3D VISION & ROBOTICS LAB

UNIST
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Lab Seminar
From Single-Image HMR to World-Grounded Video HMR

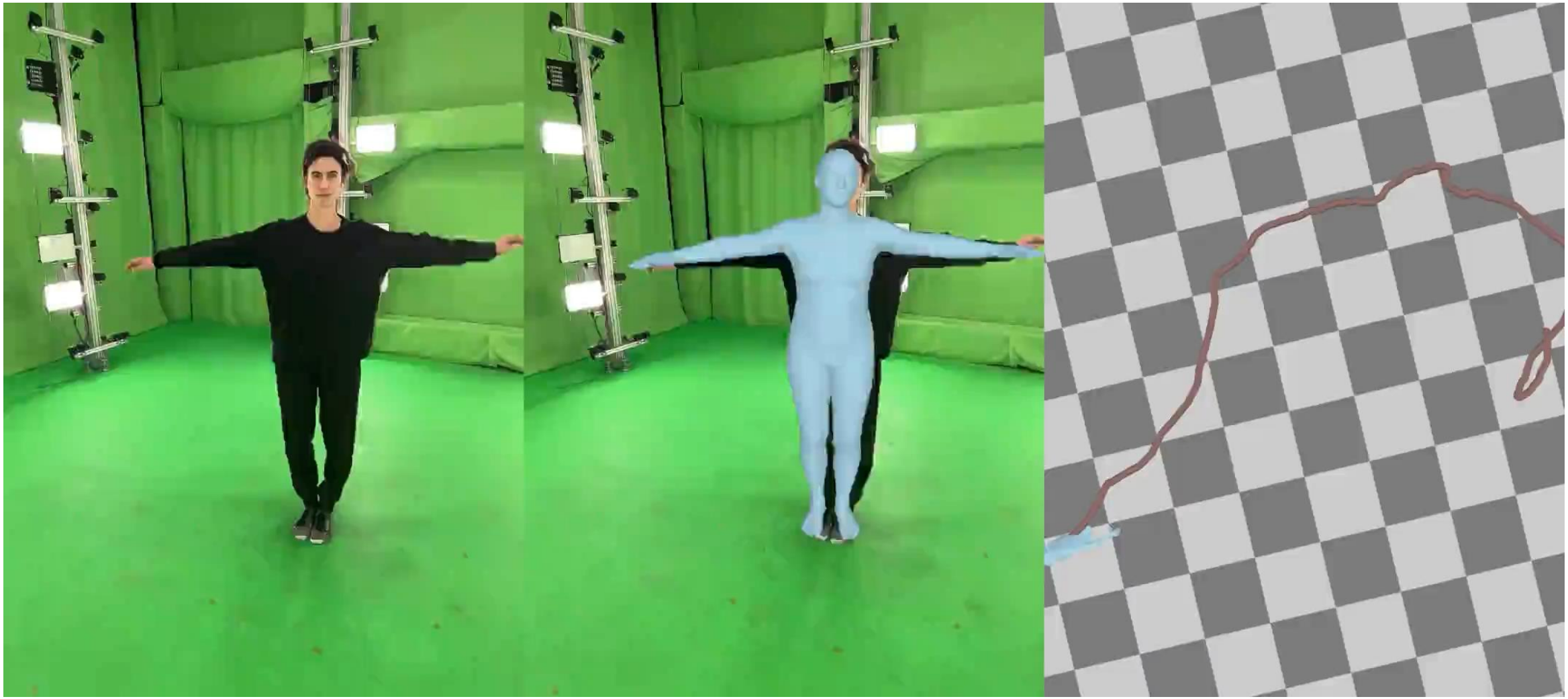
Presenter: Gyeongsu Cho

Wed Oct 1, 2025

Contents

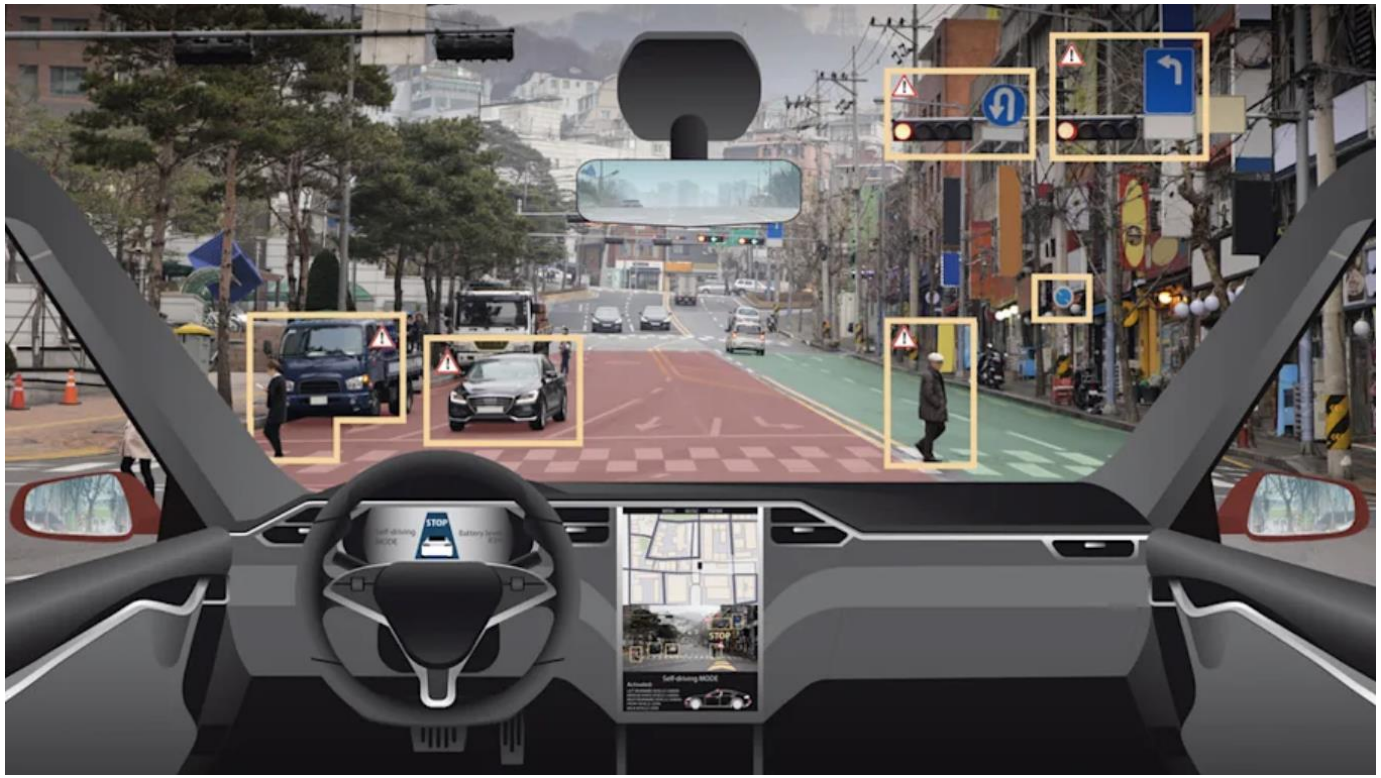
Introduction

Introduction



What is World-Grounded Video HMR?

Introduction



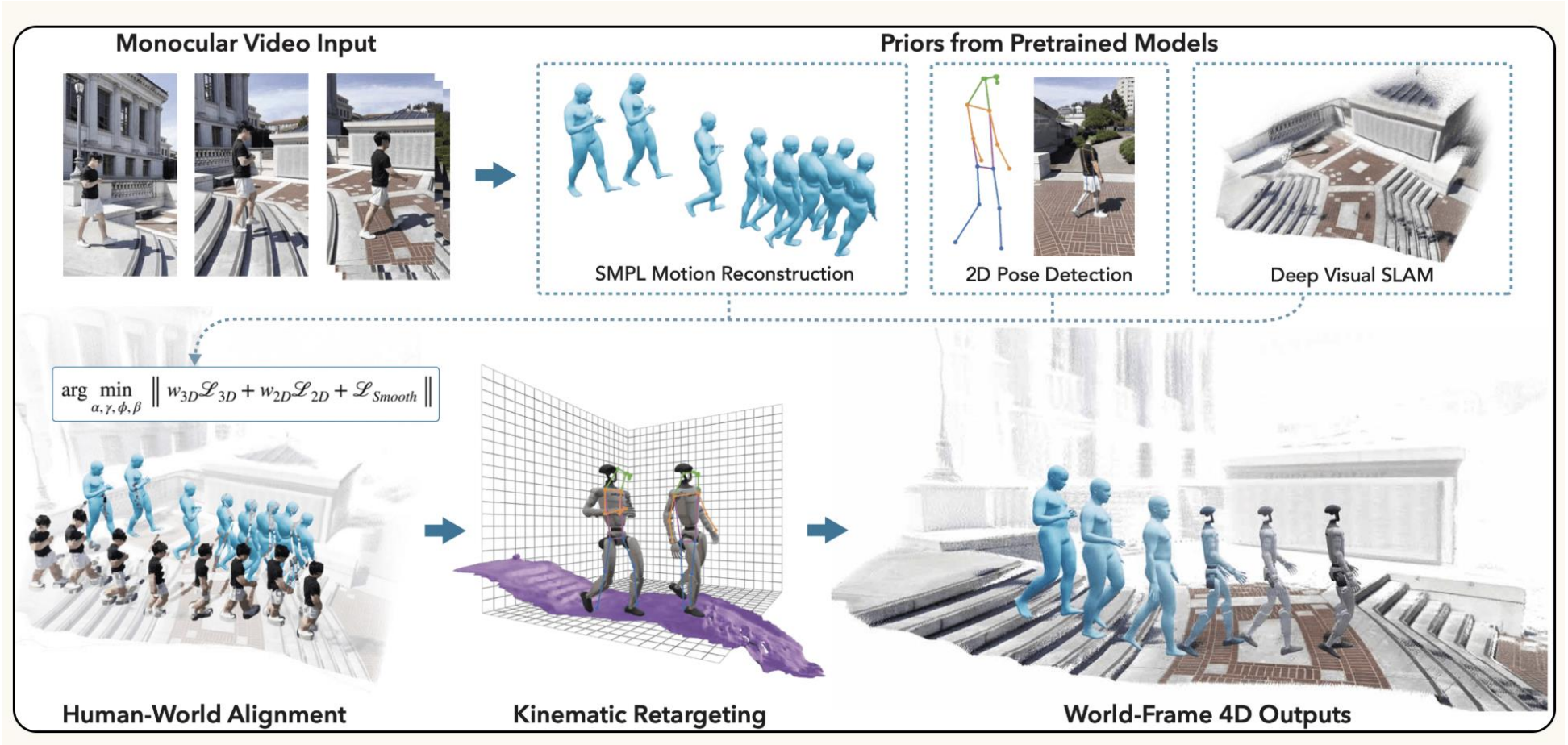
Perceiving humans in self-driving scenarios is important

Introduction

[arXiv 2025] VideoMimic

World-Grounded Video HMR helps to train humanoid robots. (real-to-sim)

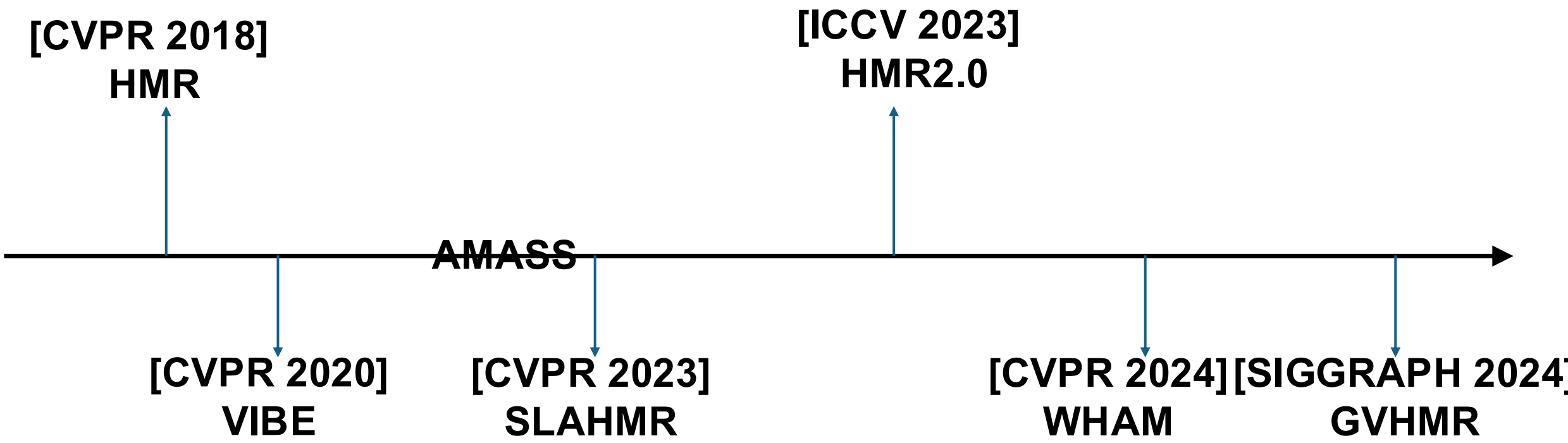
Introduction



[arXiv 2025] VideoMimic

World-Grounded Video HMR helps to train humanoid robots. (real-to-sim)

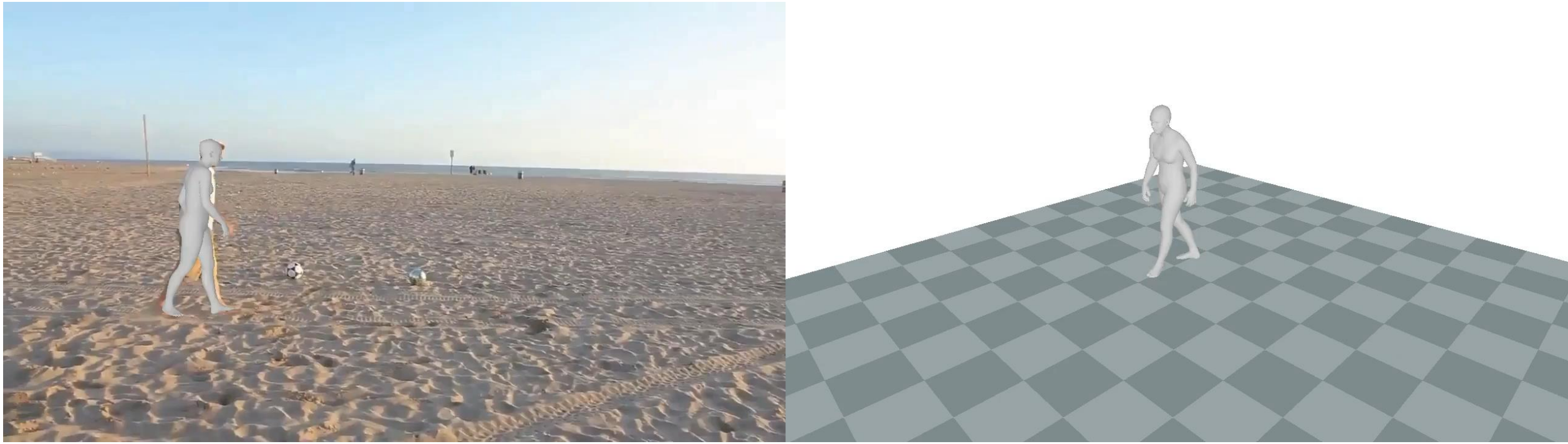
Introduction



History of human mesh recovery

Introduction

Goal of this presentation

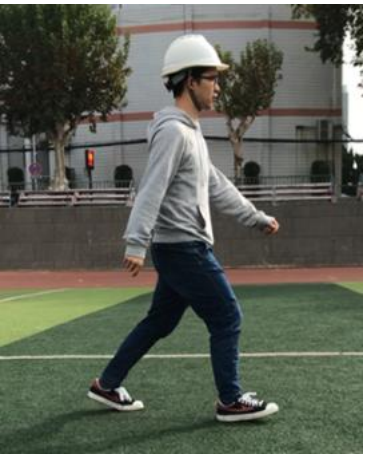


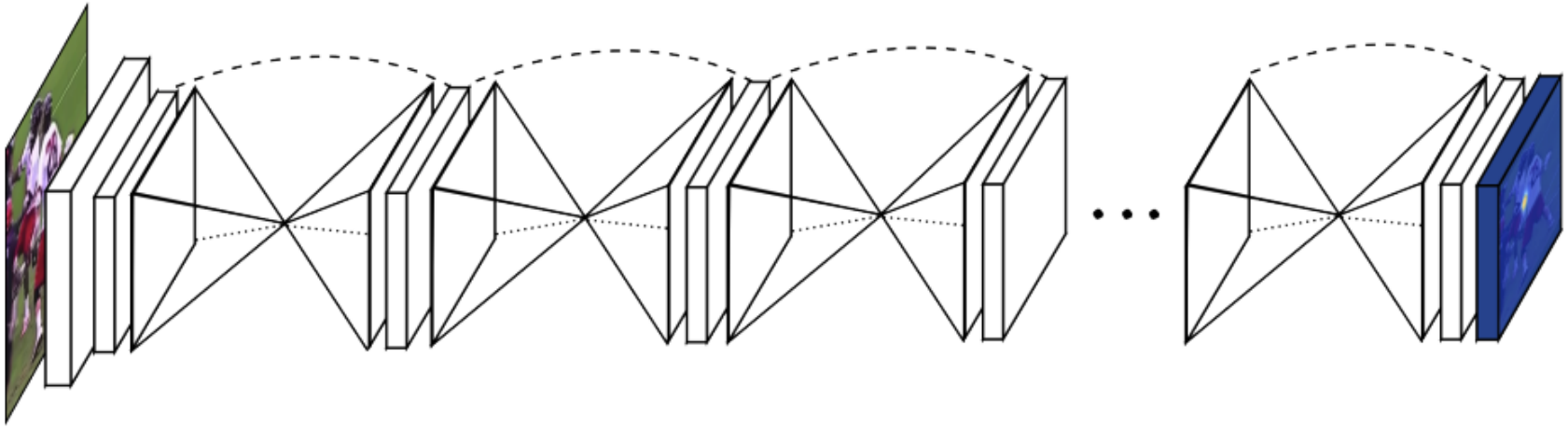
[SIGGRAPH Asia 2024] GVHMR

Understanding GVHMR.

Per-frame Human Mesh Recovery: HMR

Motivation of HMR

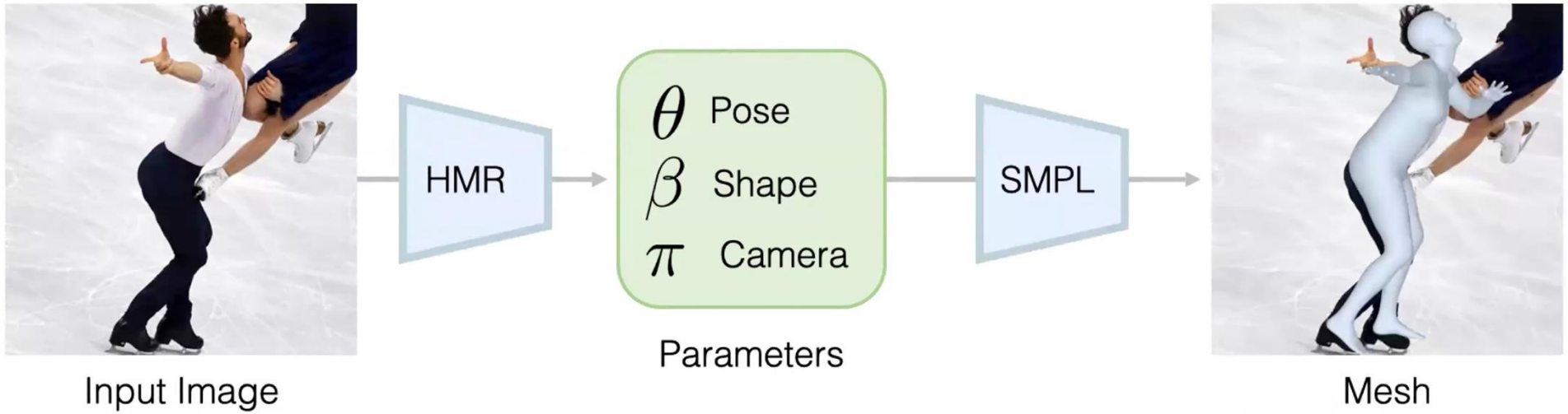




[ECCV 2016] Stacked Hourglass Networks for Human Pose Estimation

Traditional Human Pose Estimation (HPE) methods regress human keypoints.

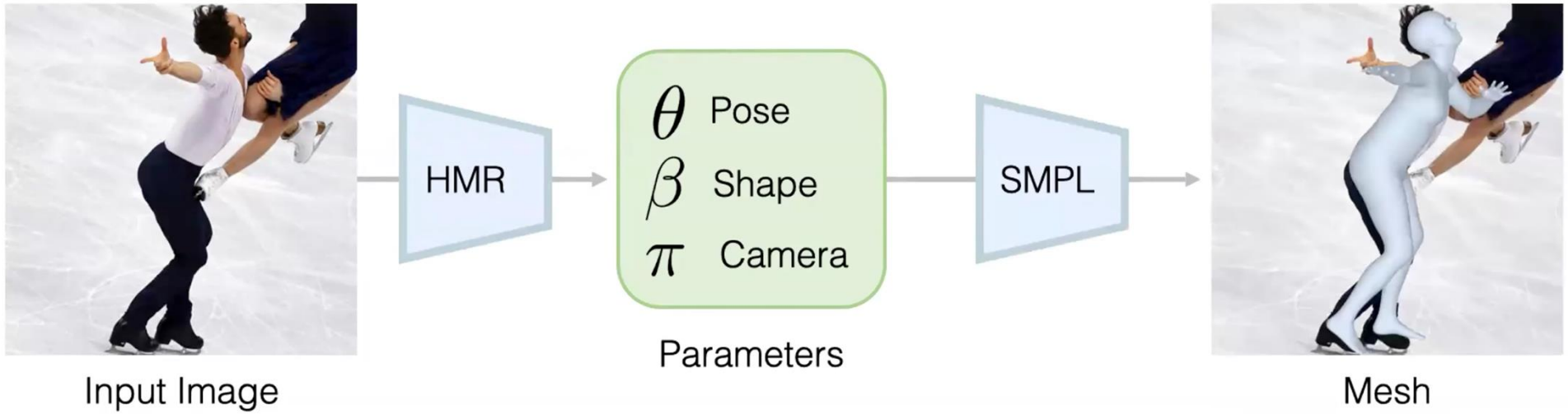
Per-frame Human Mesh Recovery: HMR

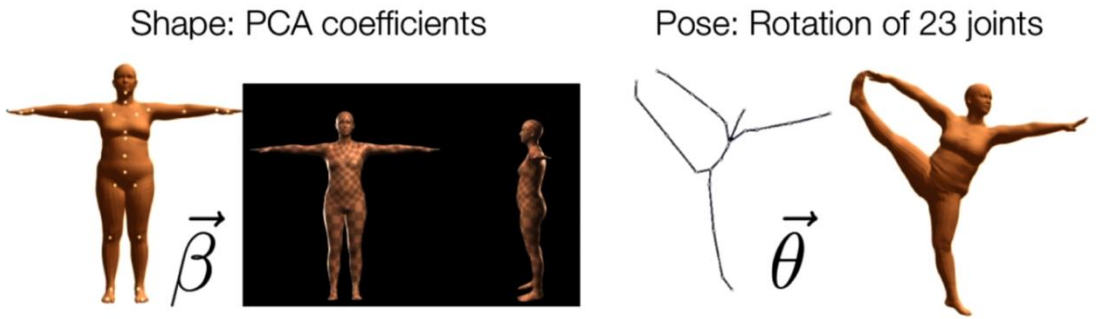


[CVPR 2018] HMR

HMR leverages the SMPL parameters.

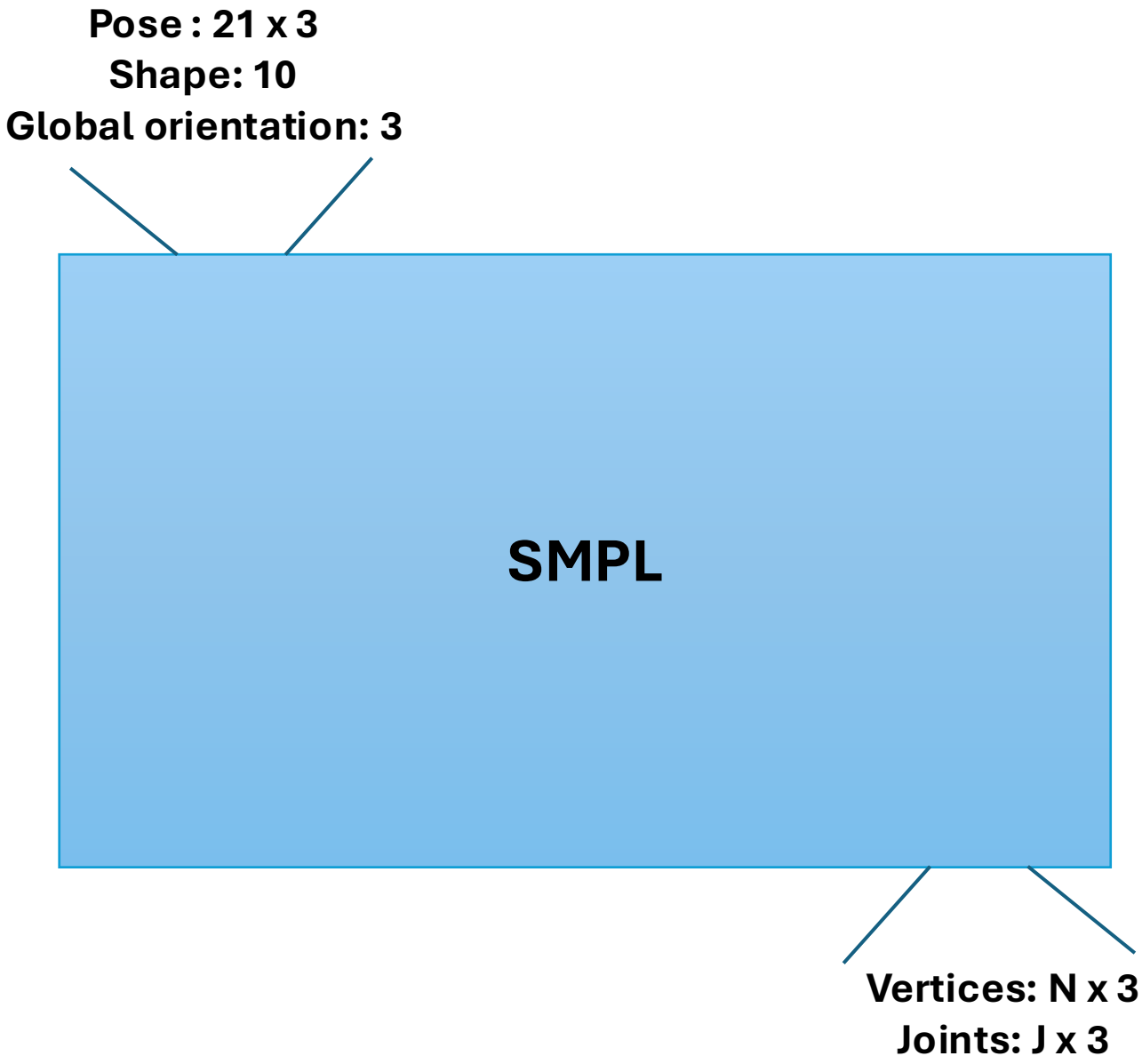
Per-frame Human Mesh Recovery: HMR





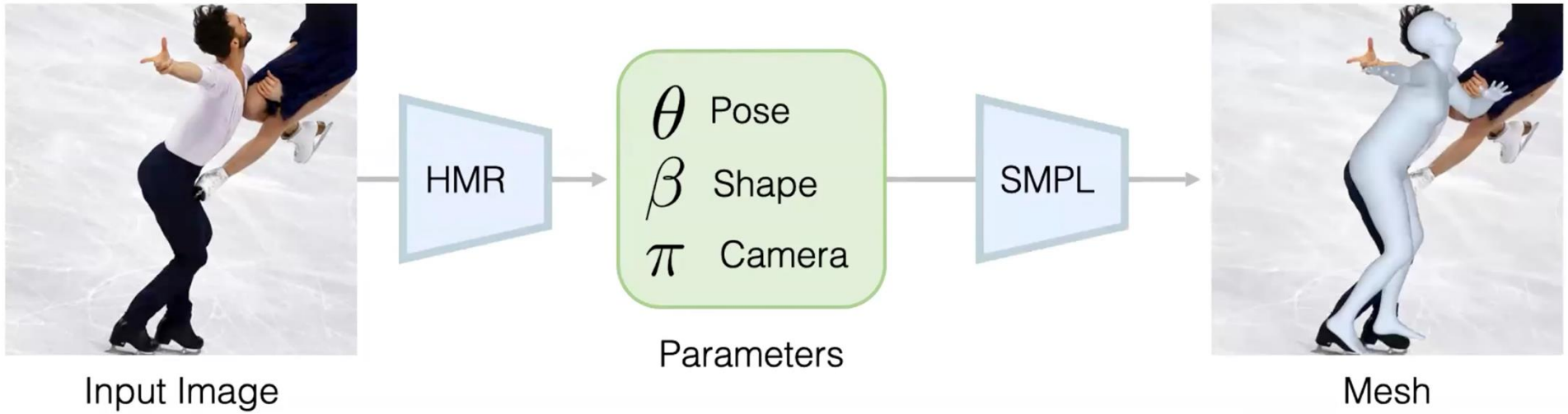
HMR leverages the SMPL parameters.

SMPL



The SMPL model is a parametric 3D human body model that represents realistic human shape and pose with a low-dimensional set of parameters

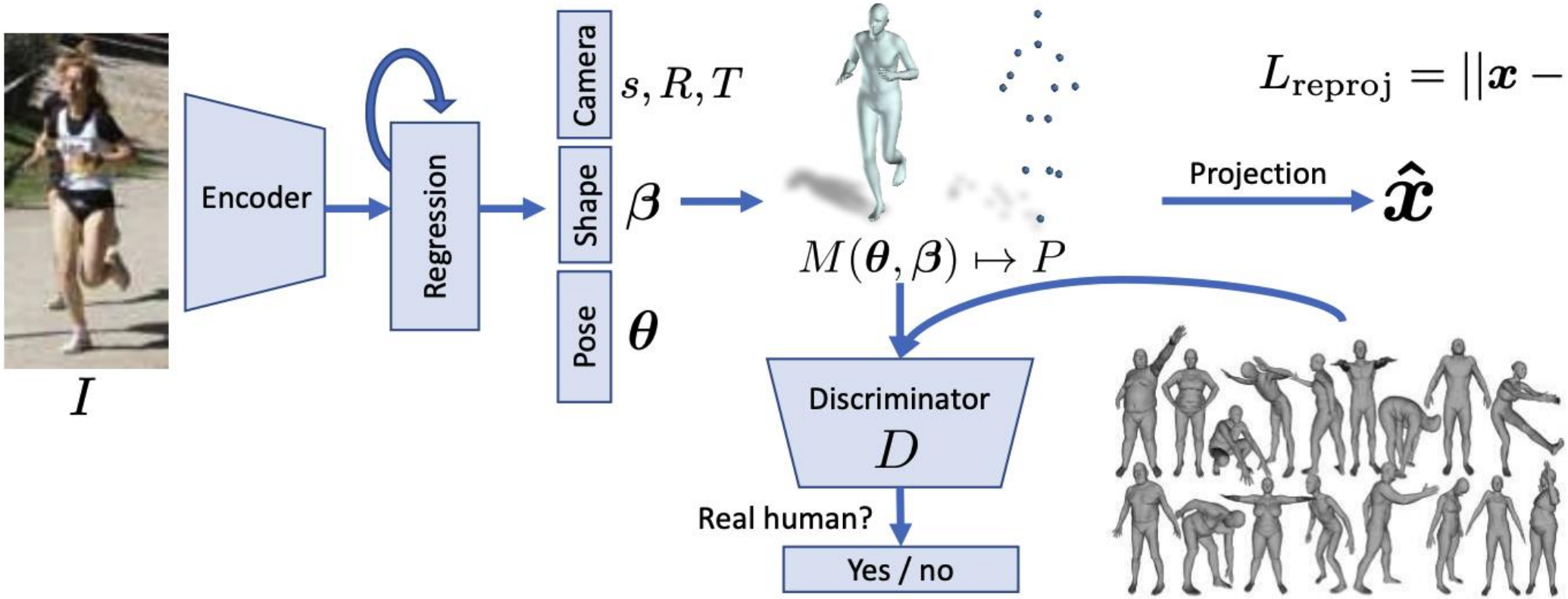
Per-frame Human Mesh Recovery: HMR



$$L_{3D} = L_{3D\ joints} + L_{3D\ smpl}$$
$$L_{joints} = ||(\mathbf{X_i} - \hat{\mathbf{X}}_\mathbf{i})||_2^2$$
$$L_{smpl} = ||[\boldsymbol{\beta}_i, \boldsymbol{\theta}_i] - [\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\theta}}_i]||_2^2.$$

HMR leverages the SMPL parameters.

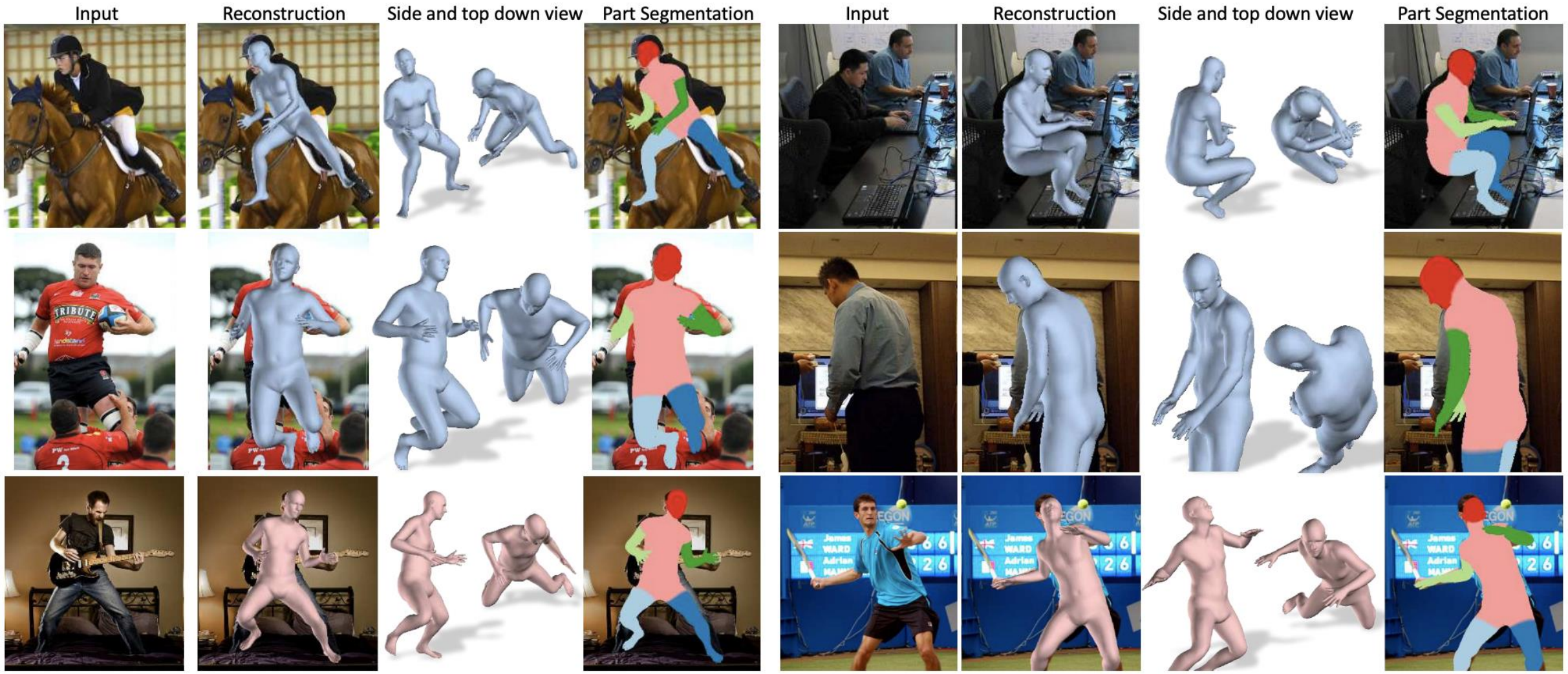
Per-frame Human Mesh Recovery: HMR



[CVPR 2018] HMR

HMR uses the prior of SMPL dataset.

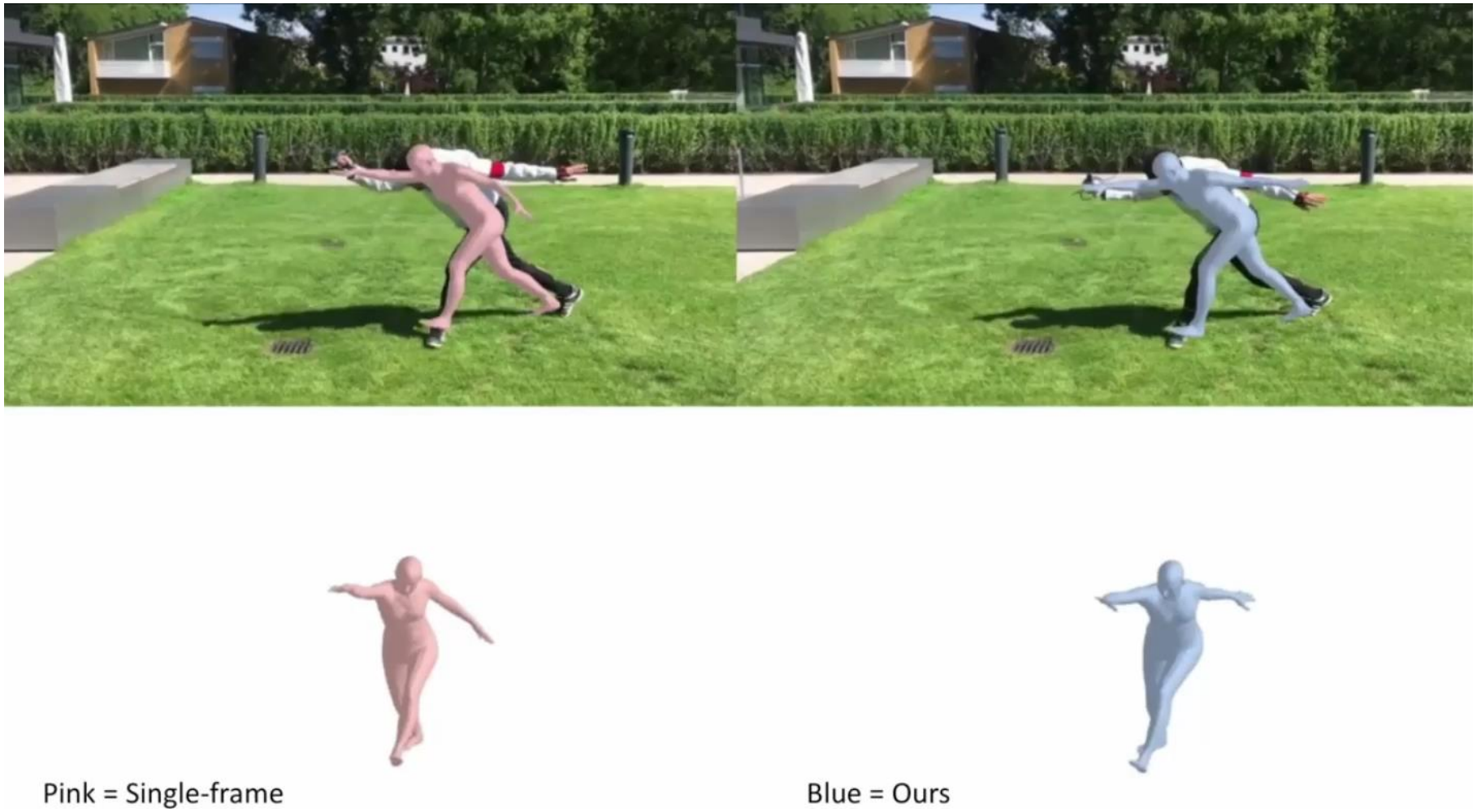
Per-frame Human Mesh Recovery: HMR



HMR enables the 3D reconstruction of human pose and shape.

Video Human Mesh Recovery: VIBE

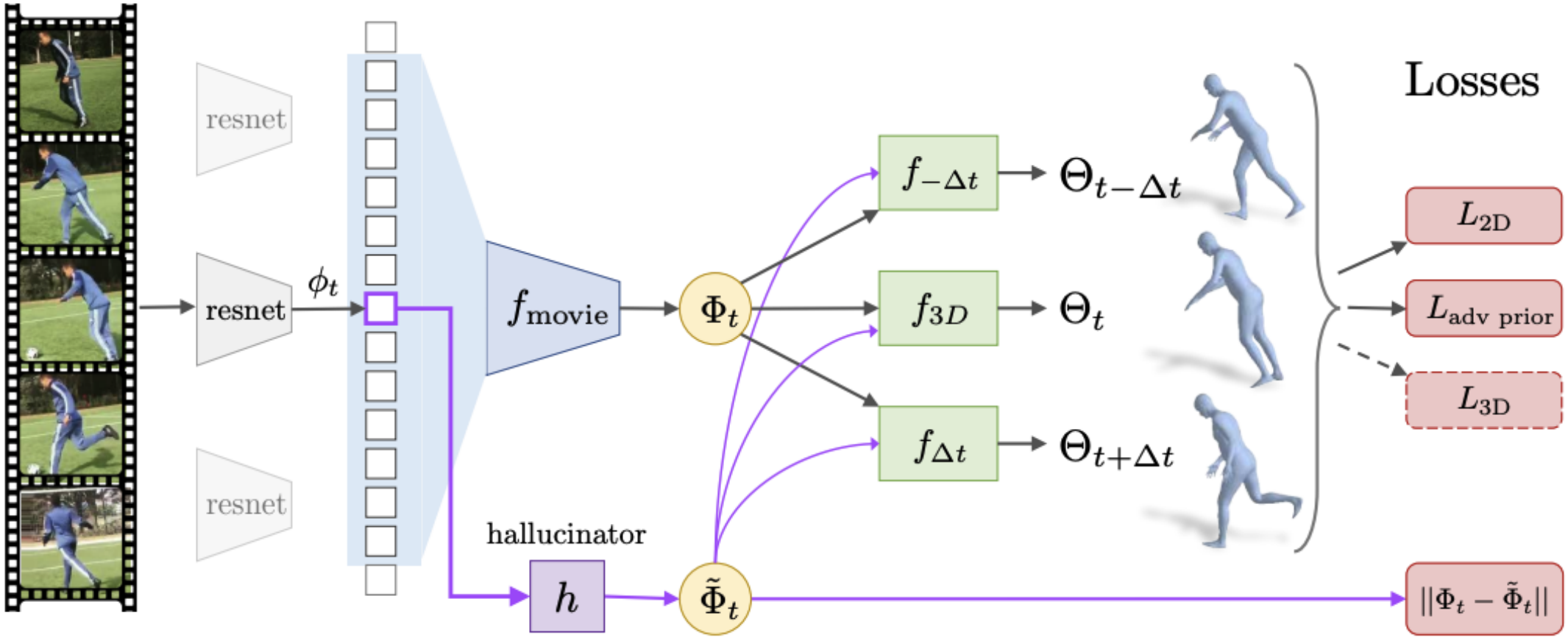
Motivation of Video HMR



[CVPR 2019] Human Dynamics

Human Dynamics leverages the temporal information.

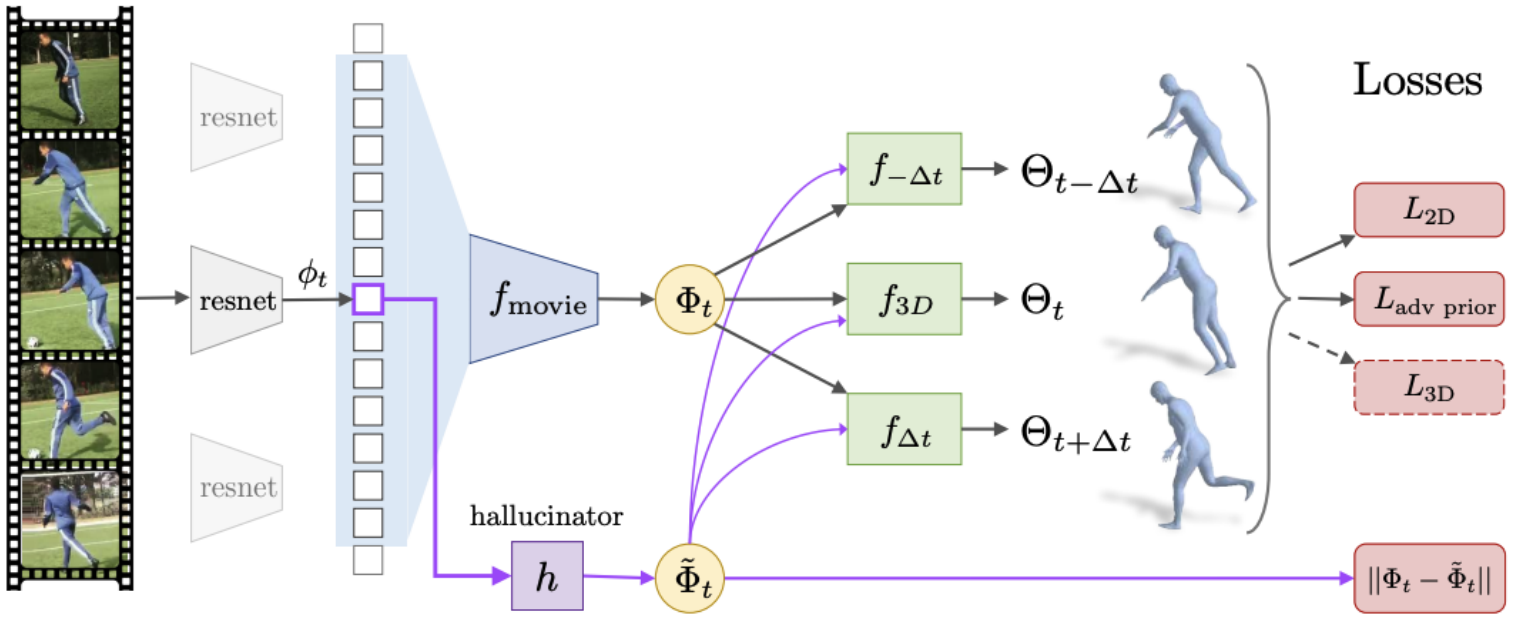
Motivation of Video HMR



[CVPR 2019] Human Dynamics

Human Dynamics leverages the temporal information.

Motivation of VIBE



[CVPR 2019] Human Dynamics

Problem of Human Dynamics:

- **Produce jittery and physically implausible motion**

→ **Use motion data prior!**

Human Dynamics produce physically implausible motion

Video Human Mesh Recovery: VIBE

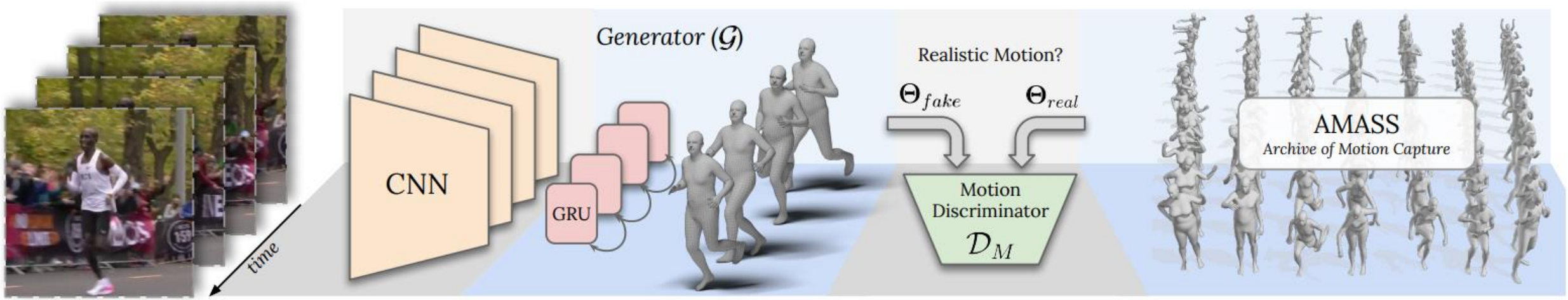
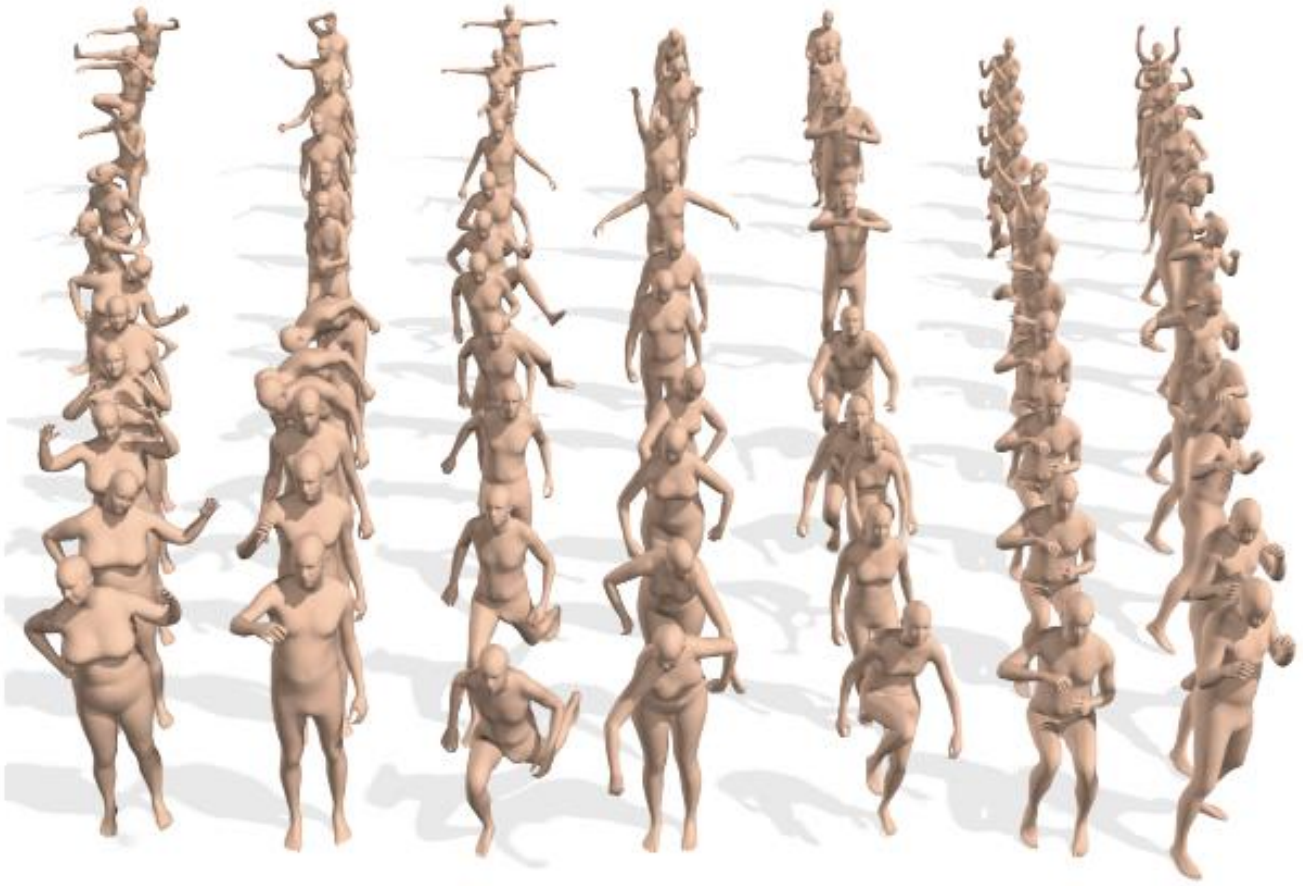


Figure 2: **VIBE architecture.** VIBE estimates SMPL body model parameters for each frame in a video sequence using a temporal generation network, which is trained together with a motion discriminator. The discriminator has access to a large corpus of human motions in SMPL format.

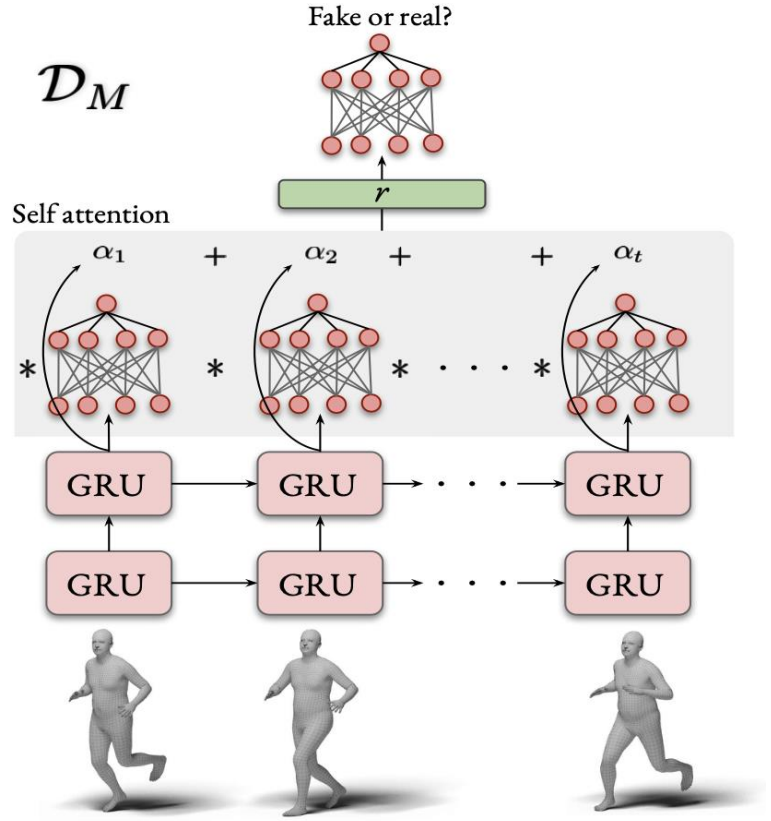
[CVPR 2020] VIBE

VIBE leverages the motion dataset prior (AMASS).

Video Human Mesh Recovery: VIBE



[ICCV 2019] AMASS



$$L_{adv} = \mathbb{E}_{\Theta \sim p_G}[(\mathcal{D}_M(\hat{\Theta}) - 1)^2]$$

$$L_{\mathcal{D}_M} = \mathbb{E}_{\Theta \sim p_R}[(\mathcal{D}_M(\Theta) - 1)^2] + \mathbb{E}_{\Theta \sim p_G}[\mathcal{D}_M(\hat{\Theta})^2]$$

AMASS is large motion capture dataset.

Video Human Mesh Recovery: VIBE

[CVPR 2020] VIBE

[CVPR 2019] Human Dynamics

World Grounded Video Human Mesh Recovery: SLAHMR

Motivation of World Grounded Video HMR

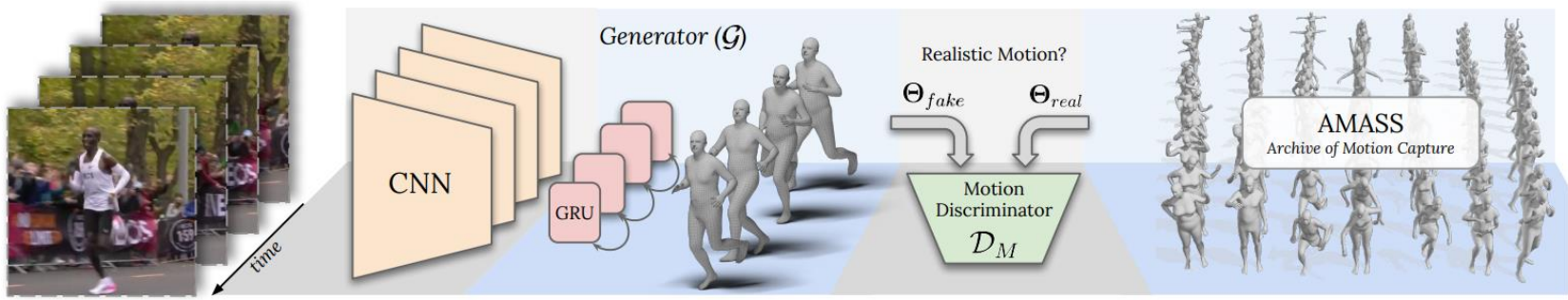


Figure 2: **VIBE architecture.** VIBE estimates SMPL body model parameters for each frame in a video sequence using a temporal generation network, which is trained together with a motion discriminator. The discriminator has access to a large corpus of human motions in SMPL format.

[CVPR 2020] VIBE

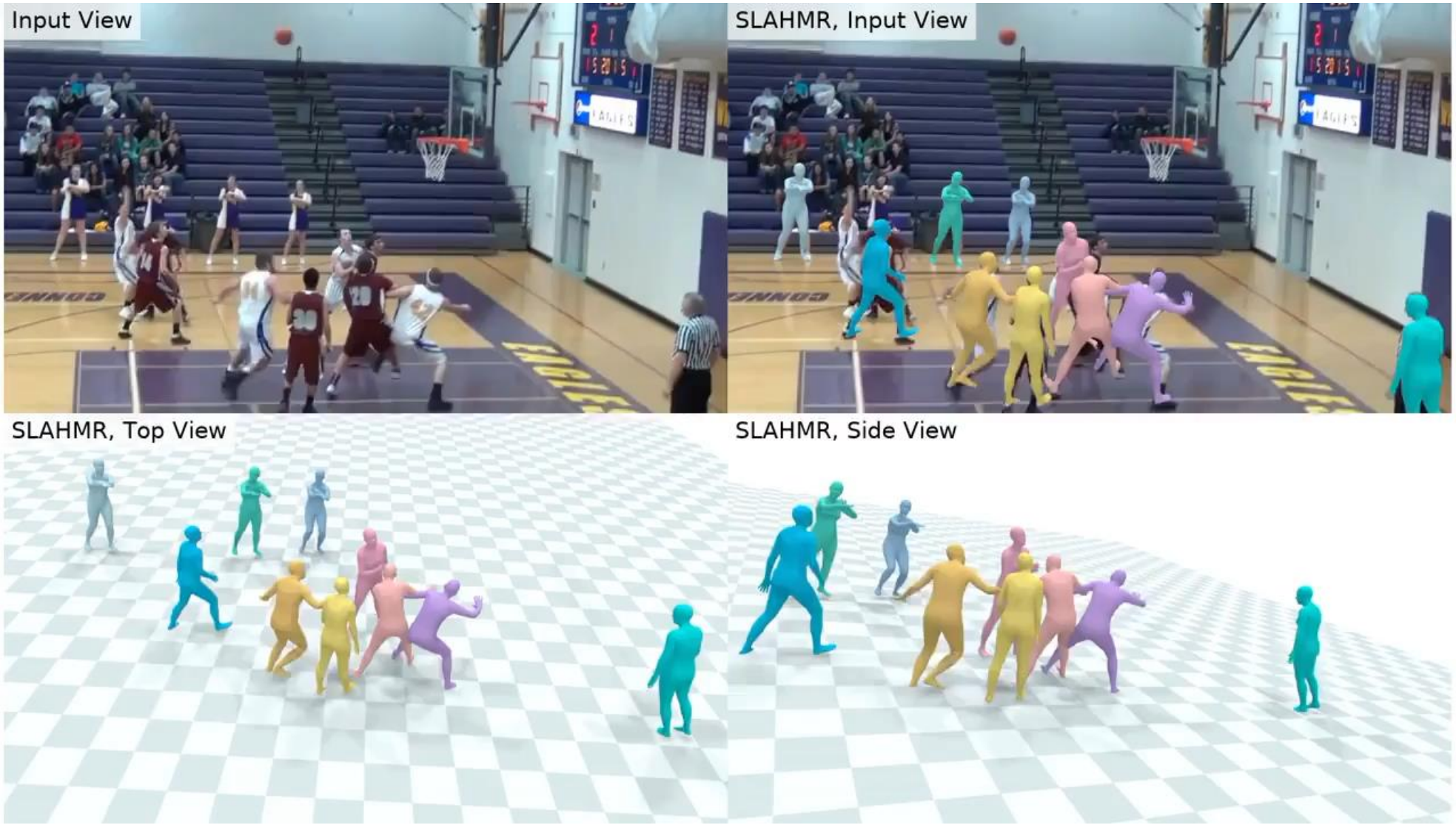
Problem of V-HMR:

- **Handle only camera space.**

→ **Leverage SLAM with HMR.**

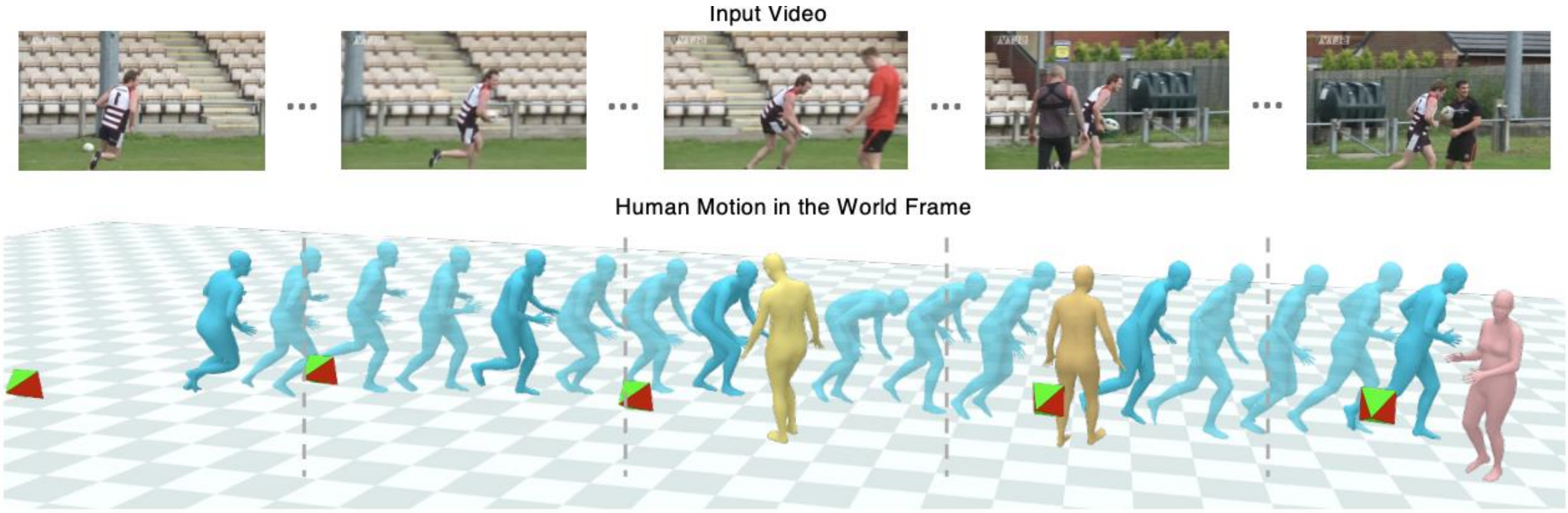
Traditional V-HMR methods predict motion in the camera coordinate frame.

World Grounded Video Human Mesh Recovery: SLAHMR



What is World-Grounded Video HMR?

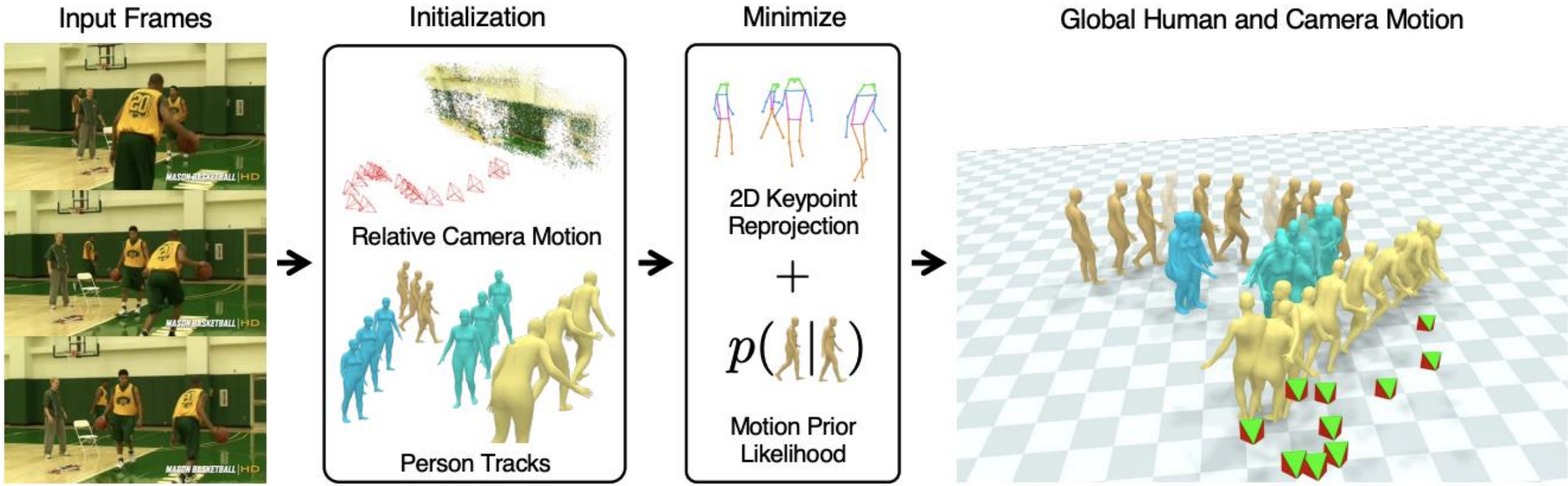
World Grounded Video Human Mesh Recovery: SLAHMR



Teaser of SLAHMR

SLAHMR predicts human mesh in world domain.

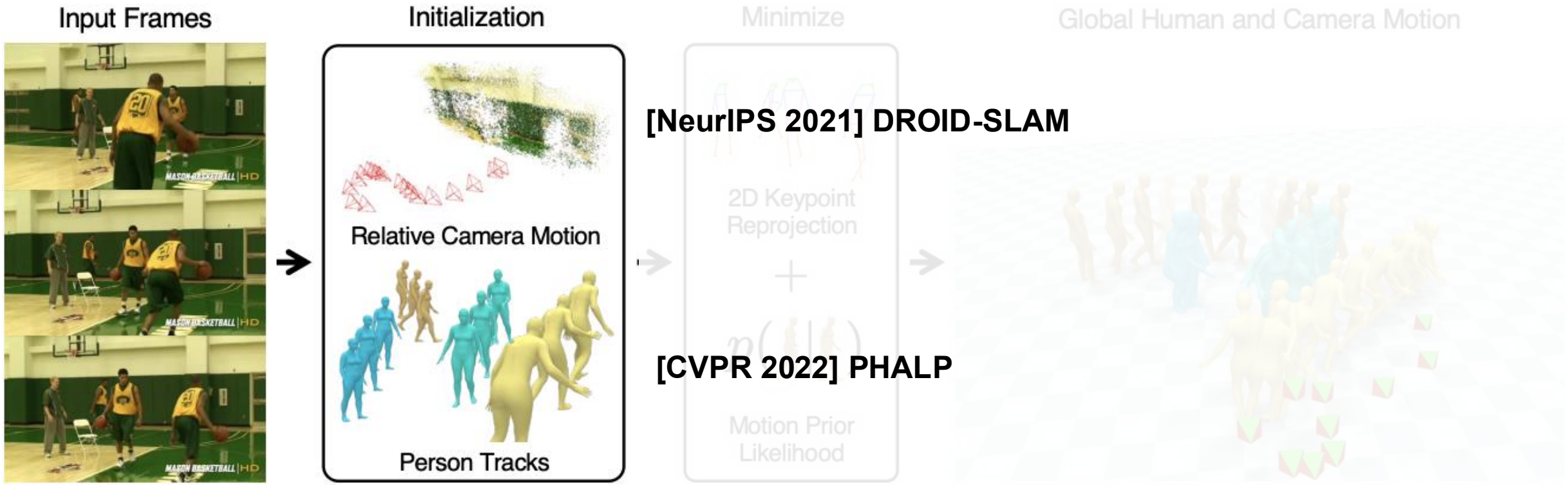
World Grounded Video Human Mesh Recovery: SLAHMR



Framework of SLAHMR

SLAHMR predicts human mesh in world domain.

World Grounded Video Human Mesh Recovery: SLAHMR



[NeurIPS 2021] DROID-SLAM

[CVPR 2022] PHALP

Framework of SLAHMR

SLAHMR leverages DROID-SLAM for camera pose estimation.

World Grounded Video Human Mesh Recovery: SLAHMR

We have: Input video, predicted 2D keypoints

Optimization object:

SMPL paramters

$$\mathbf{P}_t^i = \{\Phi_t^i, \Theta_t^i, \beta^i, \Gamma_t^i\}.$$

Global orientation(W)

Global translation(W)

SLAHMR is global optimization method.

World Grounded Video Human Mesh Recovery: SLAHMR

We have: Input video, predicted 2D keypoints

Input Video

Optimization object:

SMPL paramters

$$\mathbf{P}_t^i = \{\Phi_t^i, \Theta_t^i, \beta^i, \Gamma_t^i\}.$$

Global orientation(W)

Global translation(W)

projected 2D keypoints (pred)

predicted 2D keypoints (GT)

$$E_{\text{data}} = \sum_{i=1}^{N}\sum_{t=1}^{T} \psi_t^i \rho\big(\Pi_K(R_t \cdot {}^{\text{w}}\mathbf{J}_t^i + \alpha T_t) - \mathbf{x}_t^i\big),$$

DROID-SLAM (pred)

$${}^{\text{w}}\mathbf{J}_t^i = \mathcal{M}({}^{\text{w}}\Phi_t^i, \Theta_t^i, \beta^i) + {}^{\text{w}}\Gamma_t^i.$$

$${}^{\text{w}}\Phi_t^i = R_t^{-1\text{c}}\hat{\Phi}_t^i, \qquad {}^{\text{w}}\Gamma_t^i = R_t^{-1\text{c}}\hat{\Gamma}_t^i - \alpha R_t^{-1} T_t,$$

$$\beta_i = \hat{\beta}^i \qquad \Theta_t^i = \hat{\Theta}_t^i,$$

Camera frame SMPL (pred from PHARP)

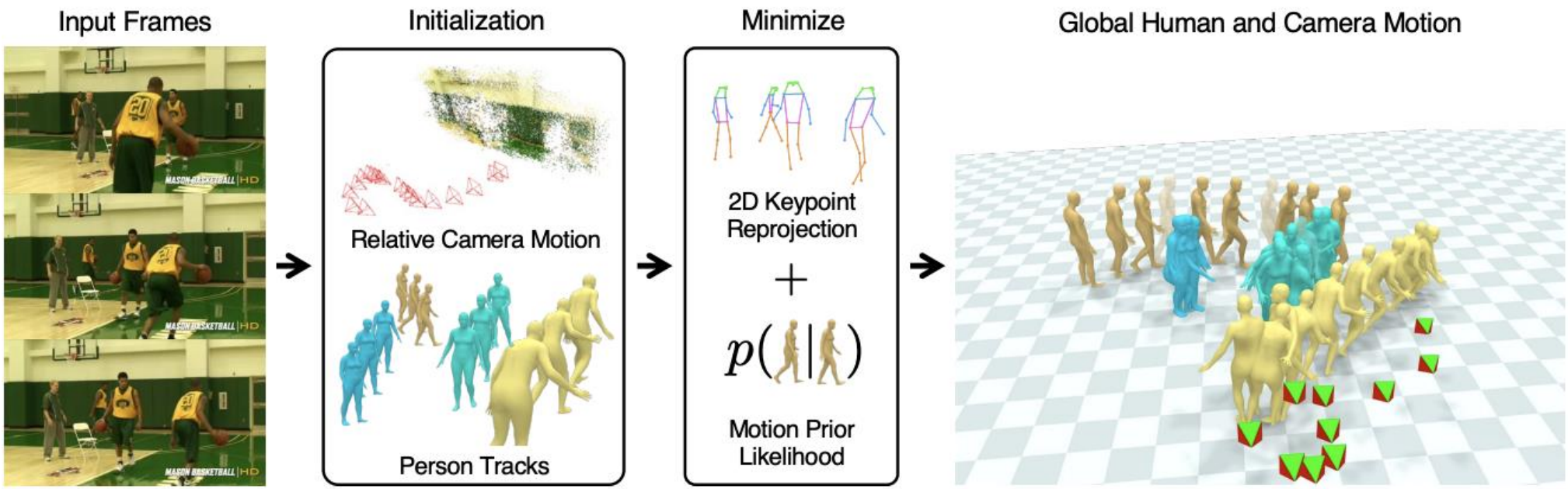
SLAHMR is global optimization method.

World Grounded Video Human Mesh Recovery: SLAHMR

Results of SLAHMR

Recent Works

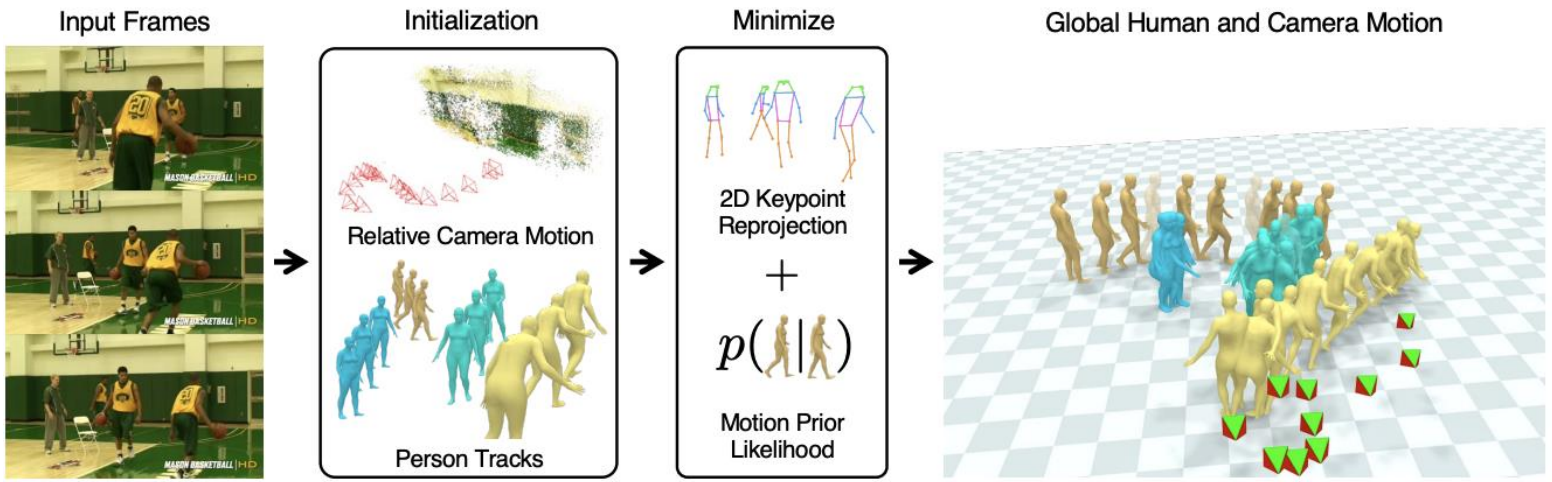
Motivation of WHAM



[CVPR 2023] SLAHMR

Global optimization method (SLAHMR) is so slow.

Motivation of WHAM



[CVPR 2023] SLAHMR

Problem of SLAHMR:

1. **Slow**

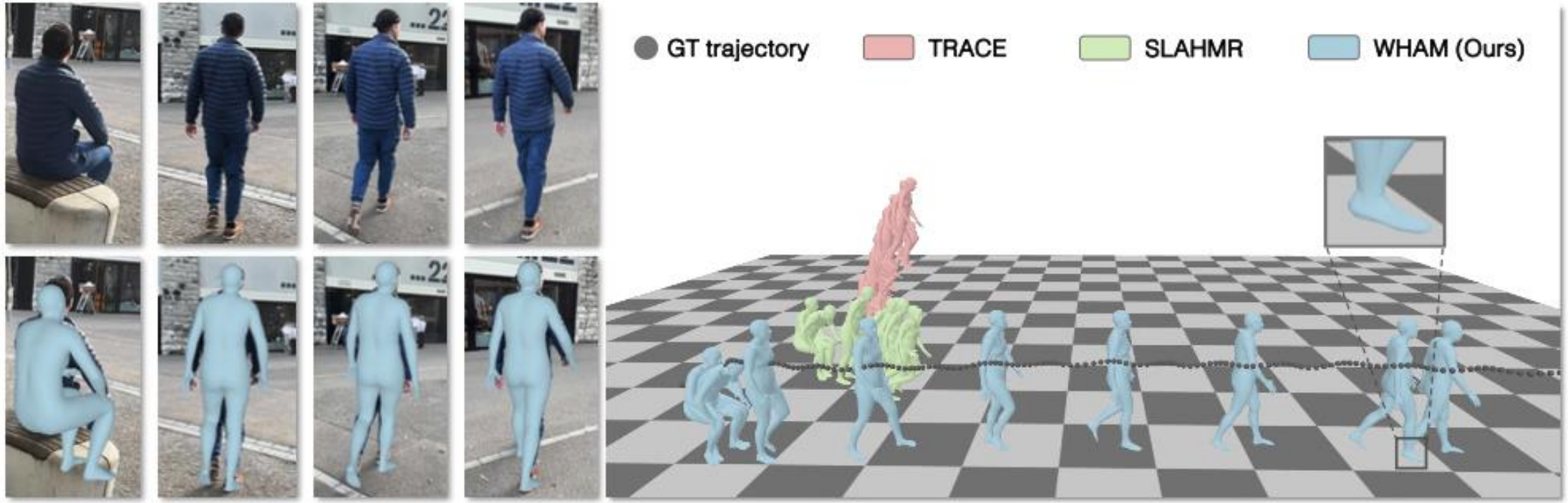
→ **Use feed forward model!**

2. **Foot sliding**

→ **Use contact module!**

Global optimization method (SLAHMR) is so slow (~ 260 minutes per 1000 frames).

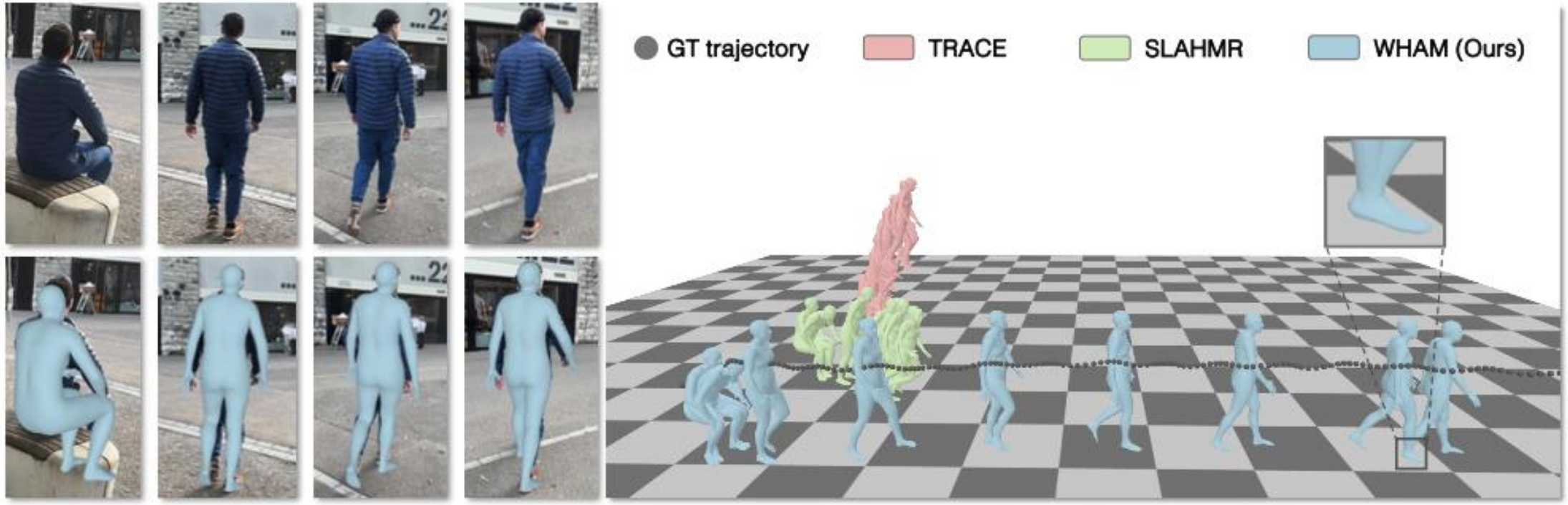
WHAM



Teaser of WHAM: SLAHMR fails to capture global 3D human traj when given in the wild videos.

WHAM is regression-based model and faster than SLAHMR .

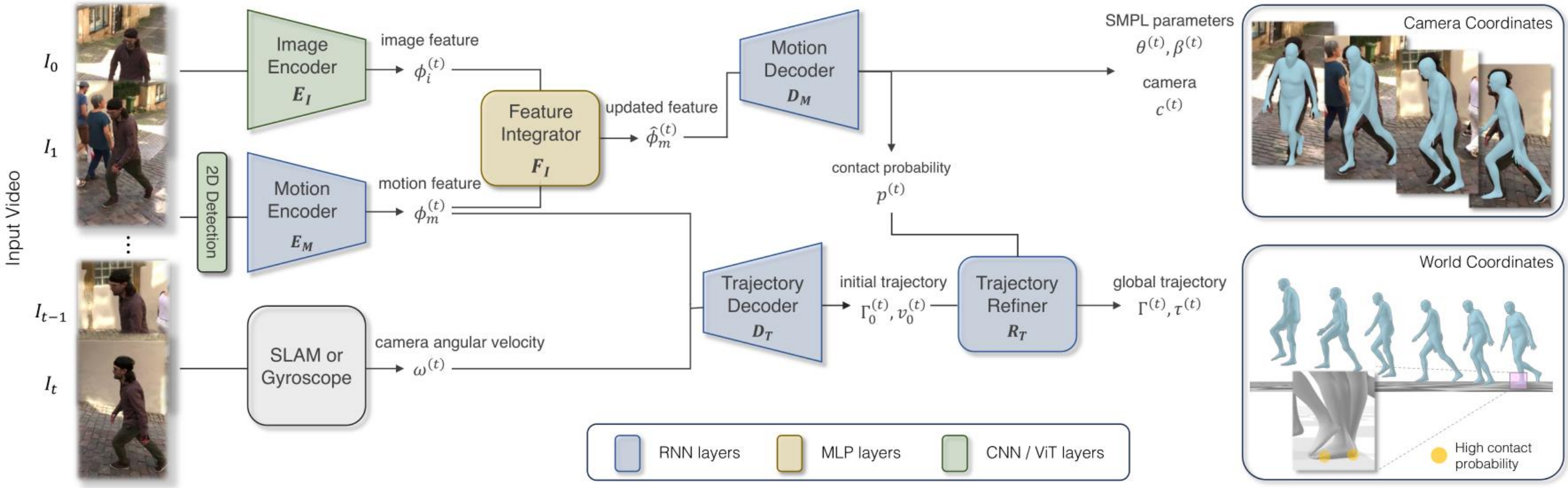
WHAM



Teaser of WHAM: SLAHMR fails to capture global 3D human traj when given in the wild videos.

WHAM is regression-based model and faster than SLAHMR

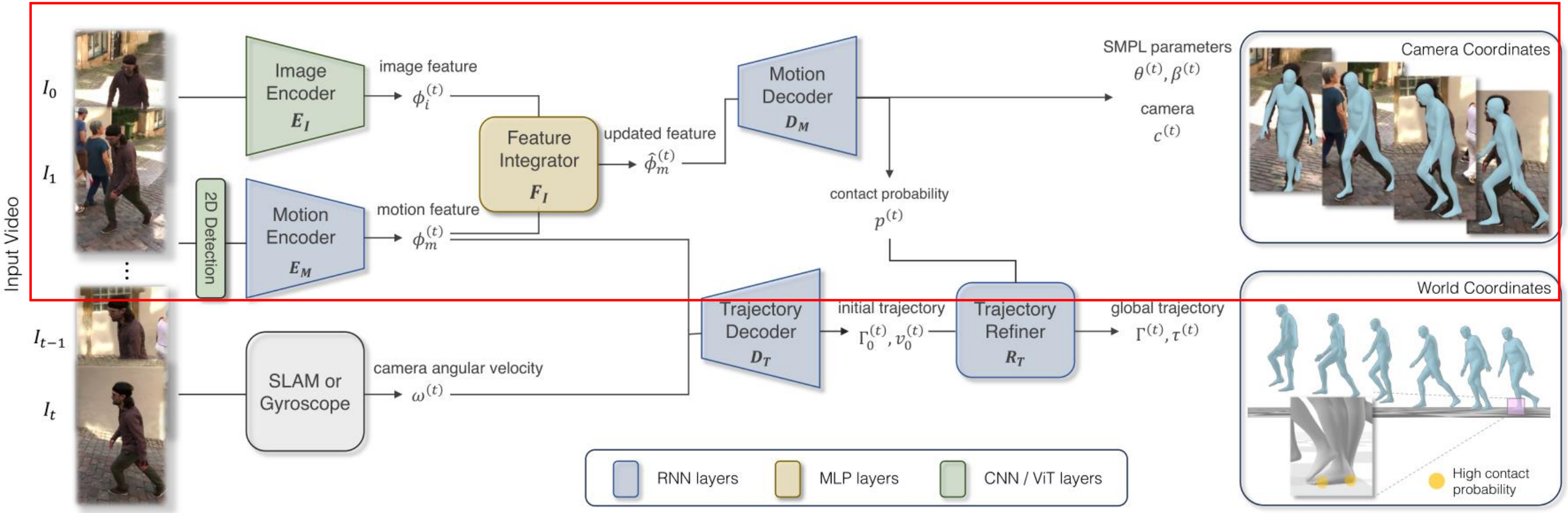
WHAM



WHAM architecture

WHAM is regression-based model and faster than SLAHMR

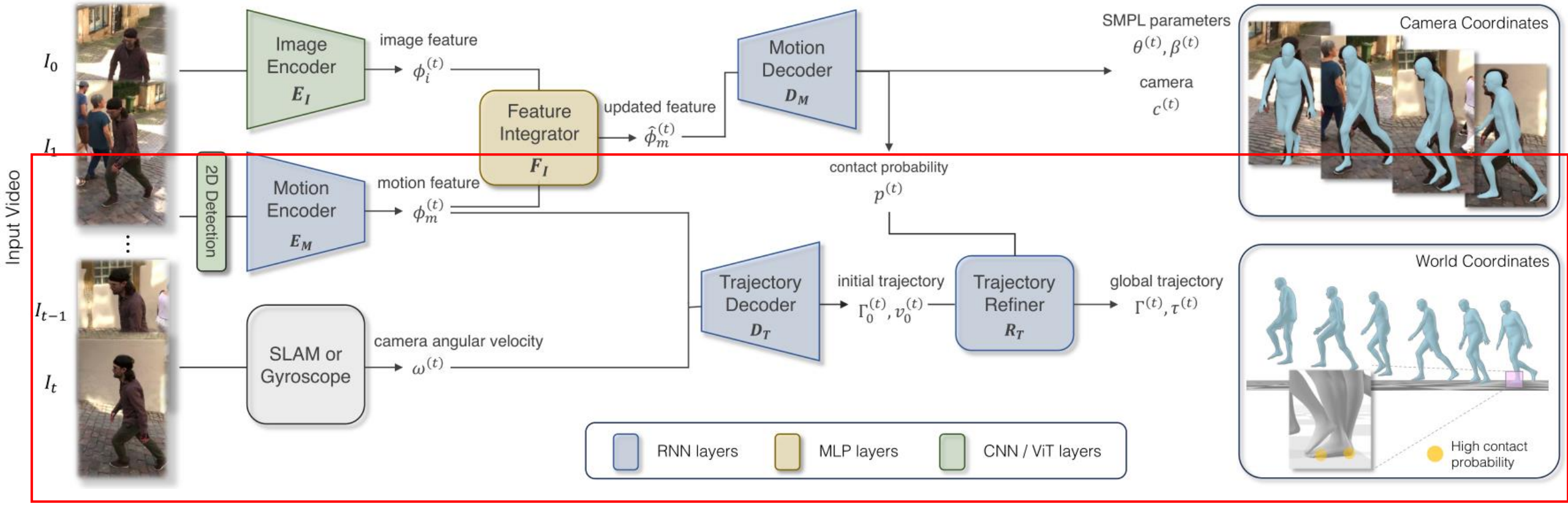
WHAM



WHAM architecture

WHAM is regression-based model and faster than SLAHMR

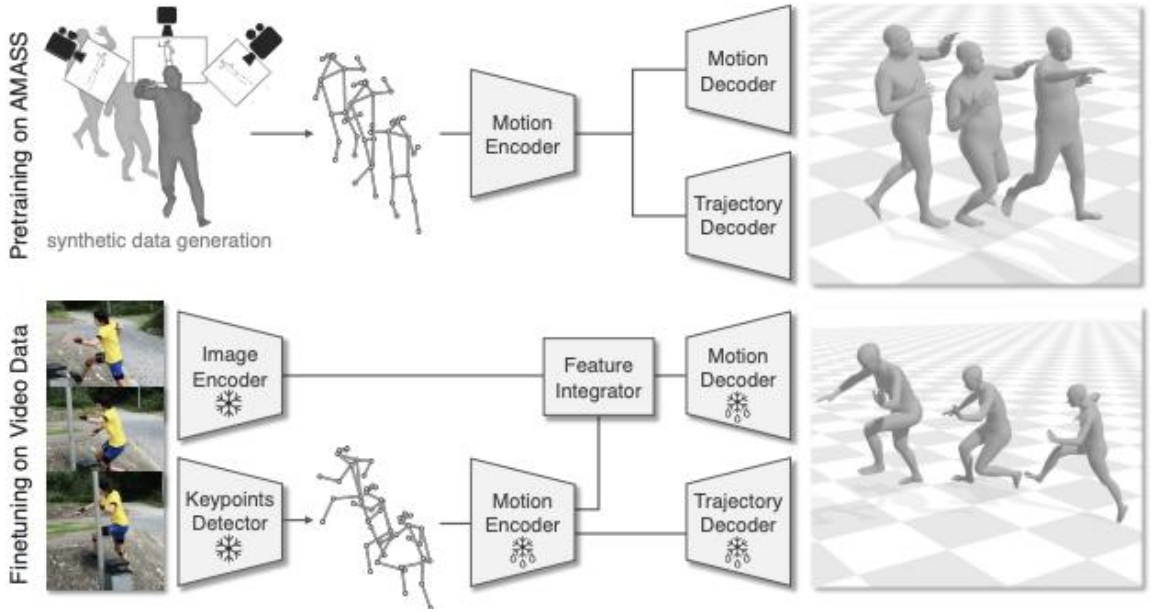
WHAM



WHAM architecture

WHAM is regression-based model and faster than SLAHMR

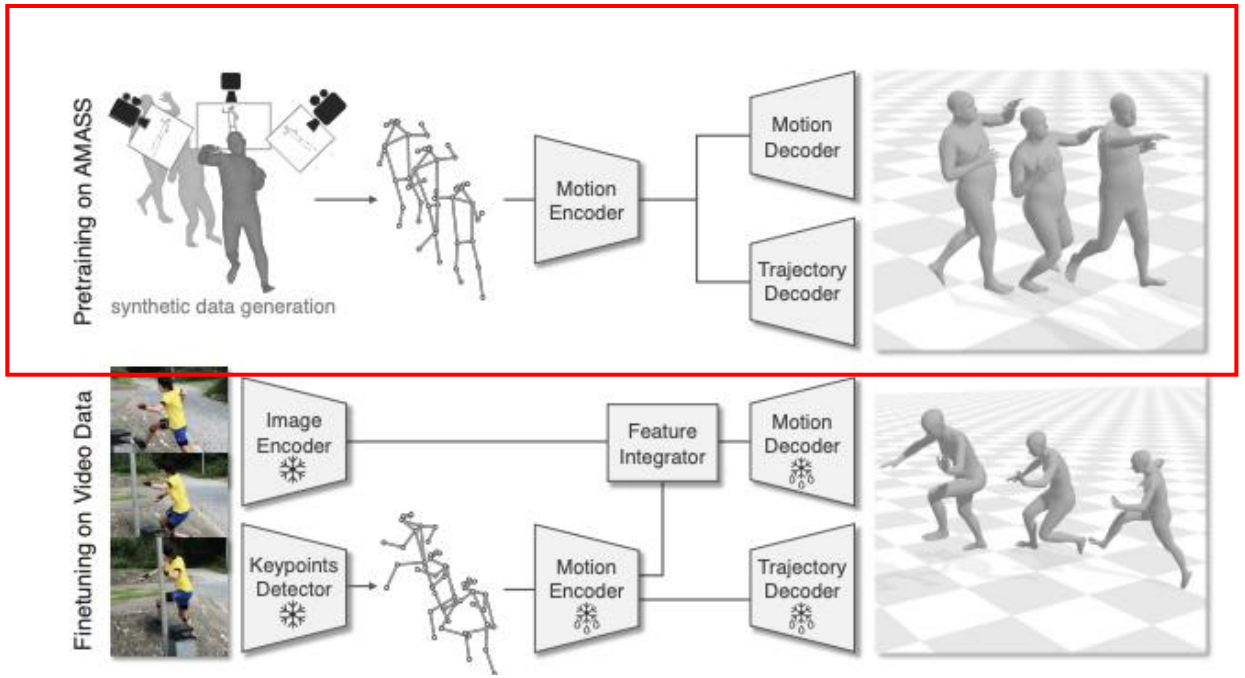
WHAM



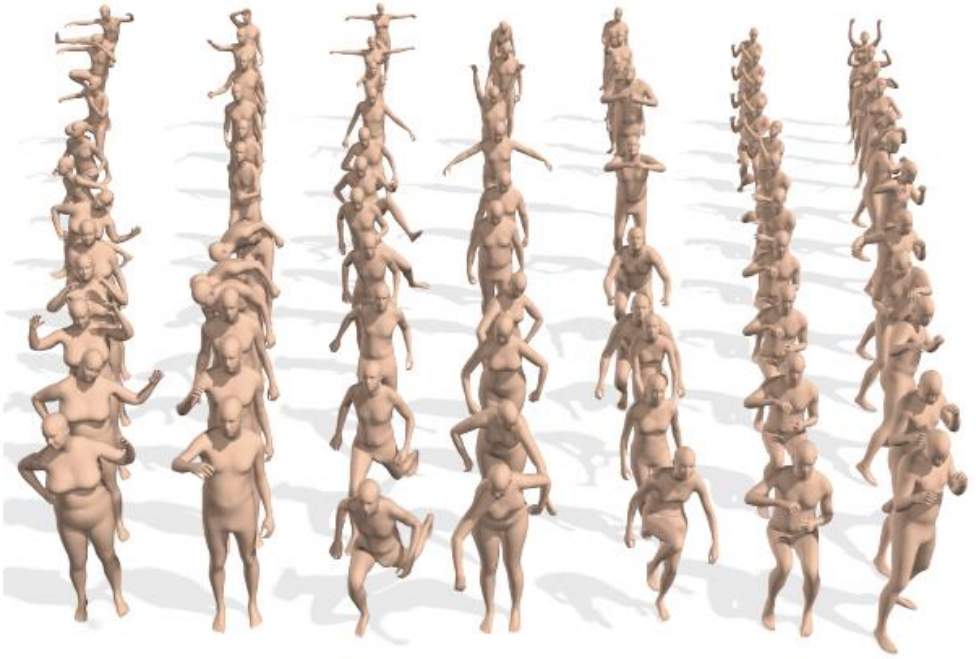
Training of WHAM

WHAM is pretrained on AMASS dataset.

WHAM



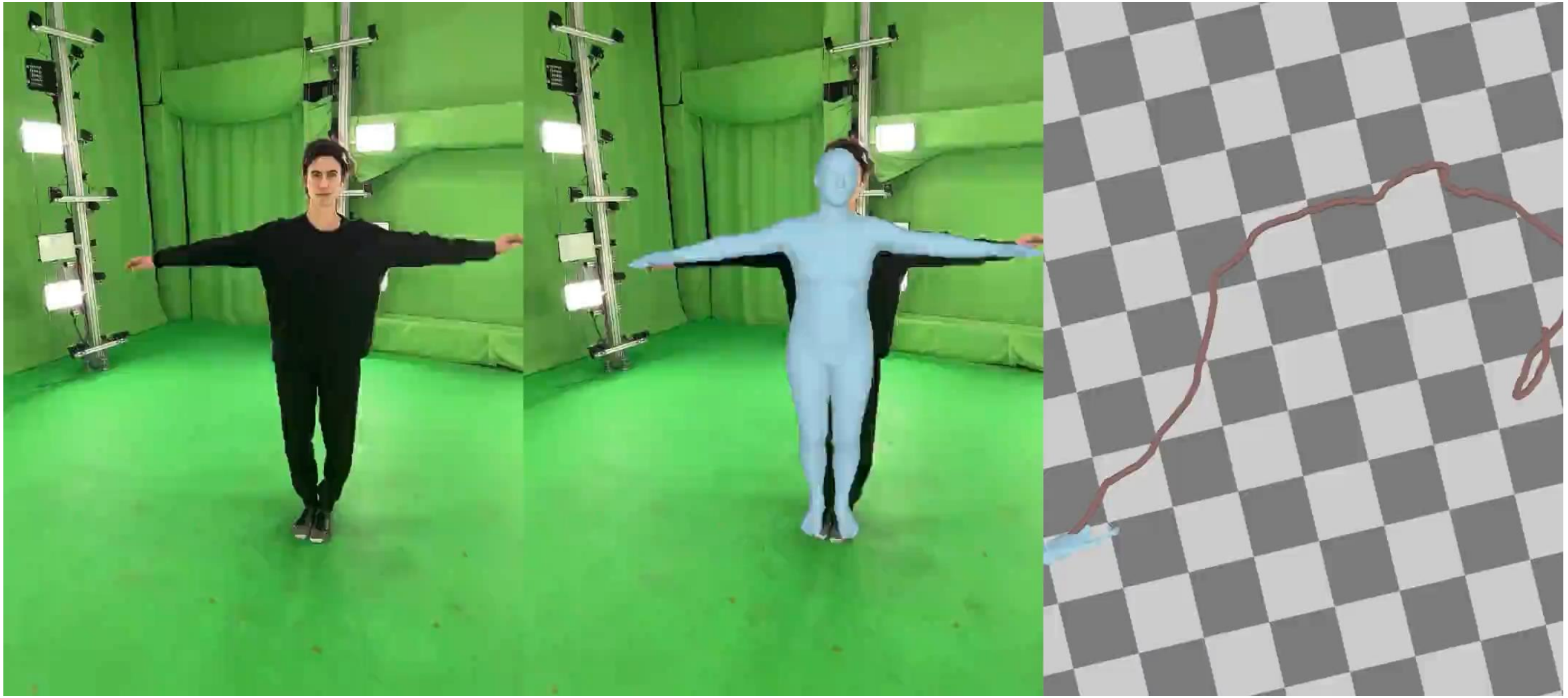
Training of WHAM



[ICCV 2019] AMASS

WHAM is pretrained on AMASS dataset instead of using prior loss.

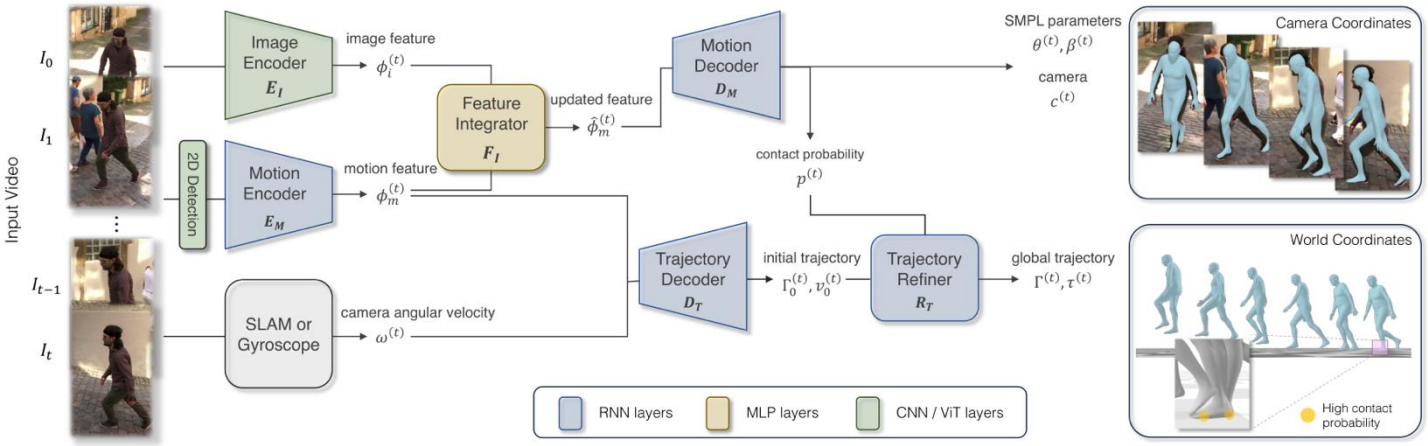
WHAM



WHAM enable feed-forward world-grounded HMR.

GVHMR

Motivation of GVHMR



[CVPR 2024] WHAM

Problem of WHAM:

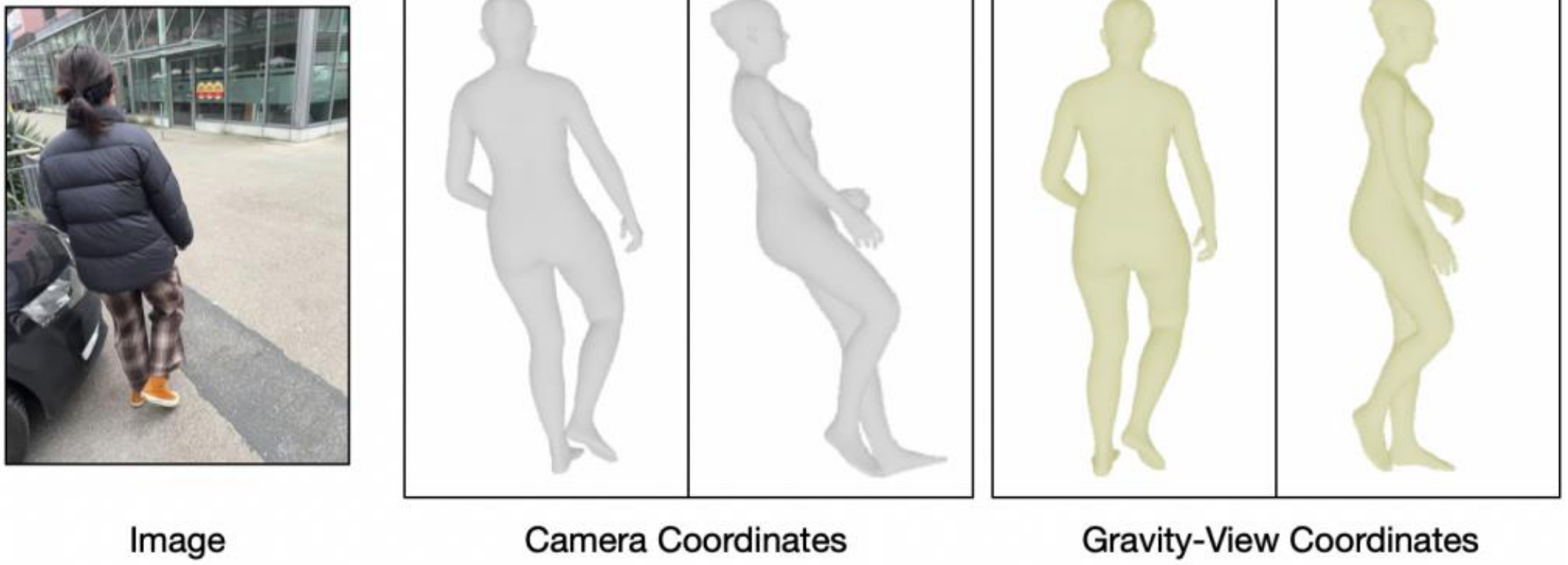
- **Large error in long-term motion**

⟶ **Use direct transformer!**

⟶ **Use Gravity-View coordinate!**

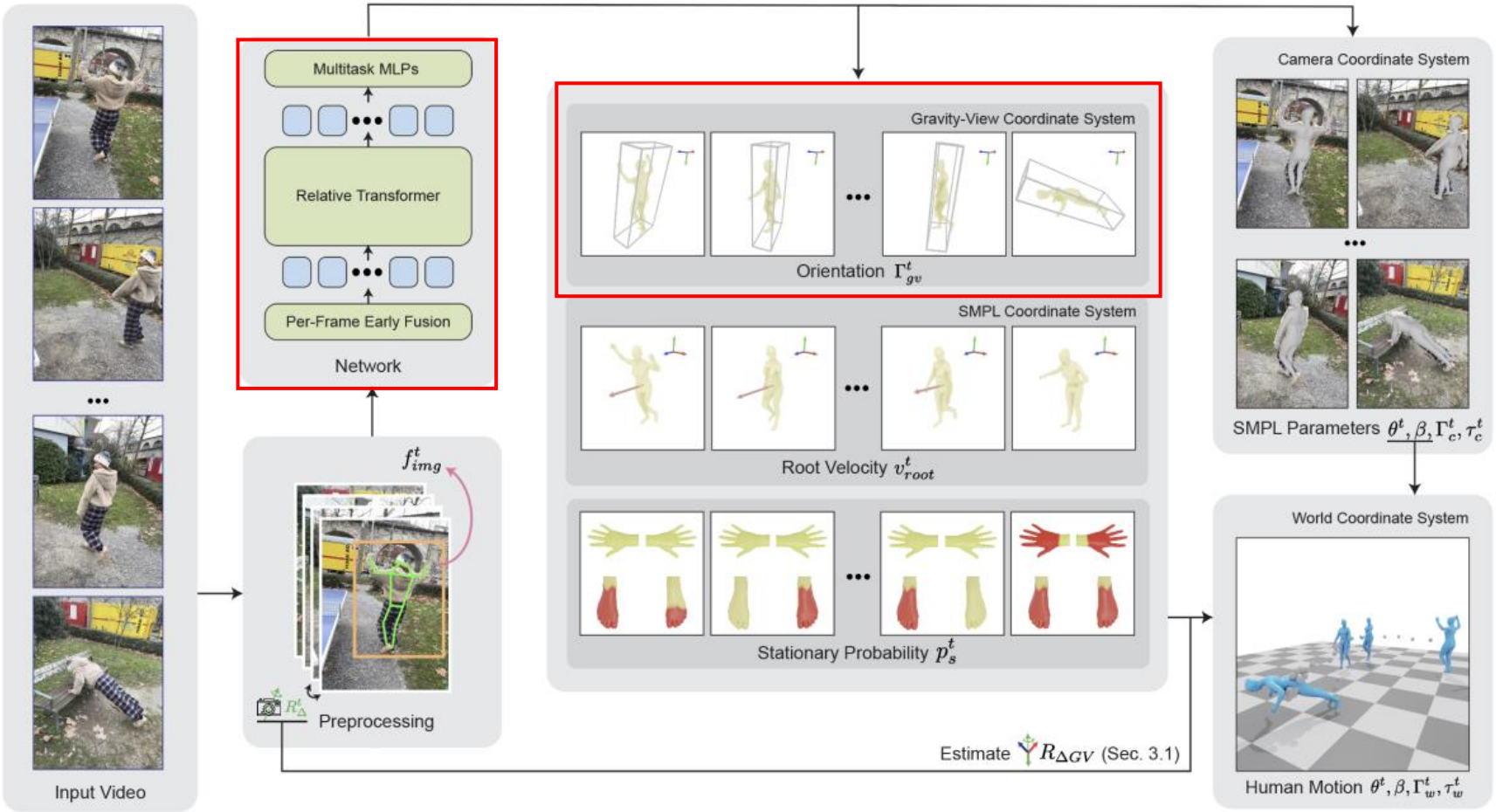
WHAM is autoregressive model, so there is error accumulation in long-term motion.

Motivation of GVHMR



In camera coordinates, a person may appear inclined due to the camera's roll and pitch movement.

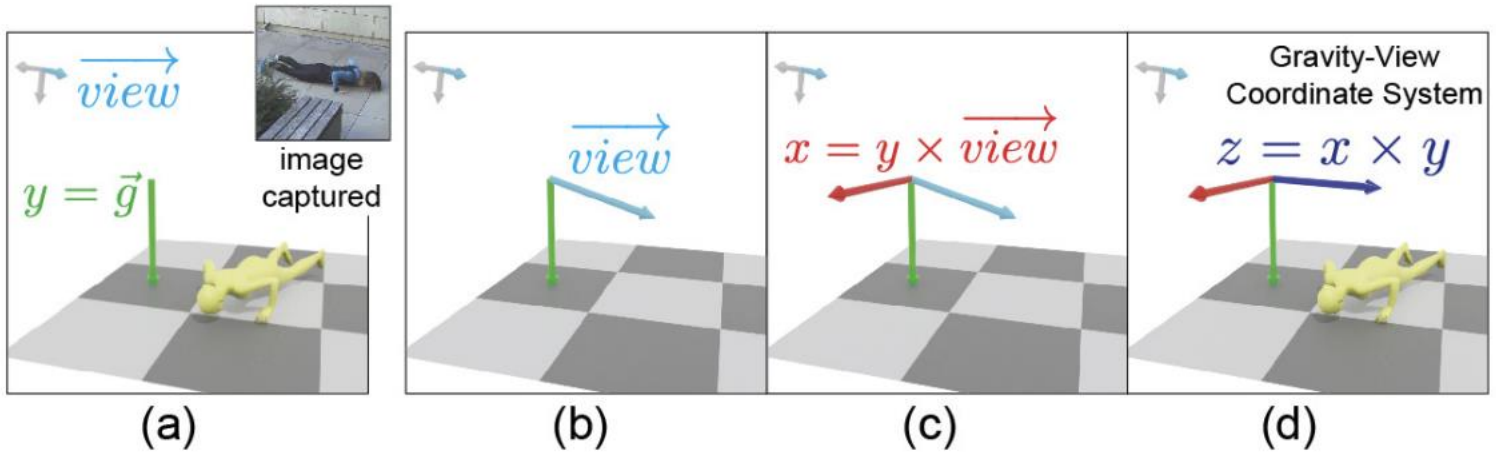
GVHMR



GVHMR pipeline

GVHMR predicts world-ground human mesh

GVHMR



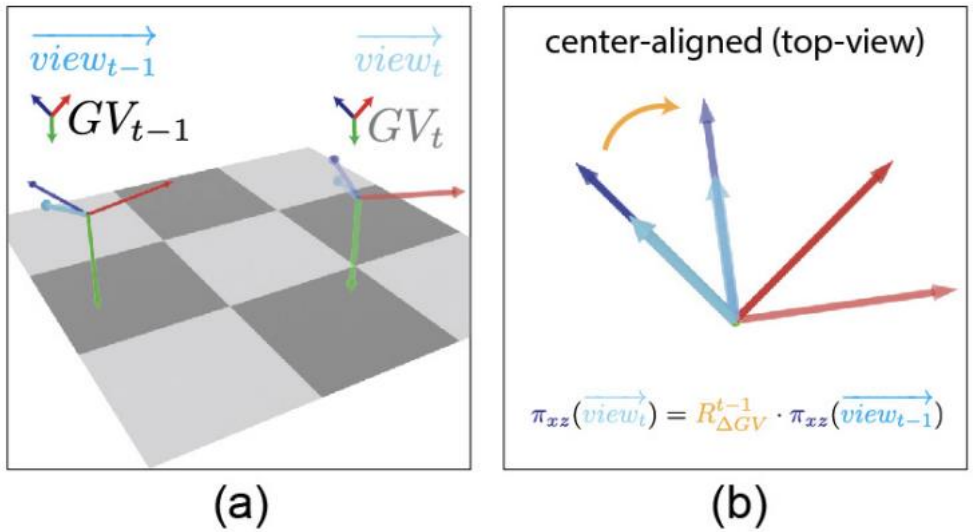
(a) y = gravity (up)

(b) view = camera forward

(c) x = y × view

(d) z = x × y

Gravity coordinates system

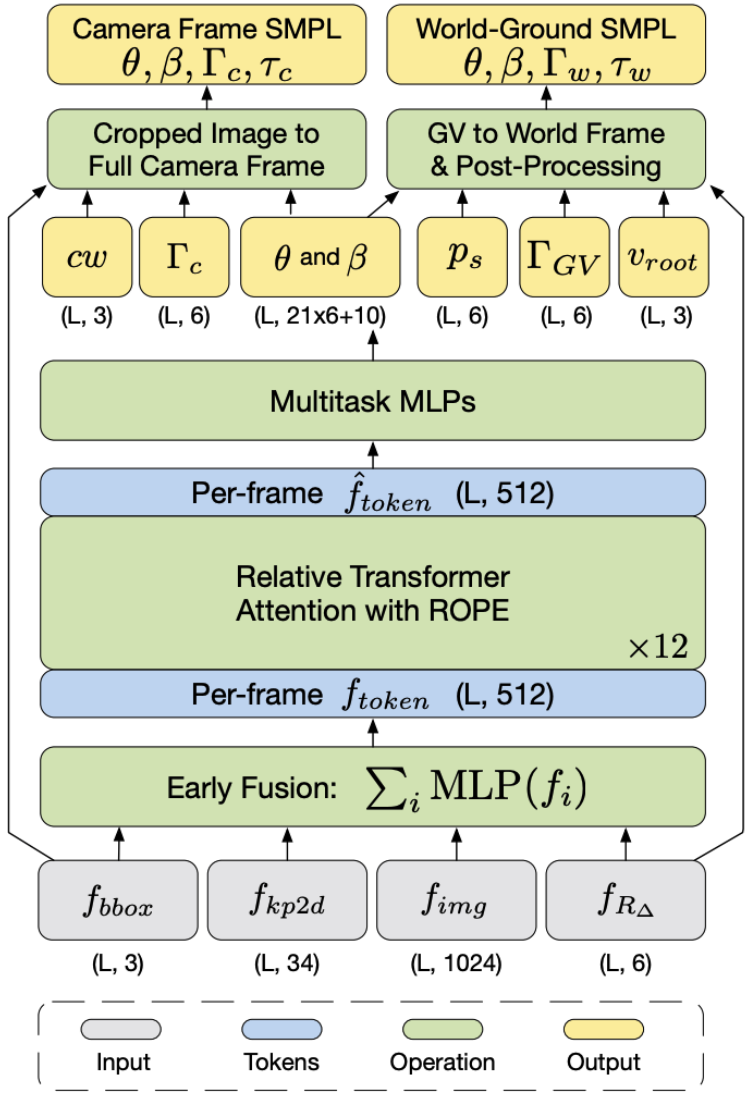


The only remaining degree of freedom is rotation about the gravity axis (yaw).

GVHMR leverages Gravity-View coordinates system (3 dof -> 1 dof) .

GVHMR

GT : camara, world SMPL paramters.



GVHMR architecture

cw: camera param Γ_c: Cam to SMPL θ, β :SMPL param Γ_GV :GV to SMPL

Bbox (L,3) from YOLO or GT

Kp2d (L,34) from ViTPose or GT

img (L,1024) from HMR2.0

Rot (L,6) from DPVO

Input: LxHxWx3 (video)

GVHMR is a transformer that maps per-frame features to SMPL parameters.

GVHMR

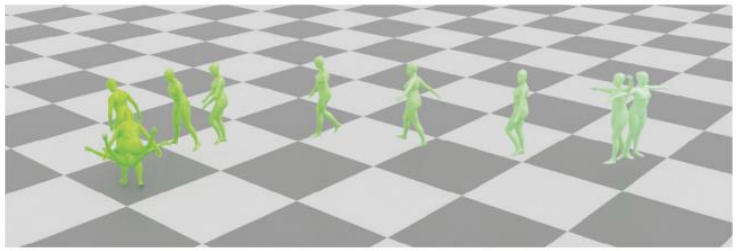
Models	RICH (24)					EMDB (24)				
	WA-MPJPE$_{100}$	W-MPJPE$_{100}$	RTE	Jitter	Foot-Sliding	WA-MPJPE$_{100}$	W-MPJPE$_{100}$	RTE	Jitter	Foot-Sliding
DPVO[Teed et al. 2024] +HMR2.0[Goel et al. 2023]	184.3	338.3	7.7	255.0	38.7	647.8	2231.4	15.8	537.3	107.6
GLAMR [Yuan et al. 2022]	129.4	236.2	3.8	49.7	18.1	280.8	726.6	11.4	46.3	20.7
TRACE [Sun et al. 2023]	238.1	925.4	610.4	1578.6	230.7	529.0	1702.3	17.7	2987.6	370.7
SLAHMR [Ye et al. 2023]	98.1	186.4	28.9	34.3	5.1	326.9	776.1	10.2	31.3	14.5
WHAM (w/ DPVO) [Shin et al. 2024]	109.9	184.6	4.1	19.7	3.3	135.6	354.8	6.0	22.5	4.4
WHAM (w/ GT gyro) [Shin et al. 2024]	109.9	184.6	4.1	19.7	3.3	131.1	335.3	4.1	21.0	4.4
Ours (w/ DPVO)	**78.8**	**126.3**	**2.4**	**12.8**	**3.0**	111.0	276.5	2.0	16.7	**3.5**
Ours (w/ GT gyro)	**78.8**	**126.3**	**2.4**	**12.8**	**3.0**	**109.1**	**274.9**	**1.9**	**16.5**	**3.5**

Qualitative results (World)

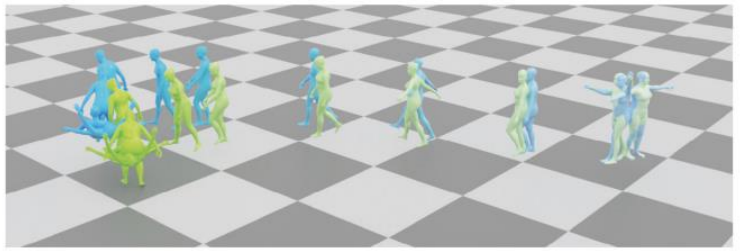
	Models	3DPW (14)				RICH (24)				EMDB (24)			
		PA-MPJPE	MPJPE	PVE	Accel	PA-MPJPE	MPJPE	PVE	Accel	PA-MPJPE	MPJPE	PVE	Accel
per-frame	SPIN [Kolotouros et al. 2019]	59.2	96.9	112.8	31.4	69.7	122.9	144.2	35.2	87.1	140.3	174.9	41.3
	PARE* [Kocabas et al. 2021a]	46.5	74.5	88.6	–	60.7	109.2	123.5	–	72.2	113.9	133.2	–
	CLIFF* [Li et al. 2022b]	43.0	69.0	81.2	22.5	56.6	102.6	115.0	22.4	68.1	103.3	128.0	24.5
	HybrIK* [Li et al. 2021]	41.8	71.6	82.3	–	56.4	96.8	110.4	–	65.6	103.0	122.2	–
	HMR2.0 [Goel et al. 2023]	44.4	69.8	82.2	18.1	48.1	96.0	110.9	18.8	60.6	98.0	120.3	19.8
	ReFit* [Wang and Daniilidis 2023]	40.5	65.3	75.1	18.5	47.9	80.7	92.9	17.1	58.6	88.0	104.5	20.7
temporal	TCMR* [Choi et al. 2021]	52.7	86.5	101.4	6.0	65.6	119.1	137.7	5.0	79.6	127.6	147.9	5.3
	VIBE* [Kocabas et al. 2020]	51.9	82.9	98.4	18.5	68.4	120.5	140.2	21.8	81.4	125.9	146.8	26.6
	MPS-Net* [Wei et al. 2022]	52.1	84.3	99.0	6.5	67.1	118.2	136.7	5.8	81.3	123.1	138.4	6.2
	GLoT* [Shen et al. 2023]	50.6	80.7	96.4	6.0	65.6	114.3	132.7	5.2	78.8	119.7	138.4	5.4
	GLAMR [Yuan et al. 2022]	51.1	–	–	8.0	79.9	–	–	107.7	73.5	113.6	133.4	32.9
	TRACE* [Sun et al. 2023]	50.9	79.1	95.4	28.6	–	–	–	–	70.9	109.9	127.4	25.5
	SLAHMR [Ye et al. 2023]	55.9	–	–	–	52.5	–	–	9.4	69.5	93.5	110.7	7.1
	PACE [Kocabas et al. 2024]	–	–	–	–	49.3	–	–	8.8	–	–	–	–
	WHAM* [Shin et al. 2024]	**35.9**	57.8	68.7	6.6	44.3	80.0	91.2	5.3	50.4	79.7	94.4	5.3
	Ours*	36.2	**55.6**	**67.2**	**5.0**	**39.5**	**66.0**	**74.4**	**4.1**	**42.7**	**72.6**	**84.2**	**3.6**

Qualitative results (Camera)

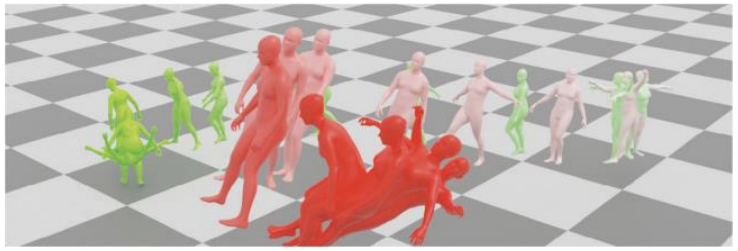
GVHMR



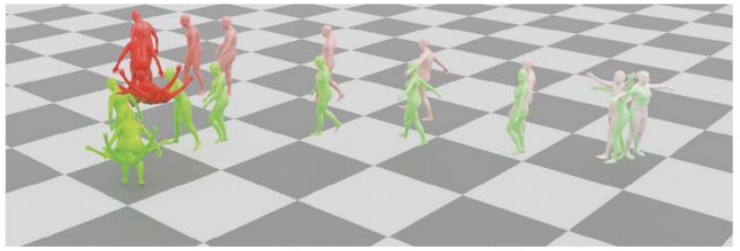
Ground Truth



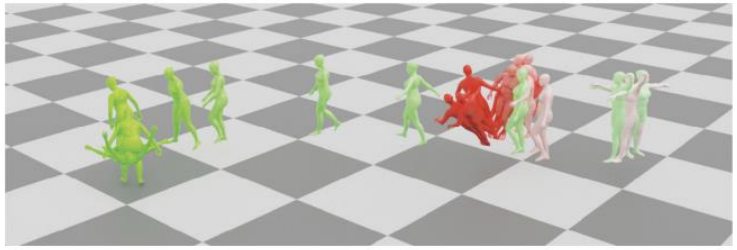
Full Model



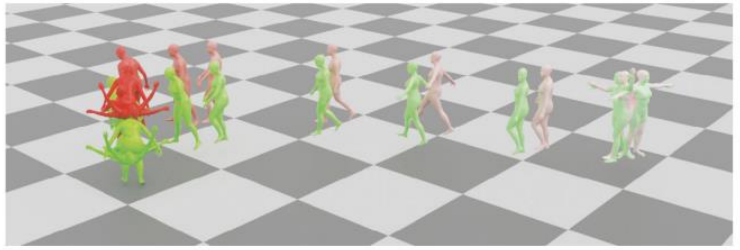
(2) w/o Γ_{GV}



(3) w/o Transformer



(5) w/o RoPE



(7) w/o Post-Processing

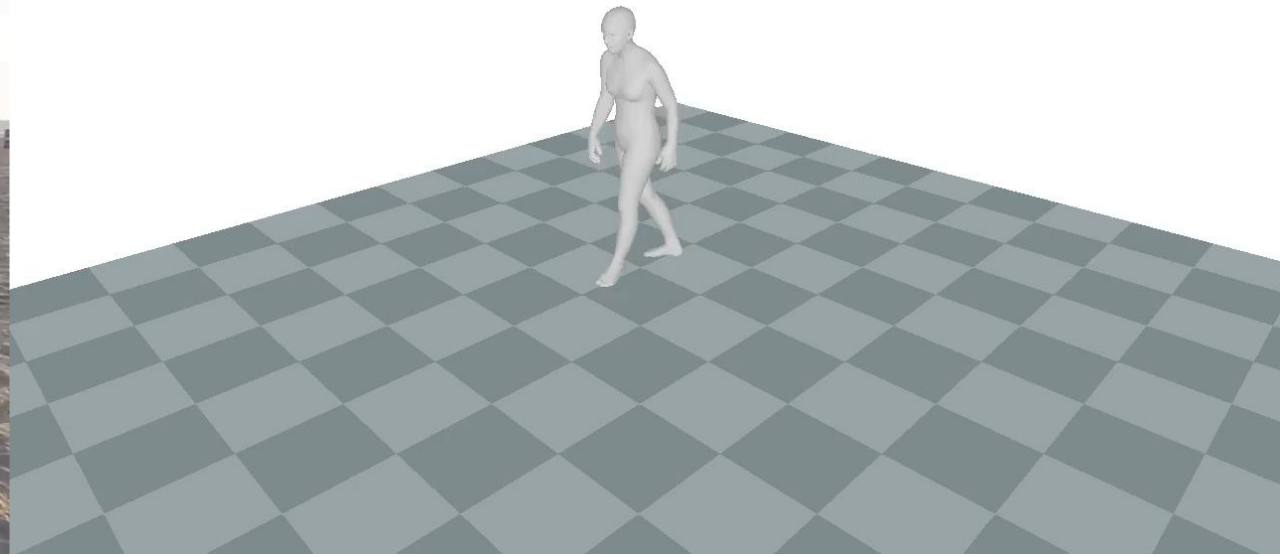
Ablation study

WHAM vs GVHMR

	[CVPR 2024] WHAM	[SIGGRAPH Asia 2024] GVHMR
Temporal modeling	• Uni-directional RNN	• Transformer (RoPE)
Training signal	• Pretrain on AMASS, finetune on video features.	• End-to-end training on multiple dataset.
Coordinate handling	• Camera coordinate -> world	• Camera +GV coordinate -> world

Conclusion

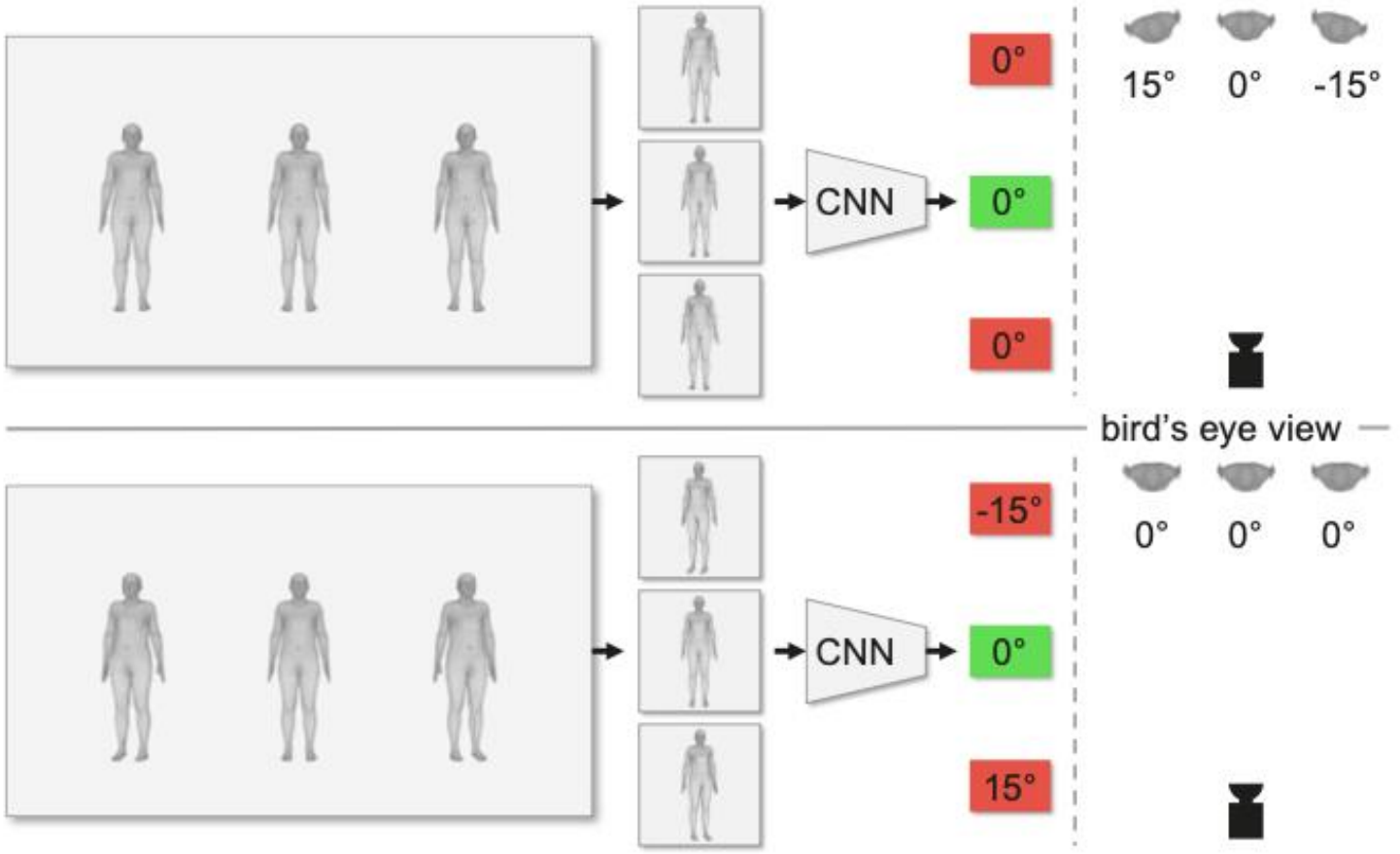
Conclusion



“Intelligence emerges from observing, interacting, and continually learning from other intelligent systems.”

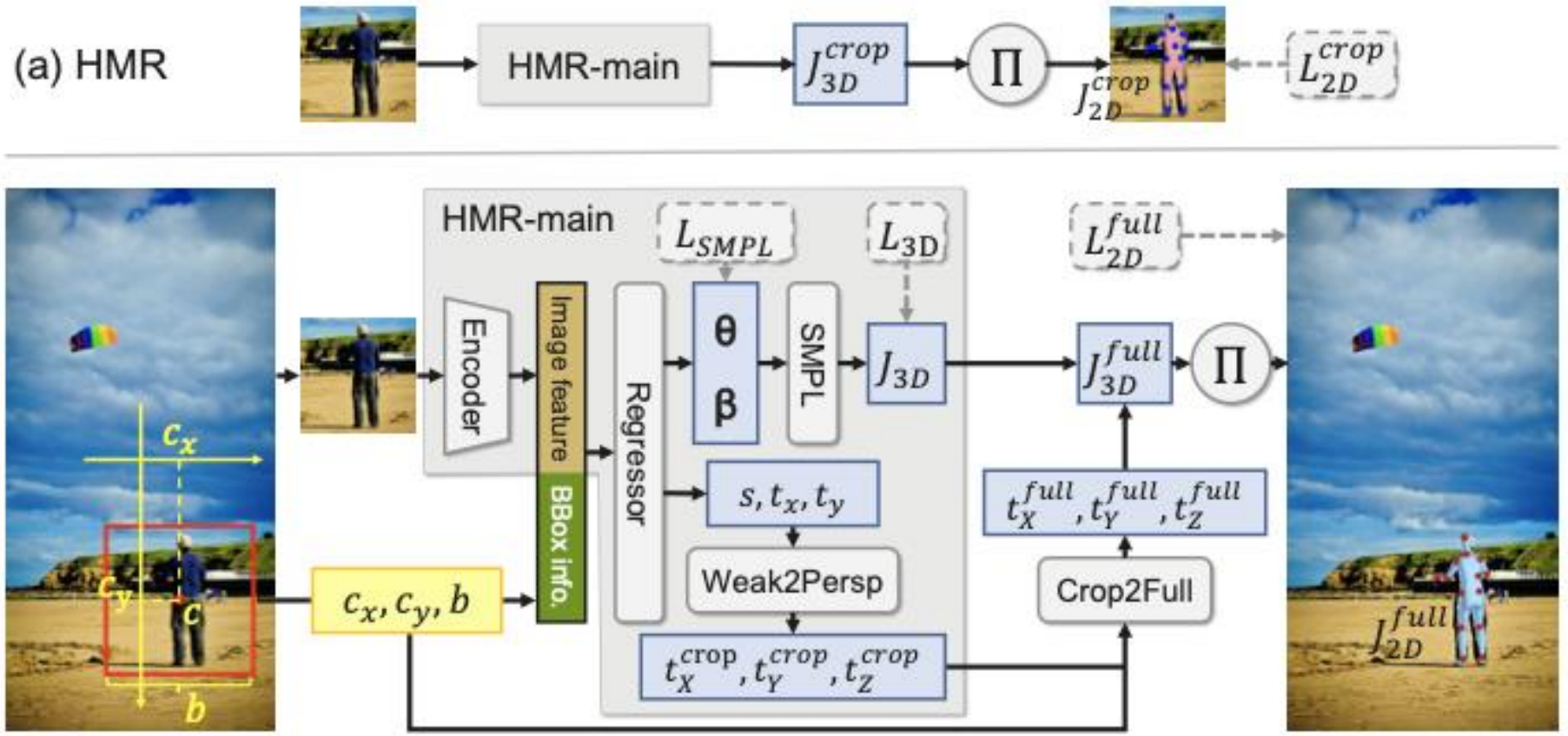
-Jathushan Rajasegaran (X.AI)

Appendix



[ECCV 2022] CLIFF

Appendix



[ECCV 2022] CLIFF

Appendix

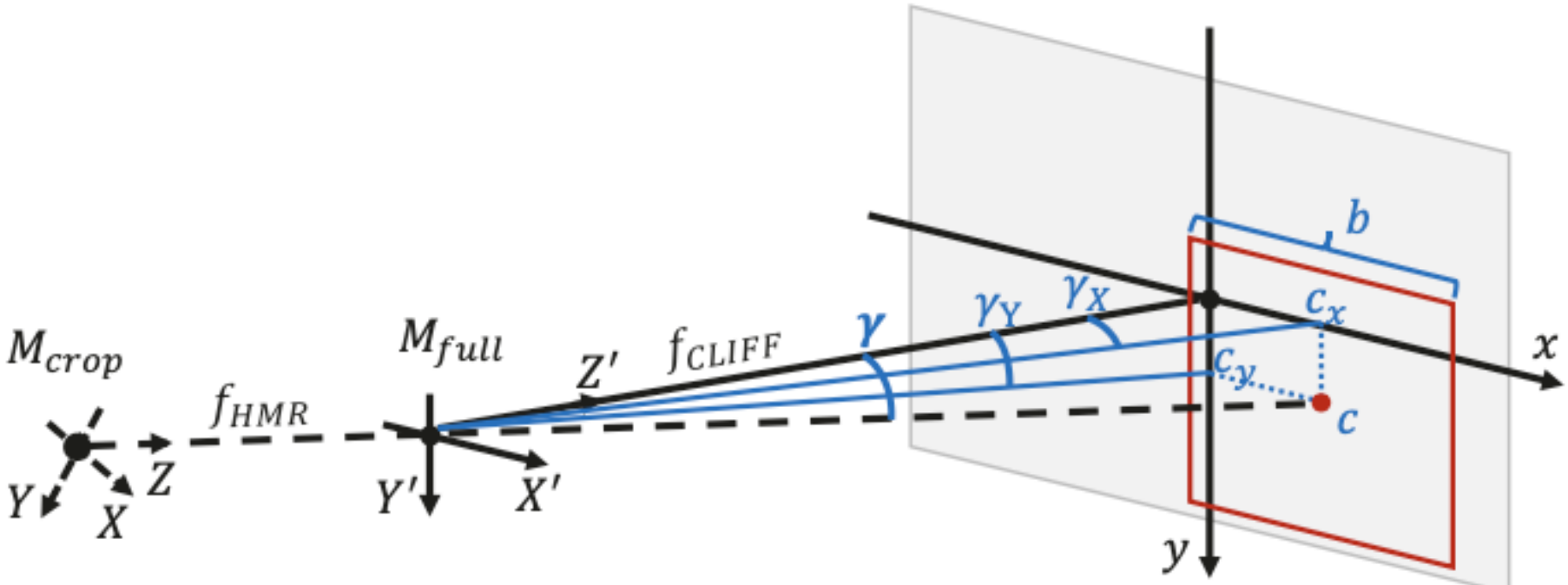
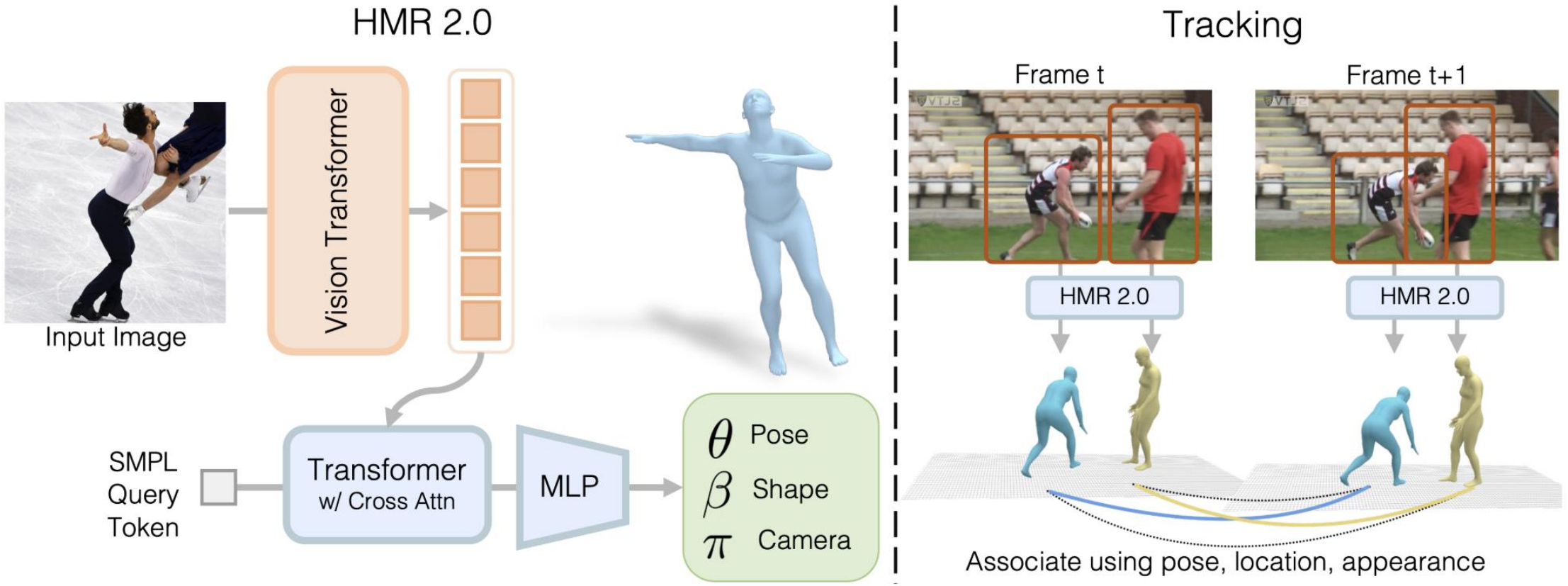


Fig. 3. The geometric relation between the virtual camera M_{crop} for the cropped image (the red rectangle) and the original camera M_{full} for the full image.

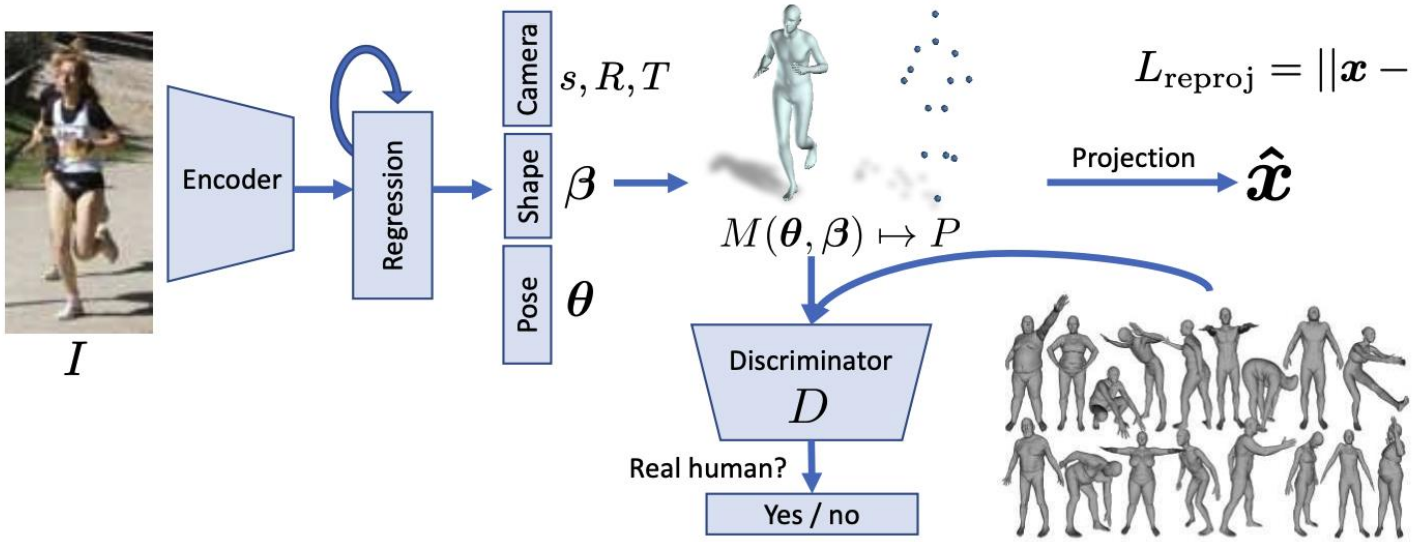
[ECCV 2022] CLIFF

Appendix



[ICCV 2023] HMR 2.0 (Human in 4D)

Appendix



$$L_{\text{reproj}} = ||x - \hat{x}||_2^2$$

$$\min L_{\text{adv}}(E) = \sum \mathbb{E}_{\Theta \sim p_E}[(D_i(E(I)) - 1)^2],$$

$$\min L(D_i) = \mathbb{E}_{\Theta \sim p_{\text{data}}}[(D_i(\Theta) - 1)^2] + \mathbb{E}_{\Theta \sim p_E}[D_i(E(I))^2].$$

[CVPR 2018] HMR

HMR uses the prior of SMPL dataset.

Appendix

$$L_{\mathcal{G}} = L_{3D} + L_{2D} + L_{SMPL} + L_{adv}$$

$$L_{3D} = \sum_{t=1}^{T} \|X_t - \hat{X}_t\|_2,$$

$$L_{2D} = \sum_{t=1}^{T} \|x_t - \hat{x}_t\|_2,$$

$$L_{SMPL} = \|\beta - \hat{\beta}\|_2 + \sum_{t=1}^{T} \|\theta_t - \hat{\theta}_t\|_2$$

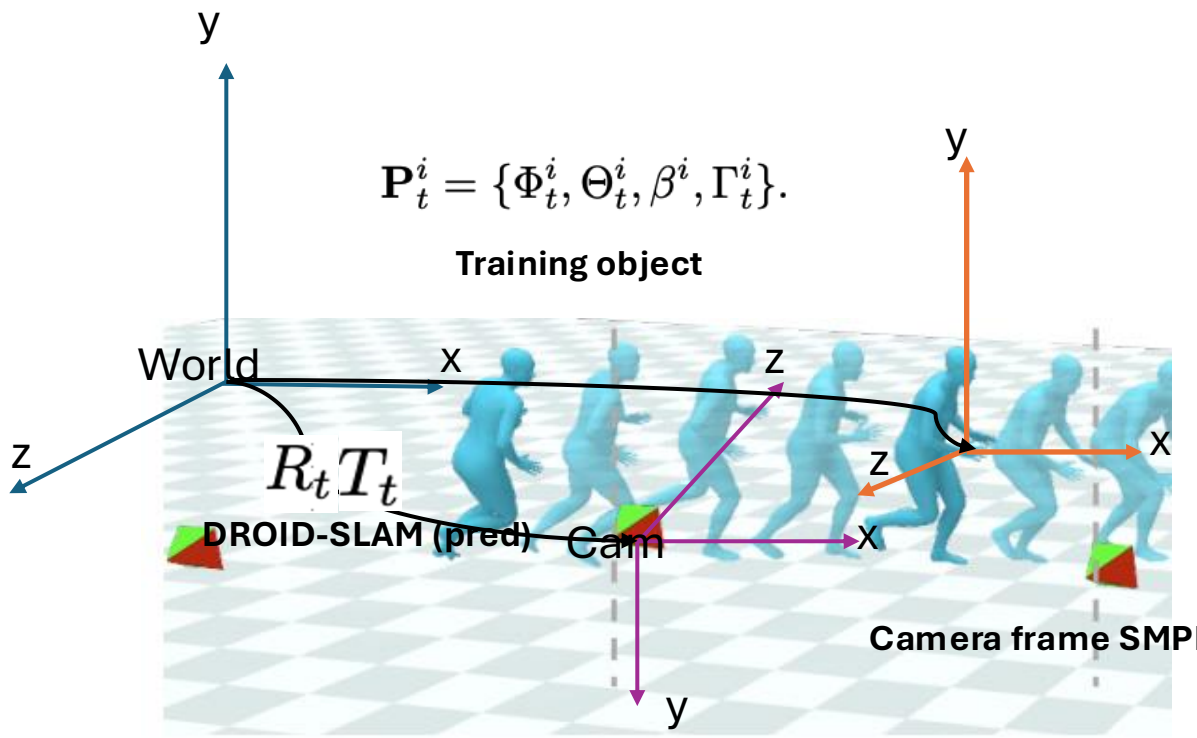
$$L_{adv} = \mathbb{E}_{\Theta \sim p_G}[(\mathcal{D}_M(\hat{\mathbf{\Theta}}) - 1)^2]$$

$$L_{\mathcal{D}_M} = \mathbb{E}_{\Theta \sim p_R}[(\mathcal{D}_M(\mathbf{\Theta}) - 1)^2] + \mathbb{E}_{\Theta \sim p_G}[\mathcal{D}_M(\hat{\mathbf{\Theta}})^2]$$

Dis: Prior loss

[CVPR 2020] VIBE

Appendix



$$\mathbf{P}_t^i = \{\Phi_t^i, \Theta_t^i, \beta^i, \Gamma_t^i\}.$$

Training object

projected 2D keypoints (pred)

$$E_{\text{data}} = \sum_{i=1}^{N} \sum_{t=1}^{T} \psi_t^i \rho\big(\Pi_K(R_t \cdot {}^{\text{w}}\mathbf{J}_t^i + \alpha T_t) - \mathbf{x}_t^i\big),$$

DROID-SLAM (pred)

$${}^{\text{w}}\mathbf{J}_t^i = \mathcal{M}({}^{\text{w}}\Phi_t^i, \Theta_t^i, \beta^i) + {}^{\text{w}}\Gamma_t^i.$$

$${}^{\text{w}}\Phi_t^i = R_t^{-1\text{c}} \hat{\Phi}_t^i, \qquad {}^{\text{w}}\Gamma_t^i = R_t^{-1\text{c}} \hat{\Gamma}_t^i - \alpha R_t^{-1} T_t,$$

$$\beta_i = \hat{\beta}^i, \qquad \Theta_t^i = \hat{\Theta}_t^i,$$

Camera frame SMPL (pred from PHARP)

SLAHMR is global optimization method.

Appendix

$$\mathbf{P}_t^i = \{\Phi_t^i, \Theta_t^i, \beta^i, \Gamma_t^i\}.$$

$$[\mathbf{V}_t^i, \mathbf{J}_t^i] = \mathcal{M}(\Phi_t^i, \Theta_t^i, \beta^i) + \Gamma_t^i.$$

$$\text{camera motion} \circ \text{human motion} = \text{net motion}.$$

$$E_{\text{data}} = \sum_{i=1}^{N} \sum_{t=1}^{T} \psi_t^i \rho\big(\Pi_K(R_t \cdot {}^{\text{w}}\mathbf{J}_t^i + \alpha T_t) - \mathbf{x}_t^i\big),$$

$$\min_{\{\{{}^{\text{w}}\Phi_t^i, {}^{\text{w}}\Gamma_t^i\}_{t=1}^T\}_{i=1}^N} \lambda_{\text{data}} E_{\text{data}}.$$

$$E_{\text{smooth}} = \sum_{i}^{N} \sum_{t}^{T} \|\mathbf{J}_t^i - \mathbf{J}_{t+1}^i\|^2.$$

$${}^{\text{w}}\Phi_t^i = R_t^{-1\text{c}} \hat{\Phi}_t^i, \qquad {}^{\text{w}}\Gamma_t^i = R_t^{-1\text{c}} \hat{\Gamma}_t^i - \alpha R_t^{-1} T_t,$$
$$\beta_i = \hat{\beta}^i, \qquad \Theta_t^i = \hat{\Theta}_t^i,$$

$$\min_{\alpha, \{\{{}^{\text{w}}\mathbf{P}_t^i\}_{t=1}^T\}_{i=1}^N} \lambda_{\text{data}} E_{\text{data}} + \lambda_\beta E_\beta + \lambda_{\text{pose}} E_{\text{pose}} + \lambda_{\text{smooth}} E_{\text{smooth}}.$$

$${}^{\text{w}}\mathbf{J}_t^i = \mathcal{M}({}^{\text{w}}\Phi_t^i, \Theta_t^i, \beta^i) + {}^{\text{w}}\Gamma_t^i.$$

SLAHMR is global optimization method.

Appendix

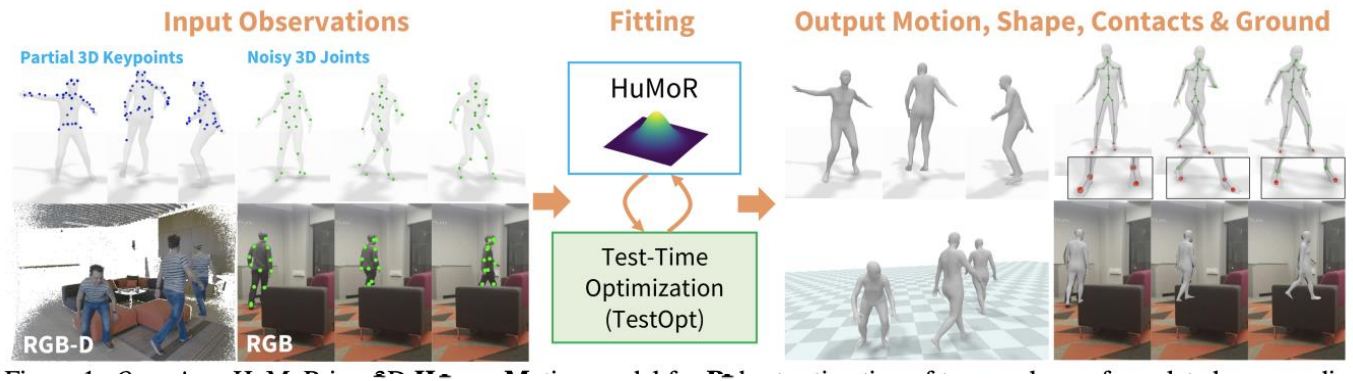
$$p_\theta(\mathbf{s}_t|\mathbf{s}_{t-1}) = \int_{\mathbf{z}_t} p_\theta(\mathbf{z}_t|\mathbf{s}_{t-1}) p_\theta(\mathbf{s}_t|\mathbf{z}_t, \mathbf{s}_{t-1})$$

$$E_{\text{CVAE}} = -\sum_i^N \sum_t^T \log \mathcal{N}(\mathbf{z}_t^i; \mu_\theta(\mathbf{s}_{t-1}^i), \sigma_\theta(\mathbf{s}_{t-1}^i)).$$

$$\mathbf{z}_t^i = \mu_\phi(\mathbf{s}_t^i, \mathbf{s}_{t-1}^i), \quad \mathbf{s}_t^i = \mathbf{s}_{t-1}^i + \Delta_\theta(\mathbf{z}_t^i, \mathbf{s}_{t-1}^i).$$

$$E_{\text{skate}} = \sum_i^N \sum_t^T \sum_j^J c_t^i(j) \|\mathbf{J}_t^i(j) - \mathbf{J}_{t+1}^i(j)\|$$

$$E_{\text{con}} = \sum_i^N \sum_t^T \sum_j^J c_t^i(j) \max(d(\mathbf{J}_t^i(j), g) - \delta, 0).$$



$$E_{\text{env}} = \lambda_{\text{skate}} E_{\text{skate}} + \lambda_{\text{con}} E_{\text{con}}.$$

$$\min_{\substack{\alpha, g, \{\mathbf{s}_0^i\}_{i=1}^N, \\ \{\{\mathbf{z}_t^i\}_{t=1}^T\}_{i=1}^N}} \lambda_{\text{data}} E_{\text{data}} + \lambda_\beta E_\beta + \lambda_{\text{pose}} E_{\text{pose}} + E_{\text{prior}} + E_{\text{env}}.$$

[CVPR 2023] SLAHMR (loss from AMASS, HuMoR)

Appendix

$\{I^t\}_{t=0}^{T}$

$\{\theta^t \in \mathbb{R}^{21\times 3}\}_{t=0}^{T}$

$\beta \in \mathbb{R}^{10}$

$\{\Gamma_c^t \in \mathbb{R}^3\}_{t=0}^{T}$

$\{\tau_c^t \in \mathbb{R}^3\}_{t=0}^{T}$

$\{\Gamma_w^t \in \mathbb{R}^3\}_{t=0}^{T}$

$\{\tau_w^t \in \mathbb{R}^3\}_{t=0}^{T}$

$$\tau_w^t = \begin{cases} [0,0,0]^T, & t = 0, \\ \sum_{i=0}^{t-1} \Gamma_w^i v_{root}^i, & t > 0. \end{cases}$$

$$\Gamma_w^t = \begin{cases} \Gamma_{GV}^0, & t = 0, \\ \prod_{i=1}^{t} R_{\Delta GV}^i \cdot \Gamma_{GV}^t, & t > 0. \end{cases}$$

[SIGGRAPH 2024] GVHMR