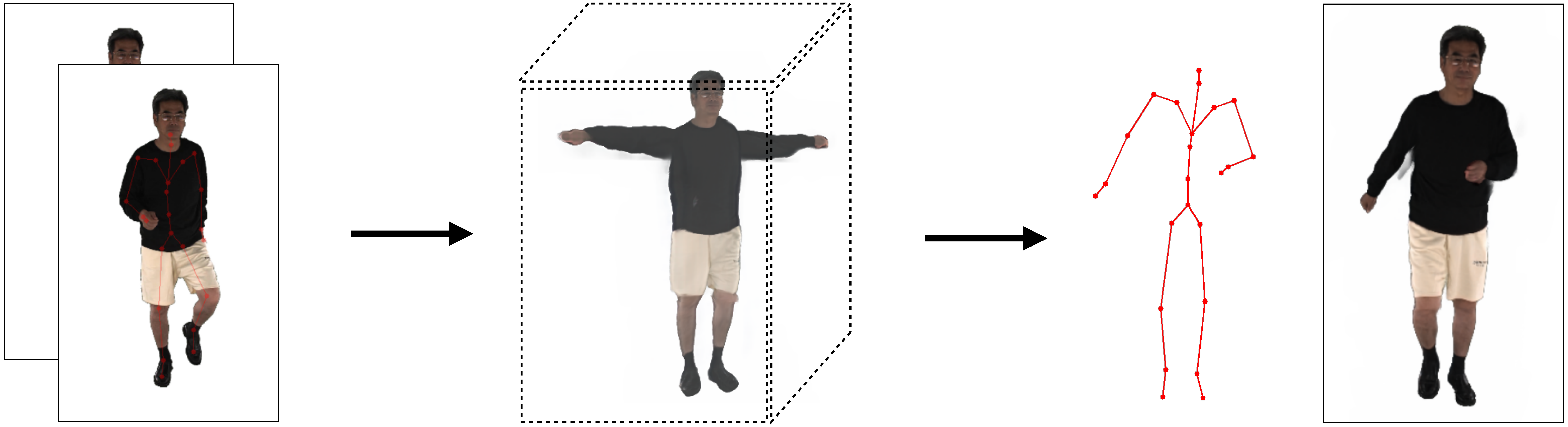
Generalized Animatable Human NeRFs from Few Views

Jakub Zadrozny, Hakan Bilen
University of Edinburgh

BMVA Symposium on 3D Vision
London | 26.02.2025

Dynamic free-viewpoint human rendering



Input: a few monocular views with pose parameters

Estimate: subject NeRF model in canonical pose

Input: target pose parameters

Output: render in target pose

Key applications include: augmented/virtual reality, movie production and immersive 3D communication.

Subject-specific vs. generalized

Subject-specific approaches

−

Requires test-time optimization

Needs extensive observations (typically ca. 30 frames)

+

Usually most accurate

Does not require large training set

Weng et al., *HumanNeRF,* CVPR 2022

Yu et al., *MonoHuman,* CVPR 2023

Hu et al., *GauHuman,* CVPR 2024

Generalized approaches (ours)

+

Uses only feed-forward passes during inference

Requires less observed views (1-3 for our model)

Learns a prior, in-paints unobserved details

−

Typically less accurate compared to subject-specific

Hu et al., *SHERF,* ICCV 2023

Li et al., *GHuNeRF,* 3DV 2024

Masuda et al., *GNH,* arxiv 2024

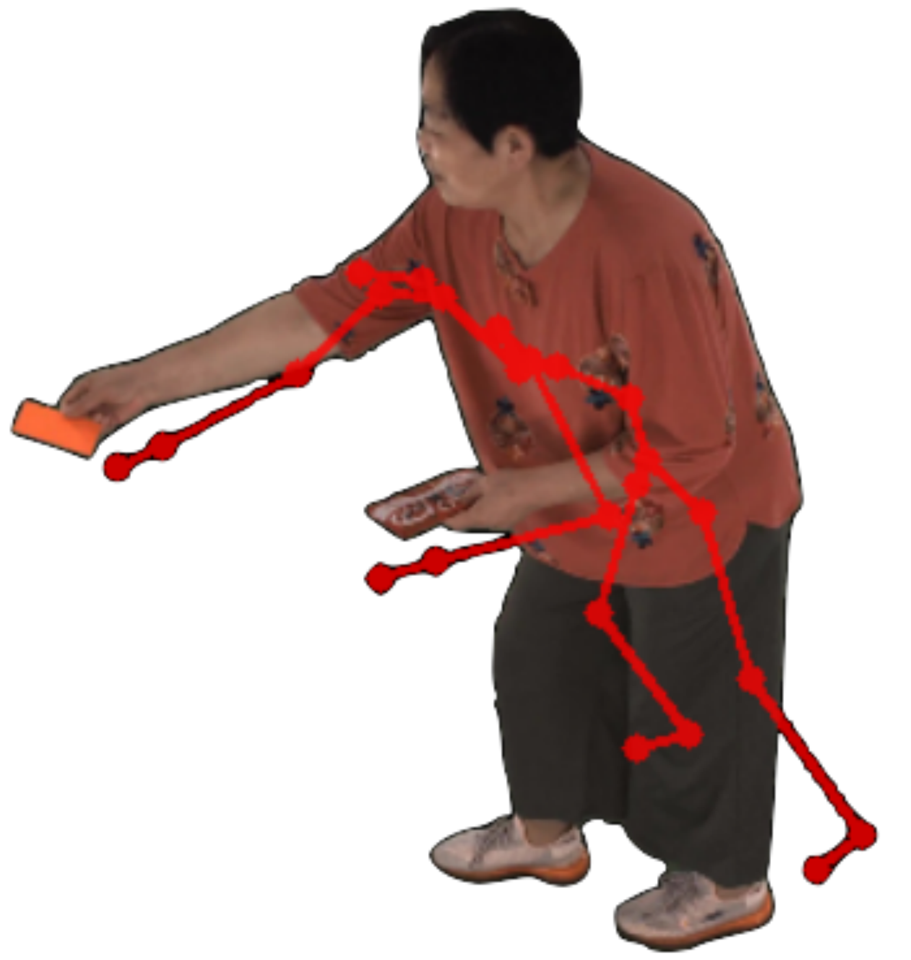
Ours

Our contributions

Limitations of existing work		Our approach
conditioning only on a single observation or on an entire video	→	improves rendering quality when given additional views
relying heavily on accurate body pose and parameters	→	improves robustness to inaccurarcies in body parameters
struggling to model loose clothing	→	still a limitation, despite some improvements in architecture

Accurate pose parameters

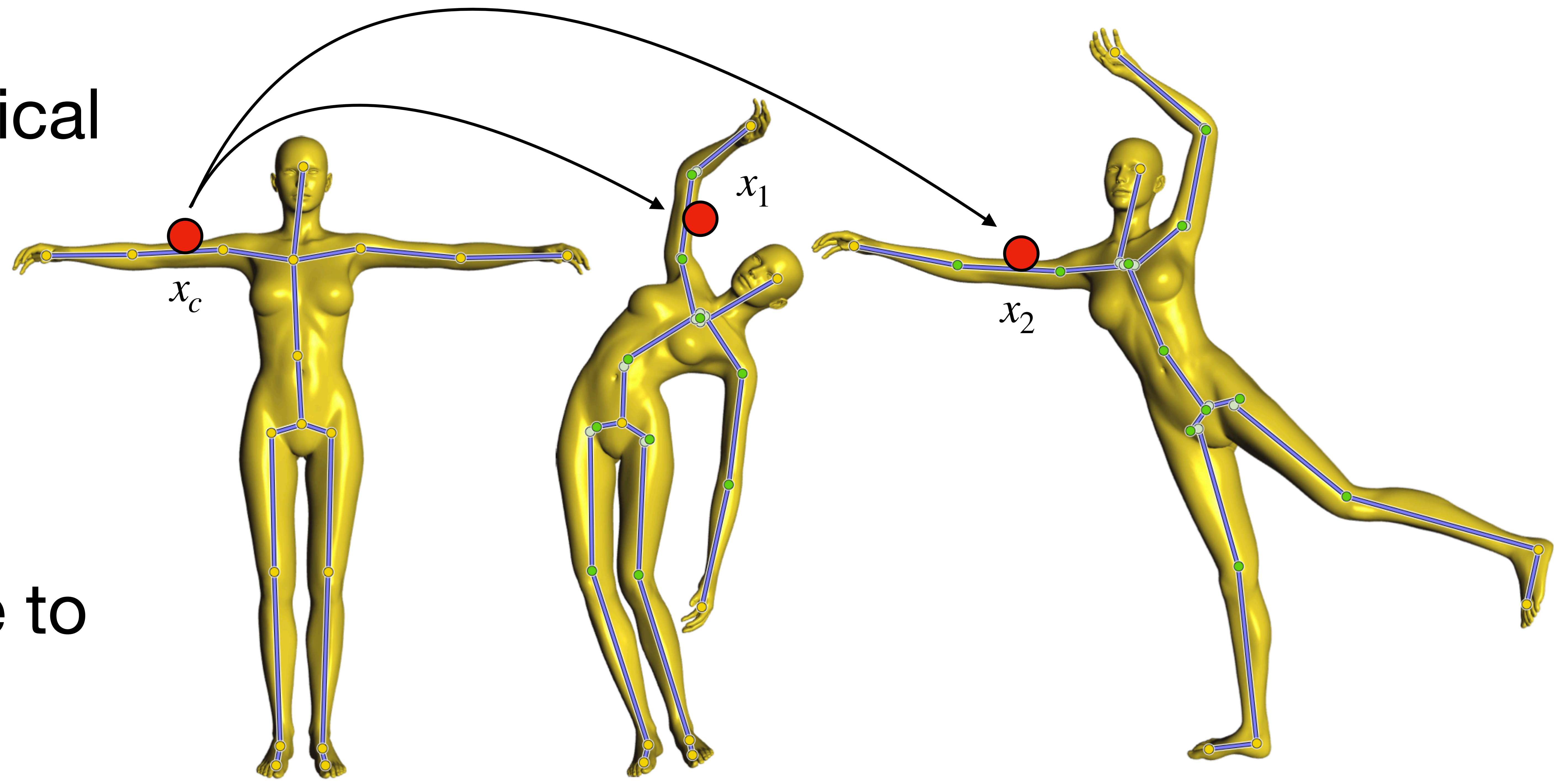
Estimated pose parameters



Parameters estimated using Li et al., *Hybrik,* CVPR 2021

Linear blend skinning

- Transforms points from the canonical pose to any given target pose
- Generally pre-defined*
- **Input:** motion/blend weights
  - softly assign points in 3D space to bones
  - *can be learned
- Can be approximately inverted



Input: skeleton shape in canonical pose

Input: joint rotations θ_1

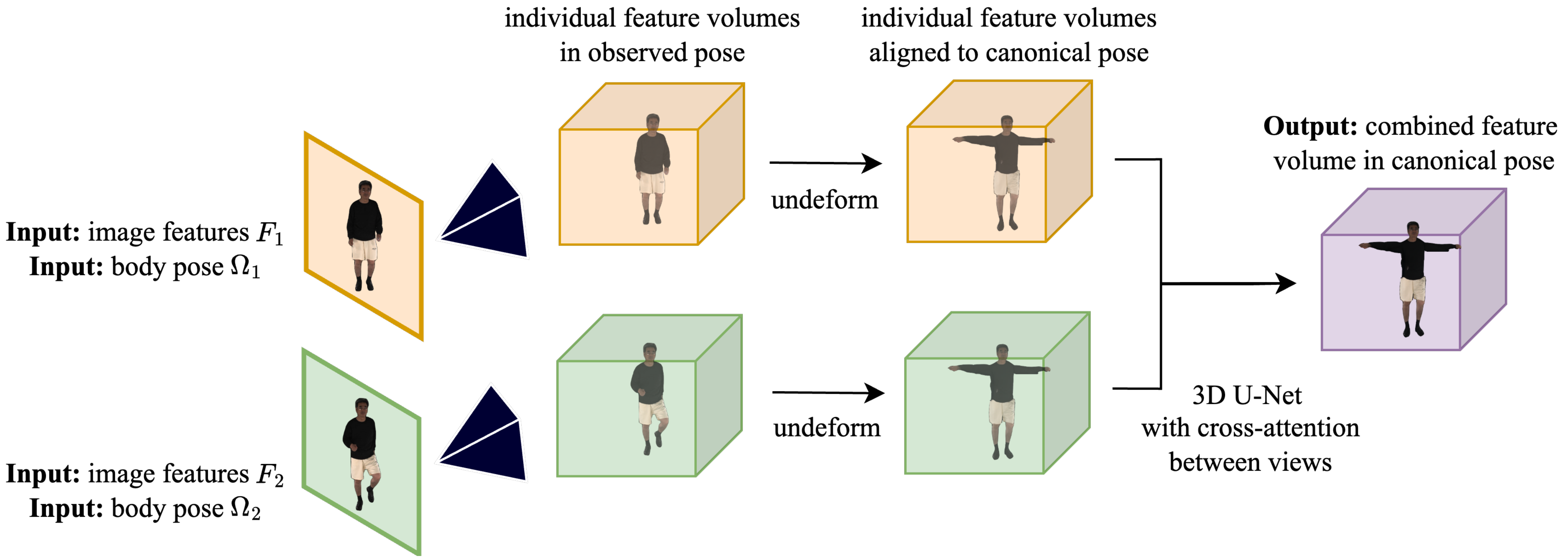
Input: joint rotations θ_2

$$LBS(x_c, \theta_i) = x_i$$

$$LBS^{-1}(x_i, \theta_i) \approx x_c$$

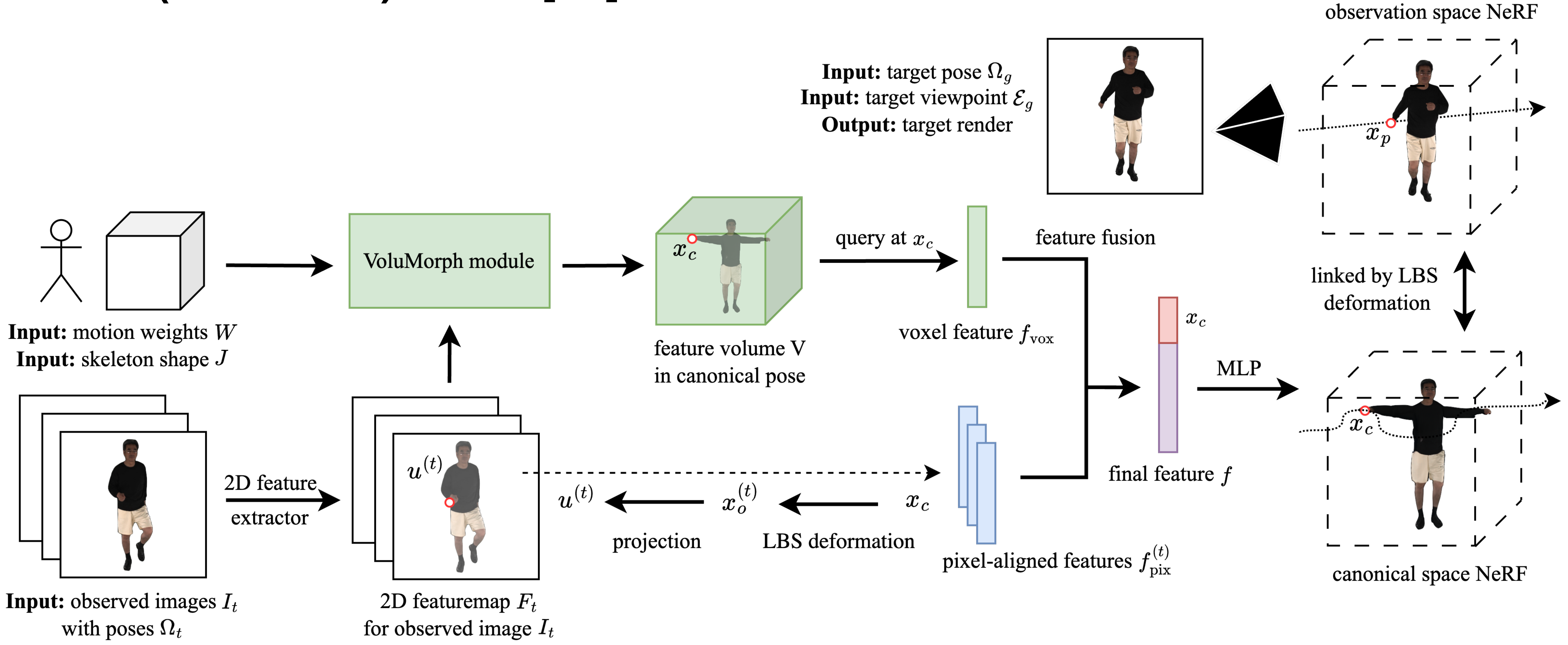
Jacobson et al., *Fast automatic skinning transformations,* ToG 2012

The VoluMorph module

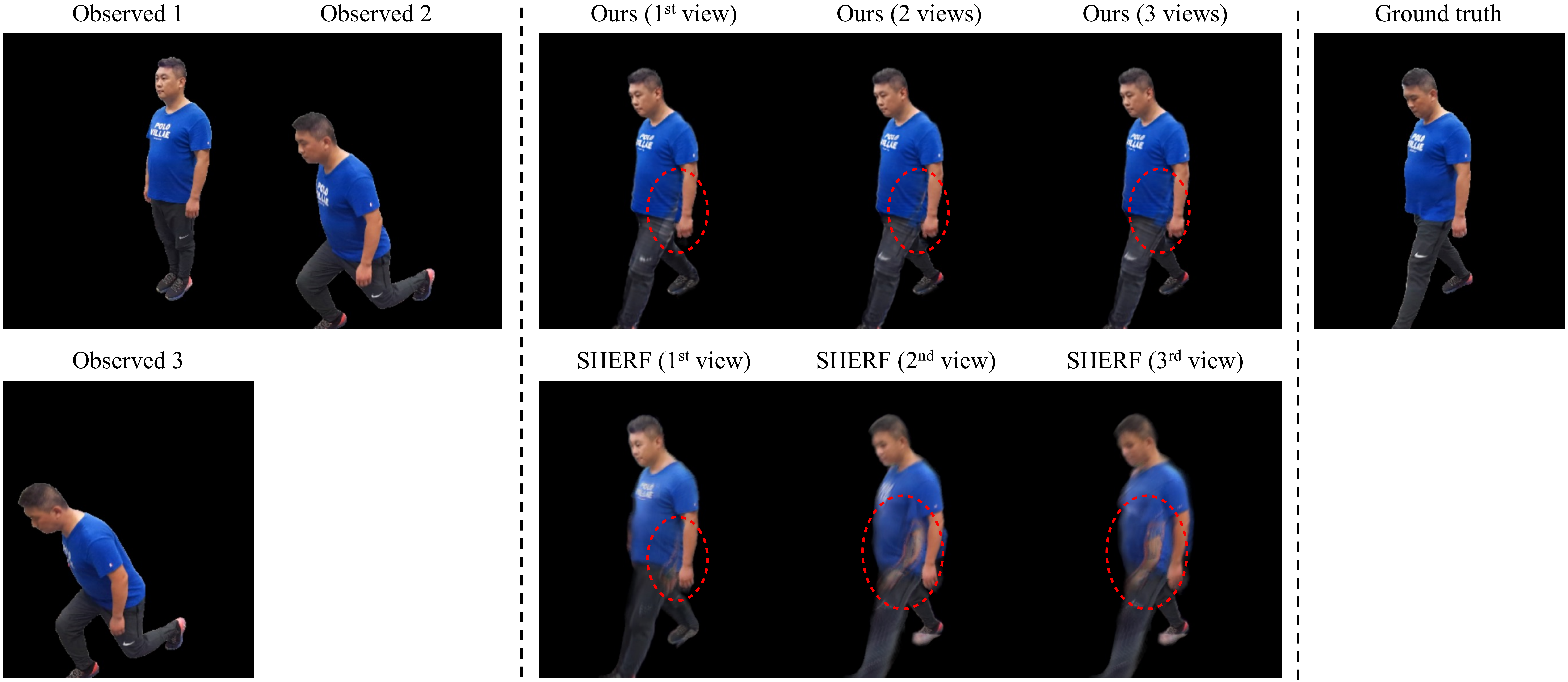


3D U-Net architecture from Szymanowicz et al., *Viewset Diffusion,* ICCV 2023.

Our (almost) full pipeline

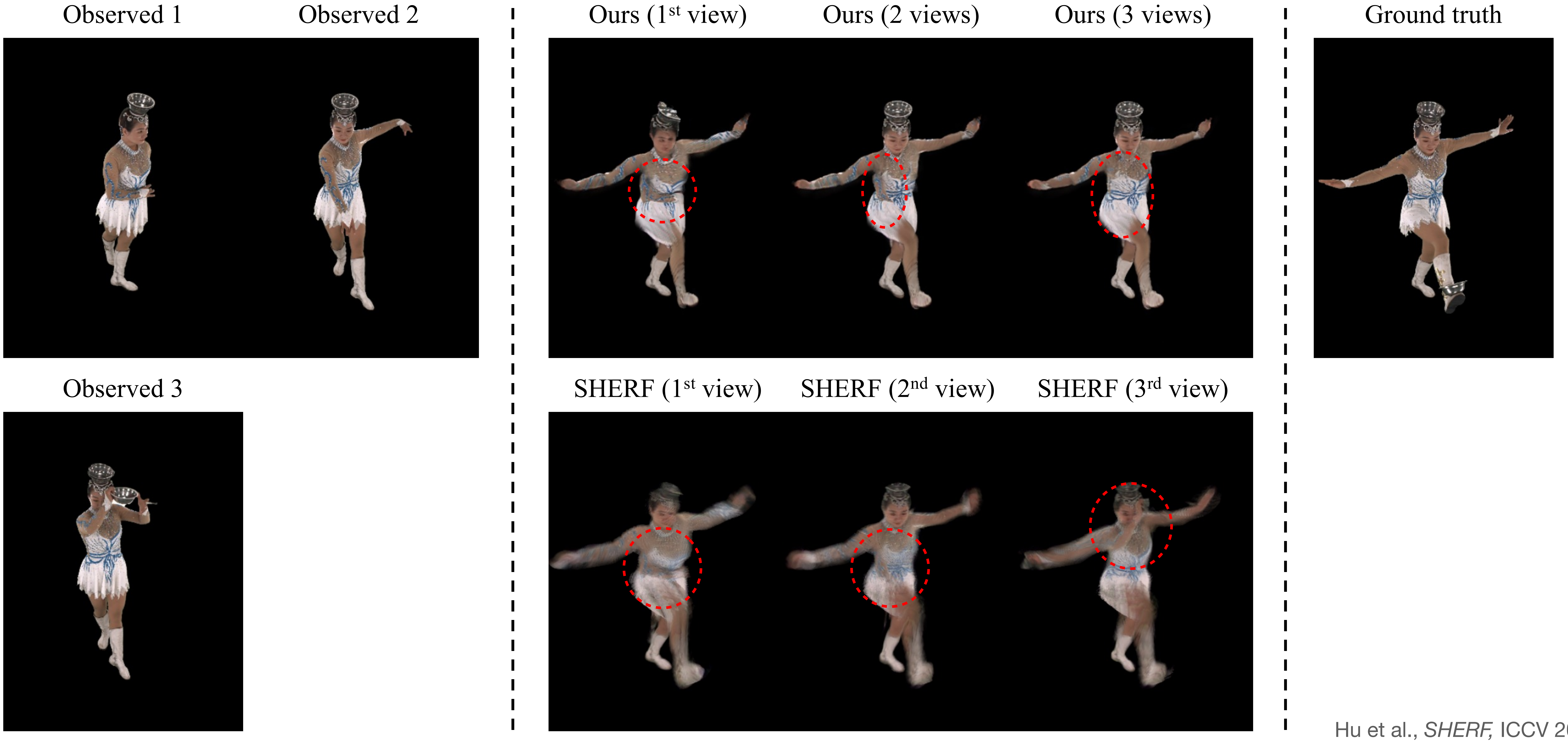


Qualitative results on HuMMan dataset



Hu et al., *SHERF,* ICCV 2023

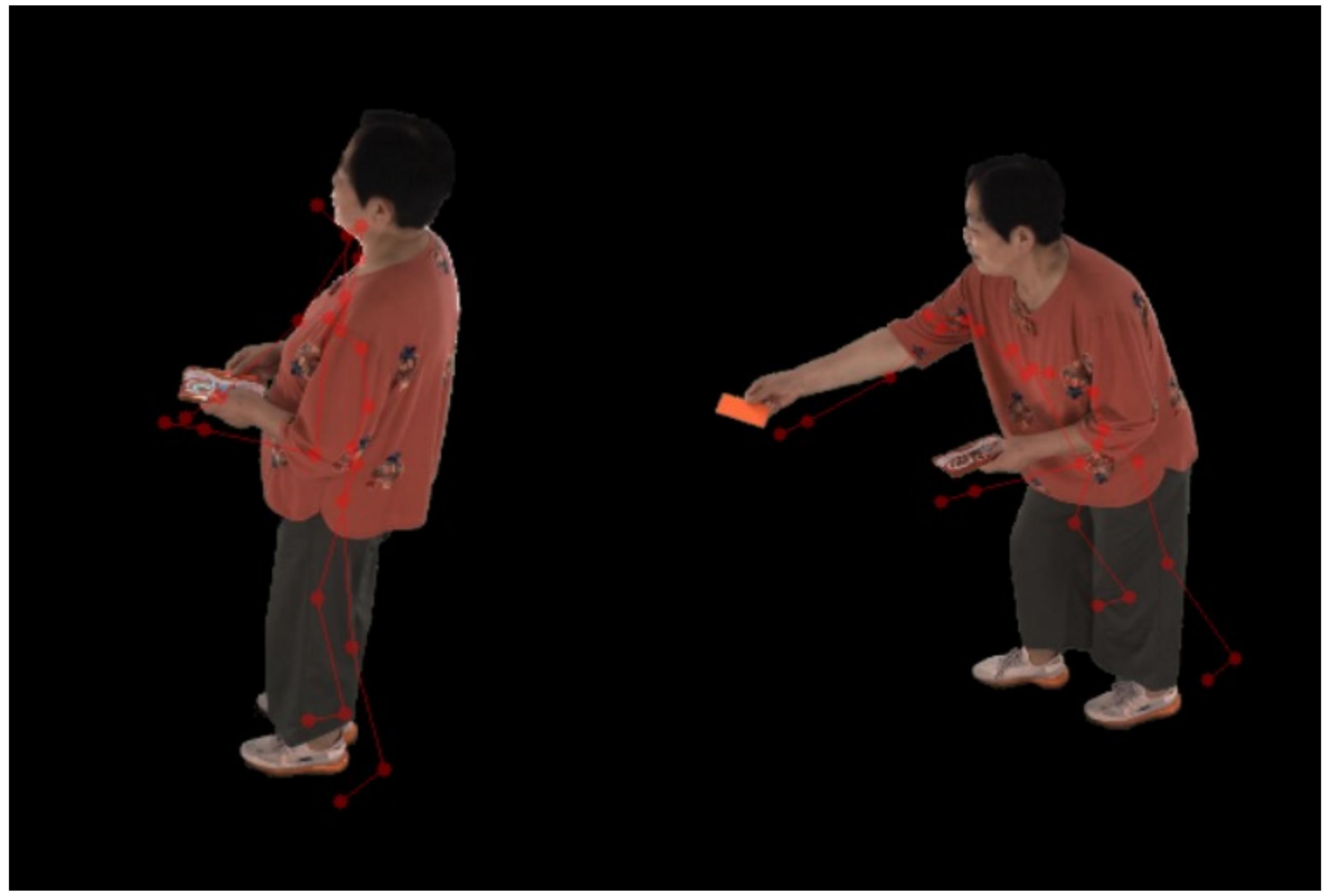
Qualitative results on DNA-Rendering dataset



Hu et al., *SHERF,* ICCV 2023

Qualitative results using estimated body parameters

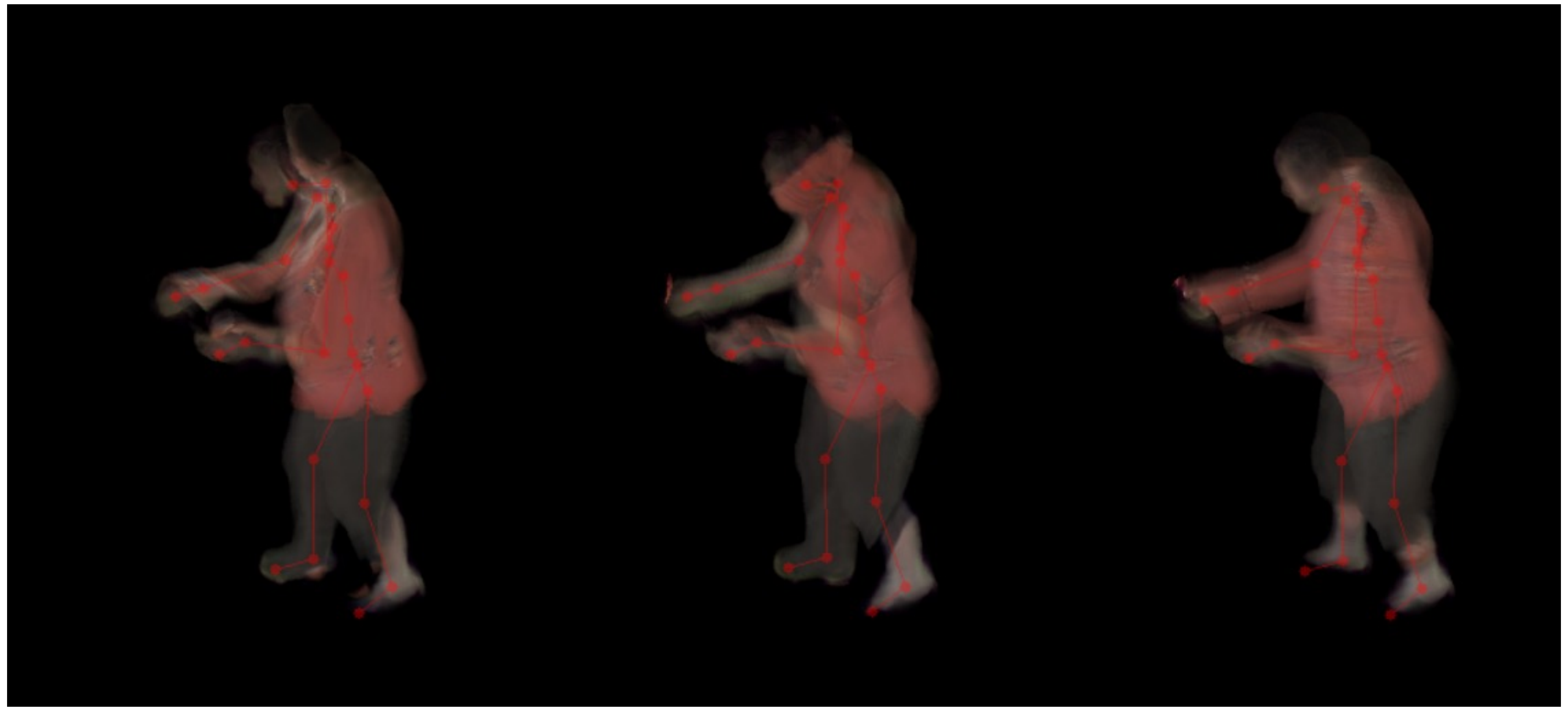
Observed 1 Observed 2



Observed 3

Ours (1^{st} view) Ours (2 views) Ours (3 views)

SHERF (1^{st} view) SHERF (2^{nd} view) SHERF (3^{rd} view)



Ground truth

Parameters estimated using Li et al., *Hybrik,* CVPR 2021

Hu et al., *SHERF,* ICCV 2023

Quantitative results

Dataset	Motion seqs. (train — test)	Actors	Cameras
HuMMan	339 (317 — 22)	153	10
DNA-Rendering	436 (372 — 64)	136	48

Method	HuMMan		DNA-Rendering			
	accurate body params		*accurate body params*		*estimated body params*	
	PSNR↑	LPIPS*↓	PSNR↑	LPIPS*↓	PSNR↑	LPIPS*↓
Ours (1 view)	26.62	34.24	27.83	40.27	**27.43**	**44.12**
Ours (2 views)	27.31	30.90	28.31	37.80	26.87	48.26
Ours (3 views)	**27.56**	29.59	**28.59**	**36.68**	27.04	47.22
Ours (4 views)	**27.56**	**29.53**	28.58	36.82	27.06	47.10
SHERF	26.87	45.03	28.49	48.22	26.93	61.97
GHuNeRF	23.88	66.27	27.80	70.84	27.19	75.13

- We evaluate rendering in **novel poses**
- Improvement on the perceptual metric, close results on PSNR (note: PSNR favours oversmoothed results over slight misalignments)
- Accurate body parameters: quality improves from 1 observation to 3, then saturates
- Estimated body parameters: best quality achieved using 1 observation

Li et al., *GHuNeRF,* 3DV 2024 Hu et al., *SHERF,* ICCV 2023

Limitations and future work

- **Limitations:**
  - our approach still struggles with estimated pose parameters
  - we do not model loose clothing accurately
- **Future work:**
  - further improve robustness to estimated body parameters and investigate the use of estimated camera poses
  - address modelling loose clothing and interactions with objects
  - improve efficiency (Gaussian splatting?)
- **Thank you!**