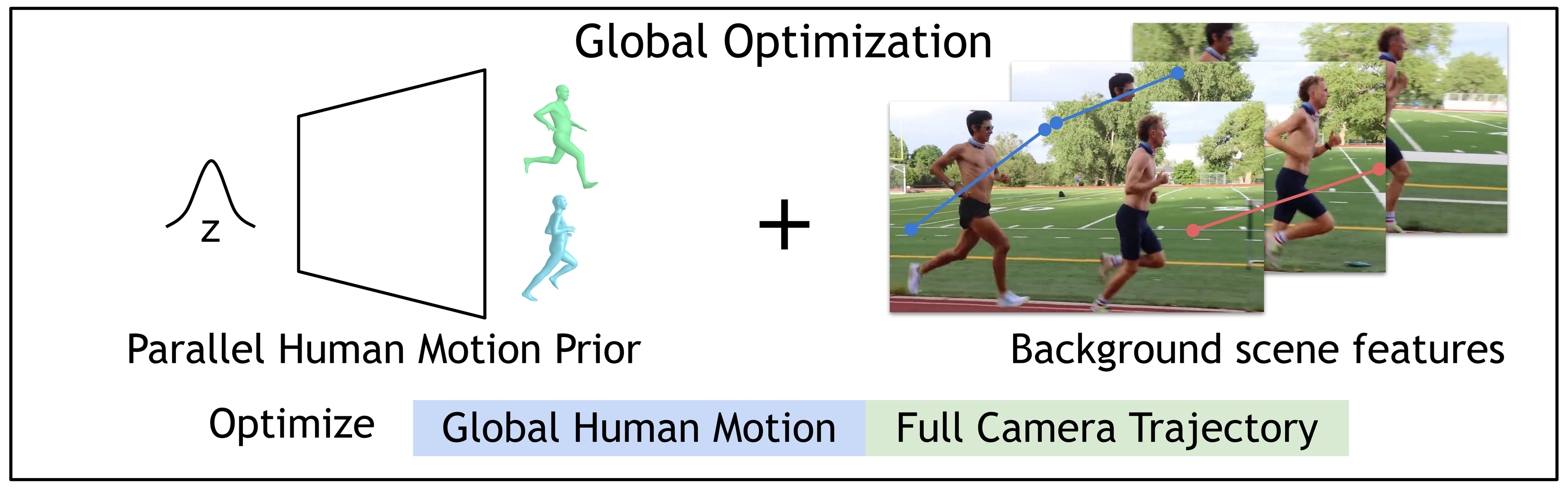
- Other methods, e.g. SLAHMR [2], rely on SLAM to jointly recover human and camera motion. However, SLAM algorithms:
  - provide camera motion up to a scale
  - assume a static scene and suffer when there are dynamic objects


Human motion breaks the static scene assumption

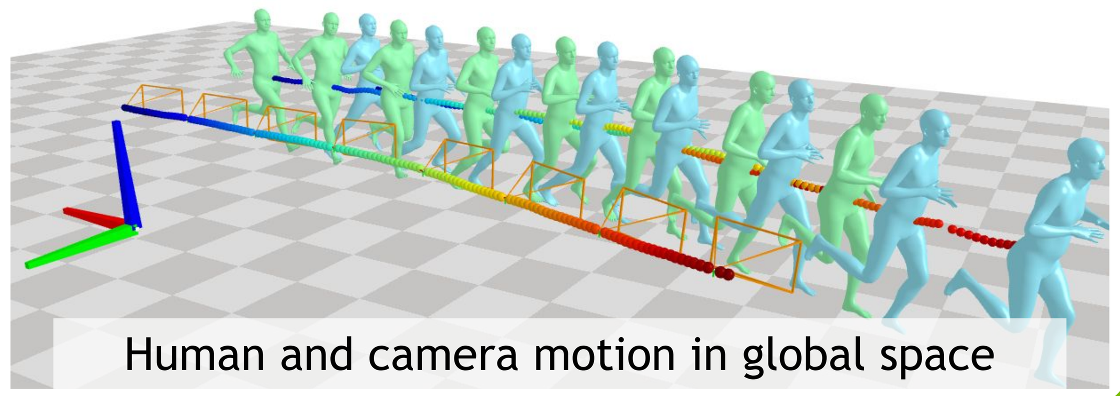
Method



- Given a video with dynamic human and camera motions, we first use off-the-shelf methods to obtain initial 2D human pose, 3D human motion, and camera motions.



- We propose a unified optimization framework that optimizes the global human motions and full camera trajectories to reduce 2D pose errors, increase motion likelihood under human motion prior, and match background features.
- The final output is coherent human and camera motion in global space.


Human and camera motion in global space

Objectives

$$\min_{\substack{\beta,\mathbf{z}\\ s,h_0,R_0,\{R_t,T_t\}_{t=1}^{T}}} E_{\text{body}} + E_{\text{scene}} + E_{\text{camera}}$$

- E_{body}: ensures that the reconstructed human motion is plausible and agrees with the image evidence.
- E_{scene}: ensures that joints with floor contact have 0 velocity and they are above ground.
- E_{camera}: ensures that the reconstructed camera motion is smooth and consistent with the static scene motion

$$E_{\text{body}} = E_{\text{2D}} + E_{\beta} + E_{\text{pose}} + E^{\text{b}}_{\text{smooth}} + E_{\text{VAE}} + E_{\text{consist}},$$

2D reprojection loss (E_{2D}), SMPL losses ($E_{\beta} + E_{\text{pose}}$), Temporal smoothness ($E^{\text{b}}_{\text{smooth}}$), Motion prior loss ($E_{\text{VAE}}$), Batch consistency ($E_{\text{consist}}$)

$$E_{\text{scene}} = E_{\text{contact}} + E_{\text{height}}$$

Contact velocity (E_{contact}), Joints above ground (E_{height})

$$E_{\text{camera}} = E_{\text{PCL}} + E^{\text{c}}_{\text{smooth}}$$

Static scene loss (E_{PCL}), Camera smoothness ($E^{\text{c}}_{\text{smooth}}$)

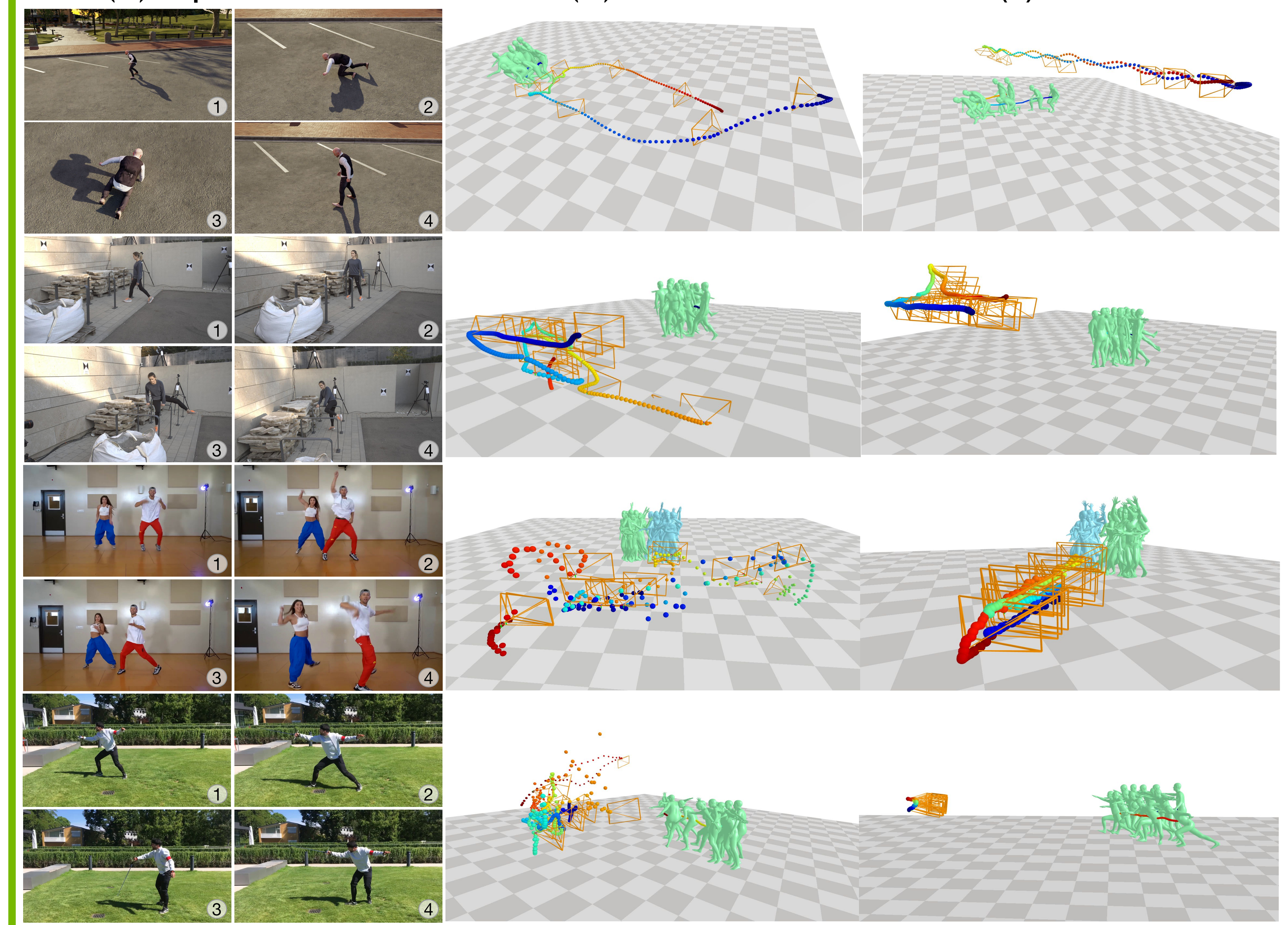
HCM dataset

- Current datasets lack annotation of both human and camera parameters.
- HCM is a synthetic dataset that has human body and camera motion annotations
- **Humans:** Renderpeople
- **Scenes:** Unreal engine
- **Motion:** AMASS [3] dataset

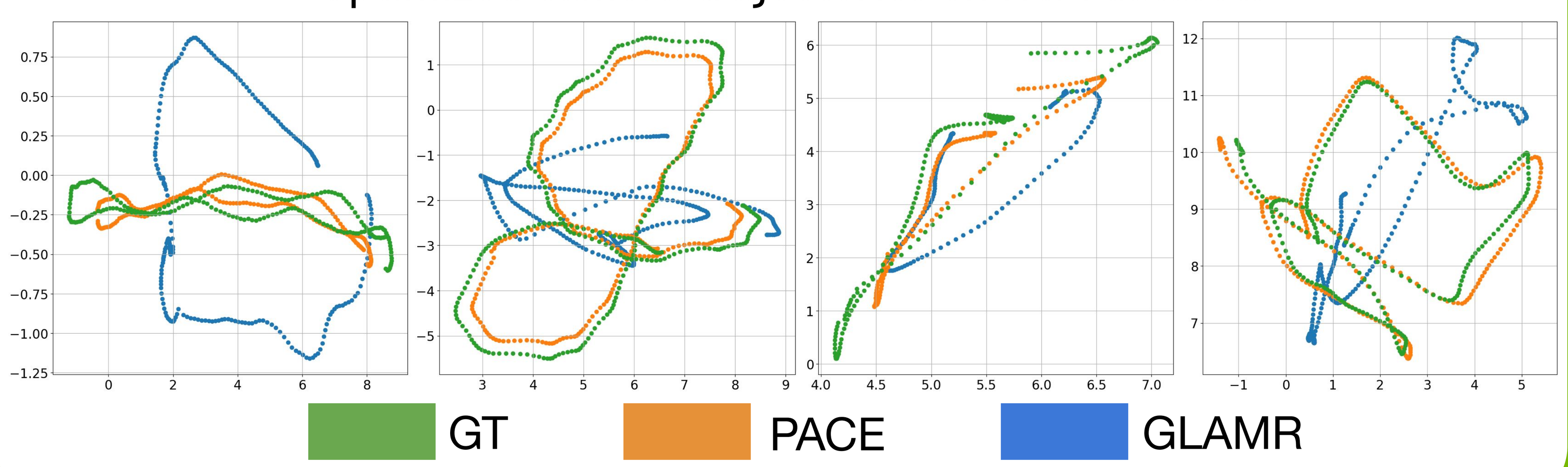
Results

Results of GLAMR and PACE on various in-the-wild videos

(a) Input videos (b) GLAMR (c) PACE



Optimized camera trajectories on HCM datasets



References

1. Yuan et al., GLAMR: Global Occlusion-Aware Human Mesh Recovery with Dynamic Cameras, CVPR 2022.
2. Ye et al., Decoupling Human and Camera Motion from Videos in the Wild, CVPR 2023.
3. Mahmood et al., AMASS: Archive of Motion Capture As Surface Shapes, ICCV 2019.

ArXiv: https://arxiv.org/abs/2310.13768 **Project page:** https://nvlabs.github.io/PACE/