CVPR
JUNE 17-21, 2024
SEATTLE, WA

HUGS: Human Gaussian Splats

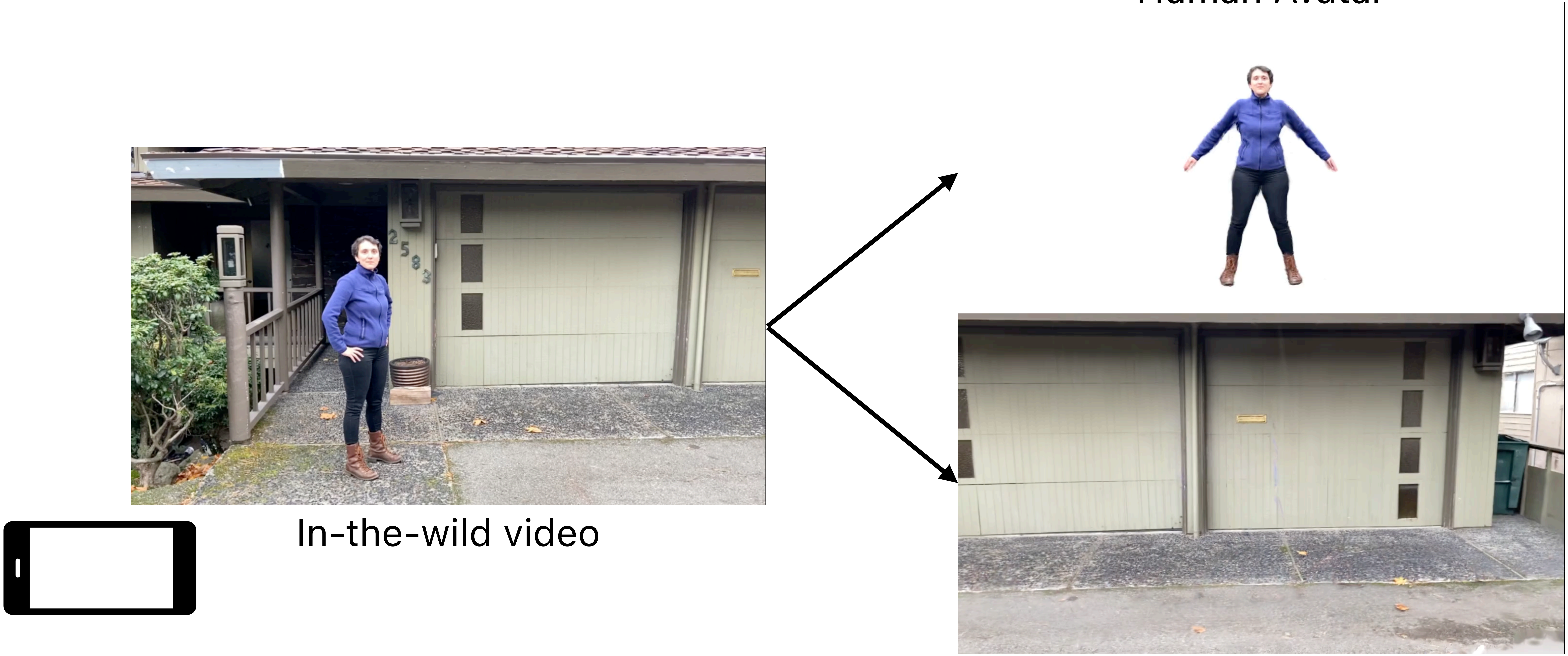
Muhammed Kocabas[1,2,3], Rick Chang[1], James Gabriel[1],
Oncel Tuzel[1], Anurag Ranjan[1]

1 2 MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS 3 ETH zürich

https://machinelearning.apple.com/research/hugs

Goal

Animatable humans and scene view synthesis



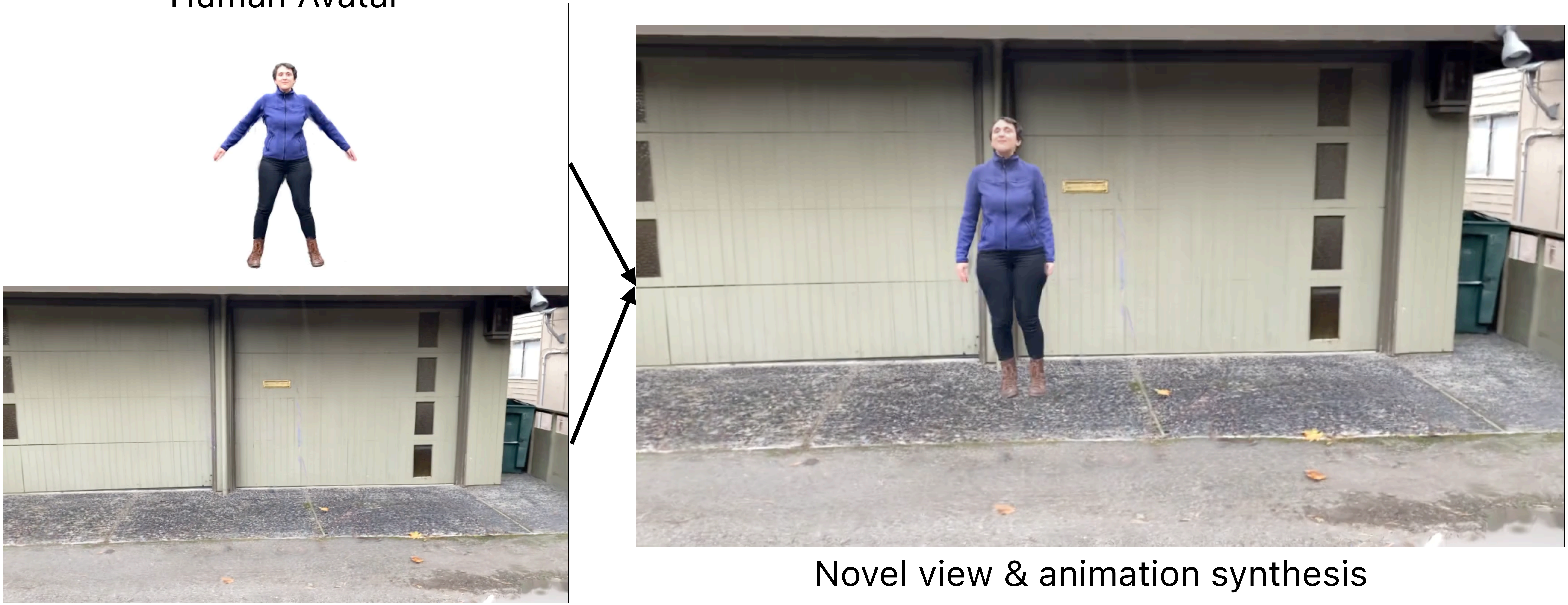
Human Avatar

In-the-wild video

Scene view synthesis

Goal

Animatable humans and scene view synthesis



Human Avatar

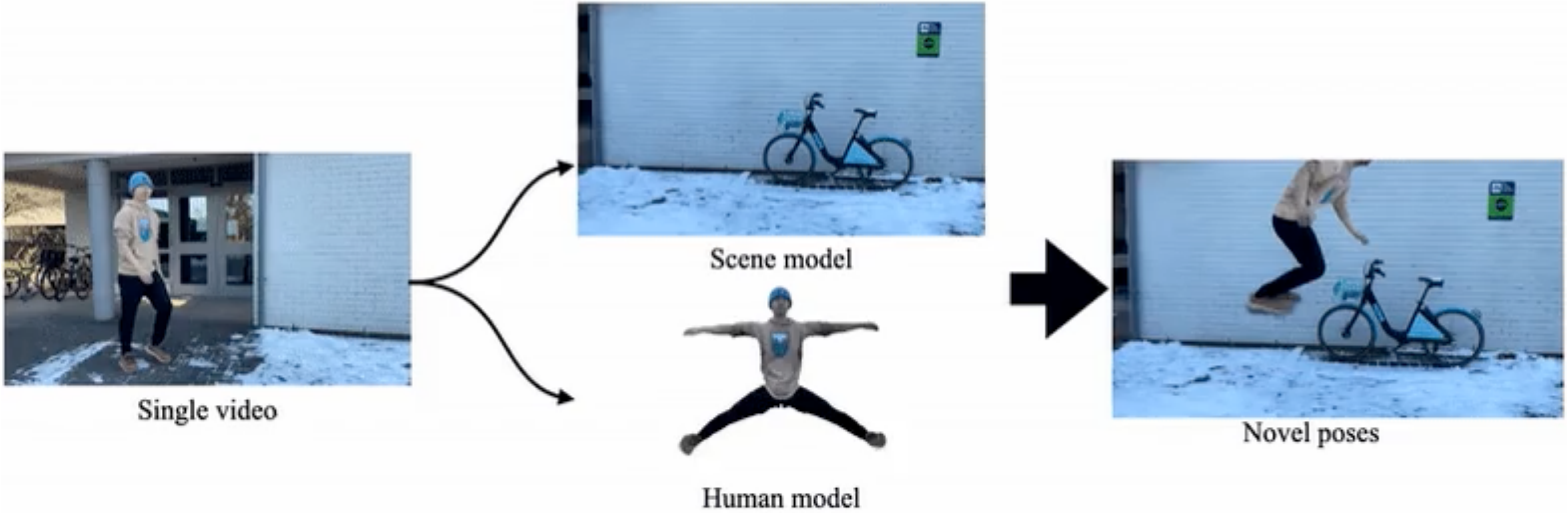
Scene view synthesis

Novel view & animation synthesis

Problem

Existing NeRF-based approaches are slow

- NeuMan: Neural Human Radiance Field from a Single Video, Jiang etal., ECCV 2020
- Train time: 3-7 days
- Render time (HD): 4 mins



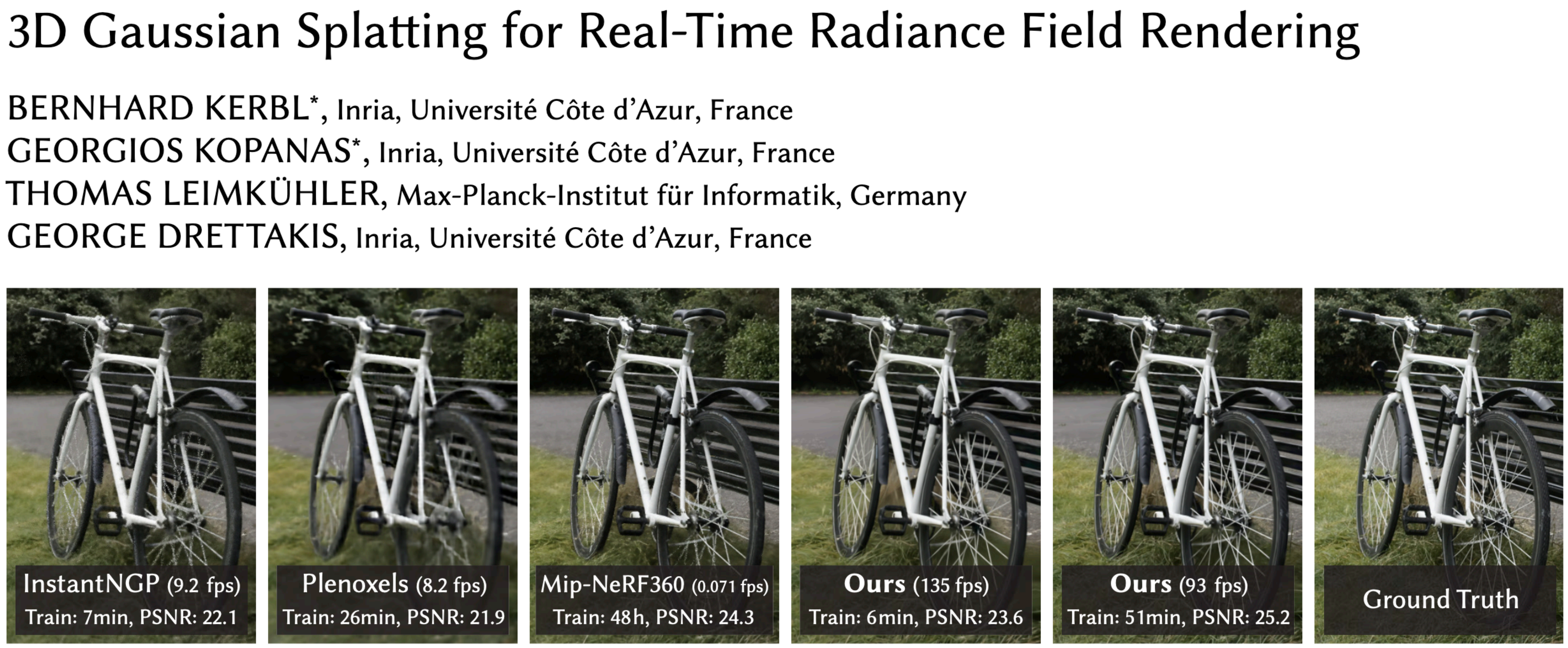
Problem

Existing NeRF-based approaches are slow

- InstantAvatar: Learning Avatars from Monocular Video, Jiang etal., CVPR 2023
- Train time: 15-20 mins
- Render time (HD): 0.5 FPS

Problem

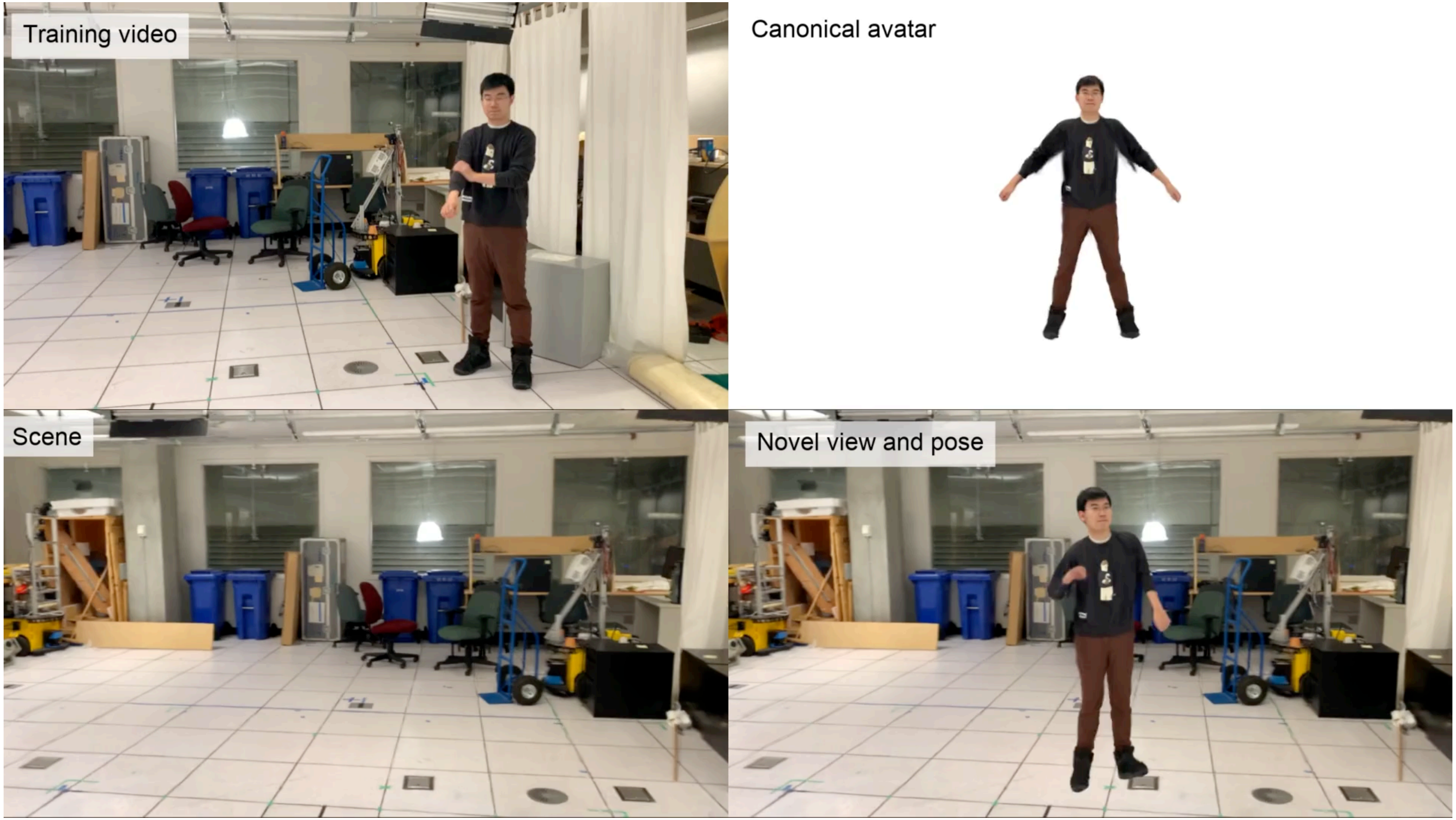
3DGS is fast with realtime rendering speed, but not animatable



3D Gaussian Splatting for Real-Time Radiance Field Rendering

BERNHARD KERBL*, Inria, Université Côte d'Azur, France
GEORGIOS KOPANAS*, Inria, Université Côte d'Azur, France
THOMAS LEIMKÜHLER, Max-Planck-Institut für Informatik, Germany
GEORGE DRETTAKIS, Inria, Université Côte d'Azur, France

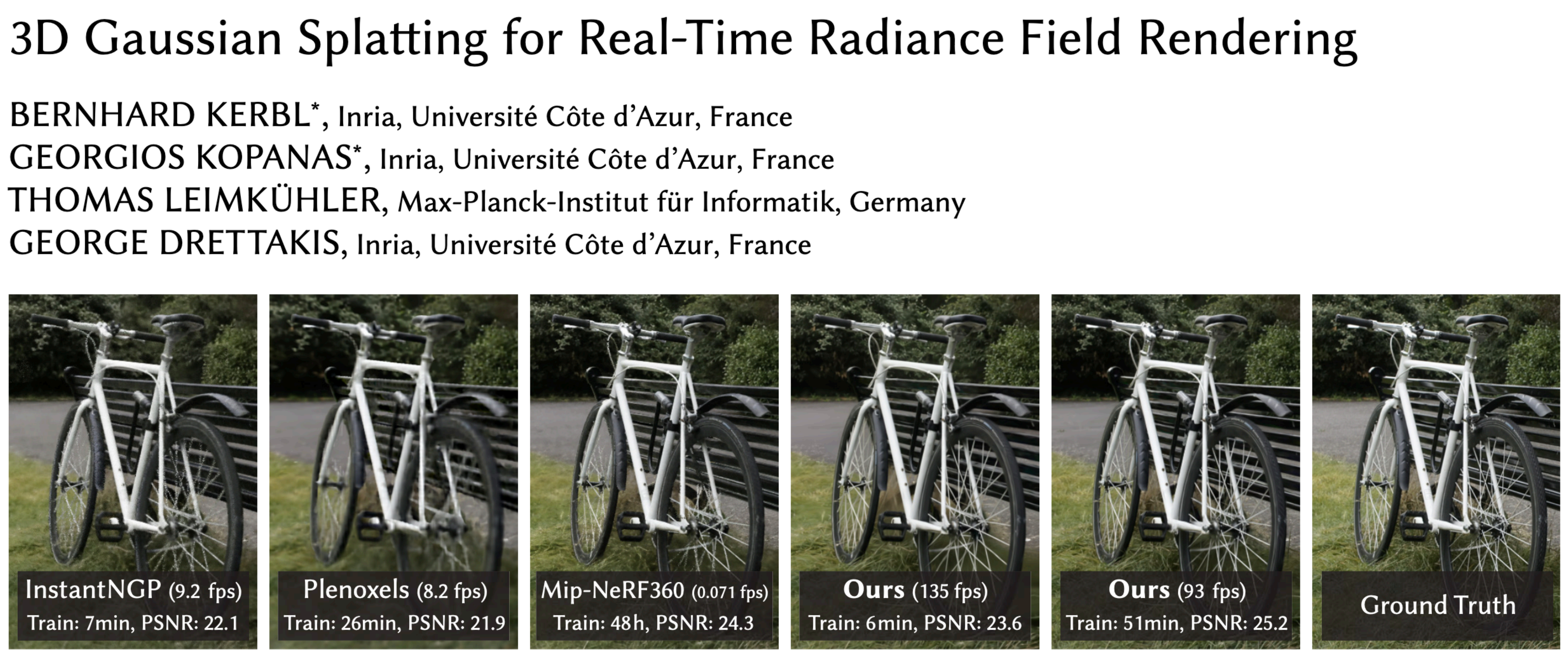
HUGS — Human Gaussian Splats



HUGS — Human Gaussian Splats

ÜRAL POR LEN

Preliminaries



3D Gaussian Splatting for Real-Time Radiance Field Rendering

BERNHARD KERBL*, Inria, Université Côte d'Azur, France
GEORGIOS KOPANAS*, Inria, Université Côte d'Azur, France
THOMAS LEIMKÜHLER, Max-Planck-Institut für Informatik, Germany
GEORGE DRETTAKIS, Inria, Université Côte d'Azur, France

InstantNGP (9.2 fps) Train: 7min, PSNR: 22.1

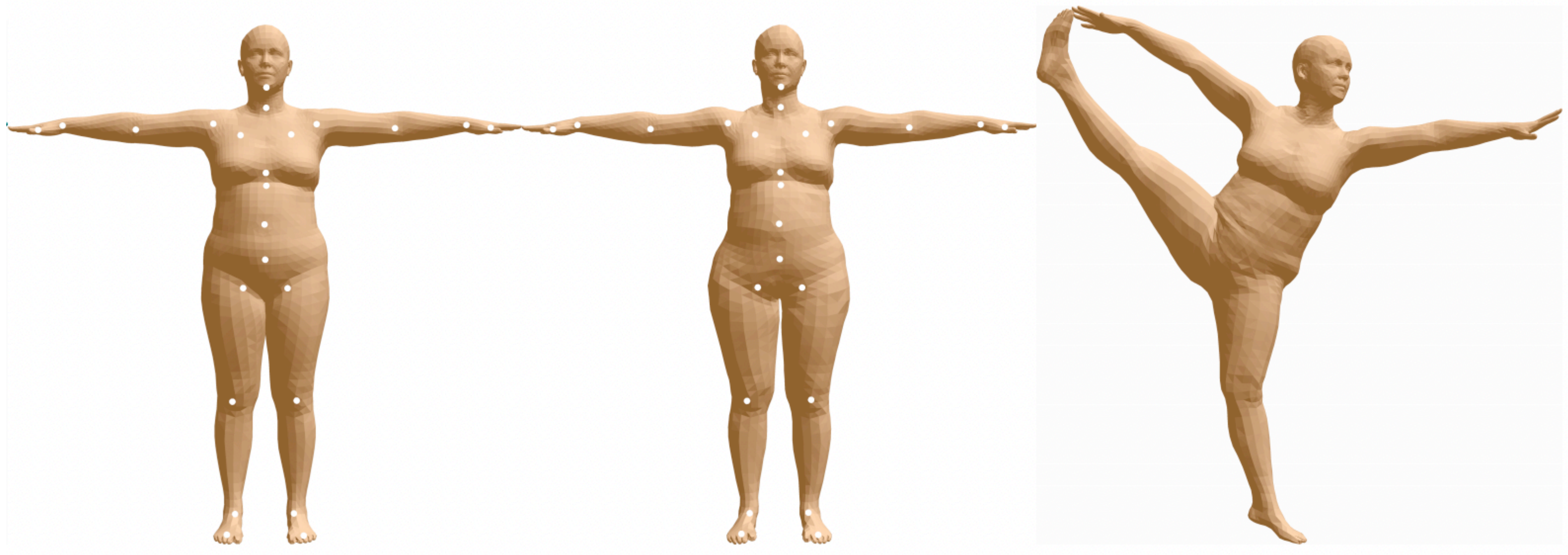
Plenoxels (8.2 fps) Train: 26min, PSNR: 21.9

Mip-NeRF360 (0.071 fps) Train: 48h, PSNR: 24.3

Ours (135 fps) Train: 6min, PSNR: 23.6

Ours (93 fps) Train: 51min, PSNR: 25.2

Ground Truth

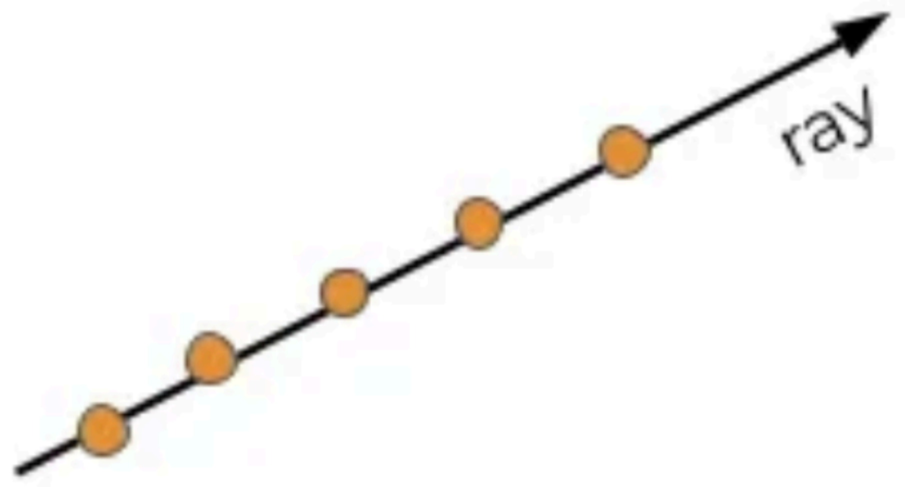


(b) $\bar{\mathbf{T}} + B_S(\vec{\beta}),\ J(\vec{\beta})$ **(c)** $T_P(\vec{\beta},\vec{\theta}) = \bar{\mathbf{T}}+B_S(\vec{\beta})+B_P(\vec{\theta})$ **(d)** $W(T_P(\vec{\beta},\vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W})$

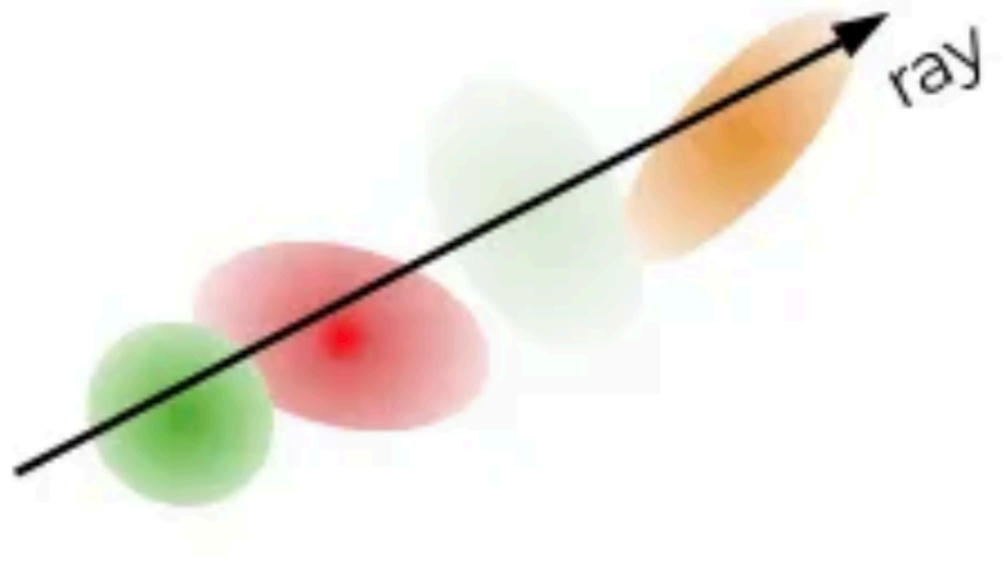
Preliminary: 3D Gaussian Splatting (3DGS)

- NeRF (Neural Radiance Fields):
  - a NN encodes the radiance field
  - Rendering is performed using raymarching (costly)

NeRF: raymarching



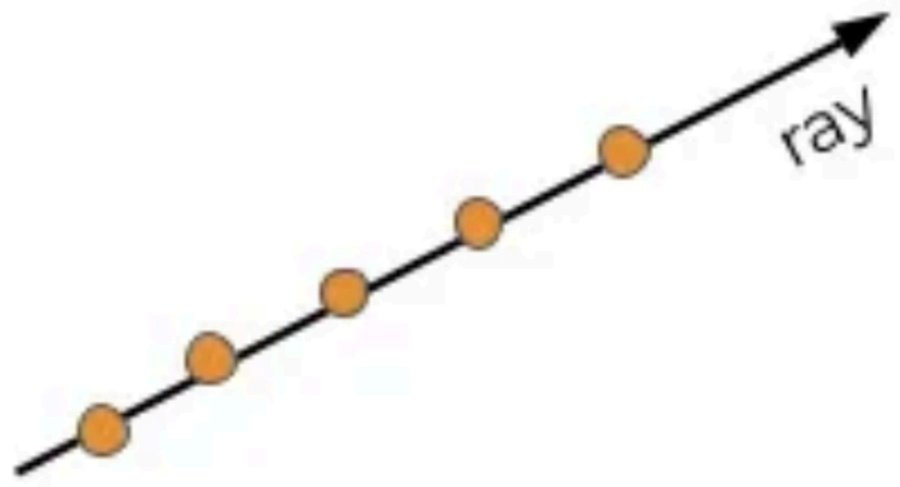
3DGS: rasterization



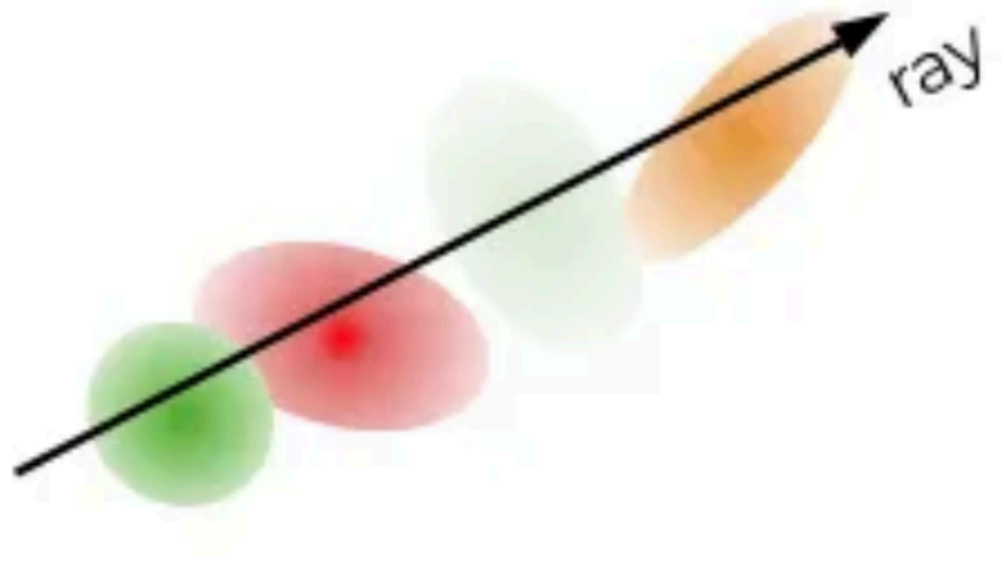
Preliminary: 3D Gaussian Splatting (3DGS)

- 3DGS
  - 3D Gaussian primitives encode the baked radiance field
  - Rendering is performed using rasterization (fast)

NeRF: raymarching

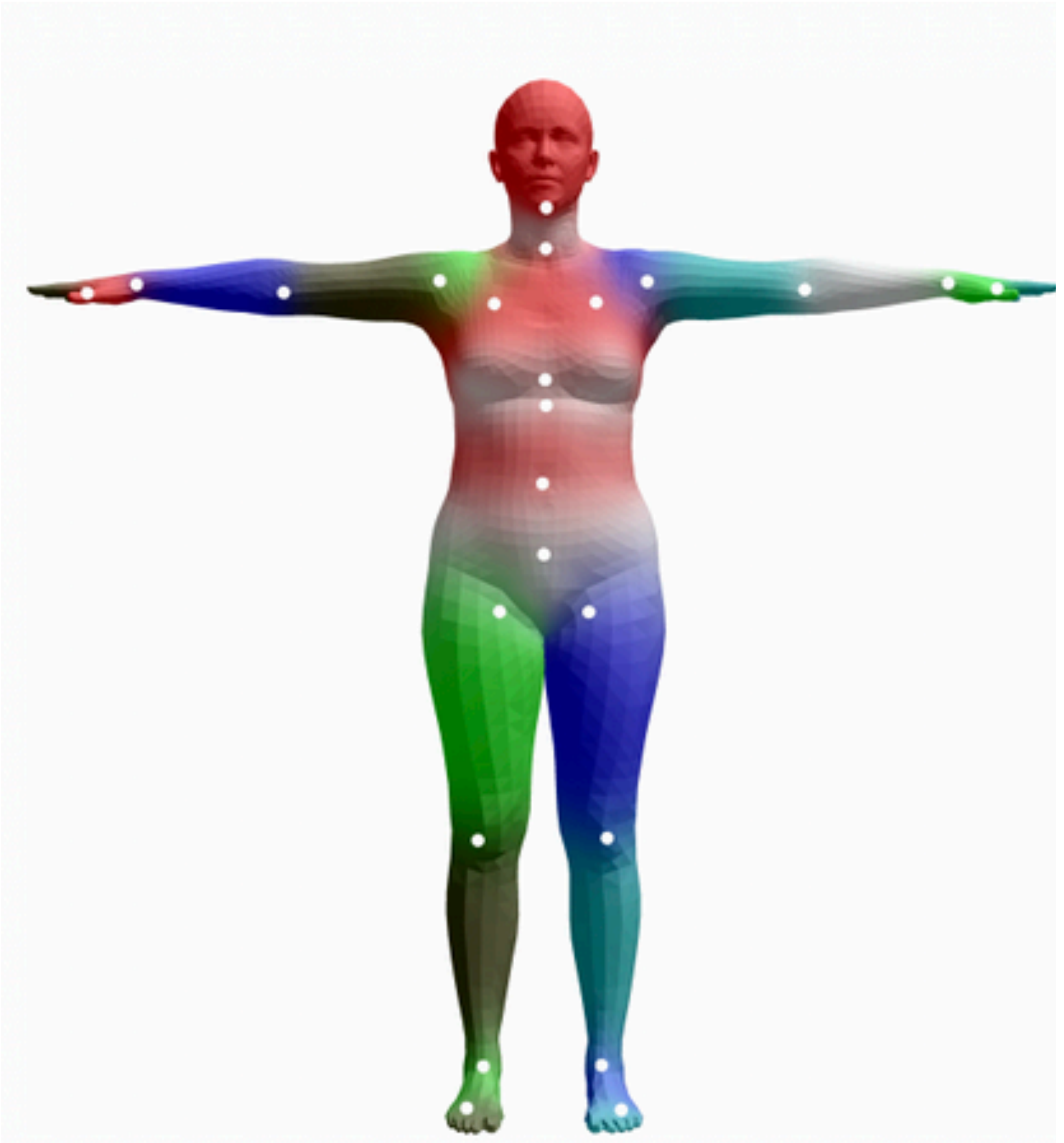


3DGS: rasterization

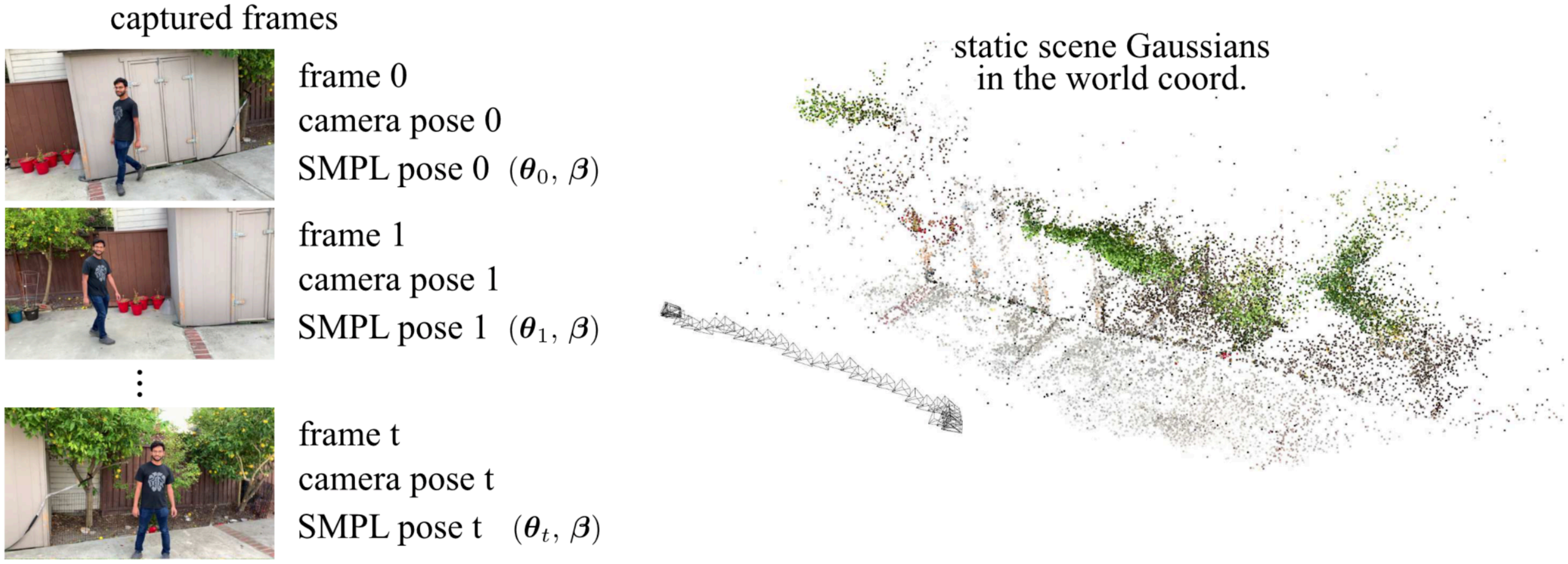


Preliminary: SMPL body model

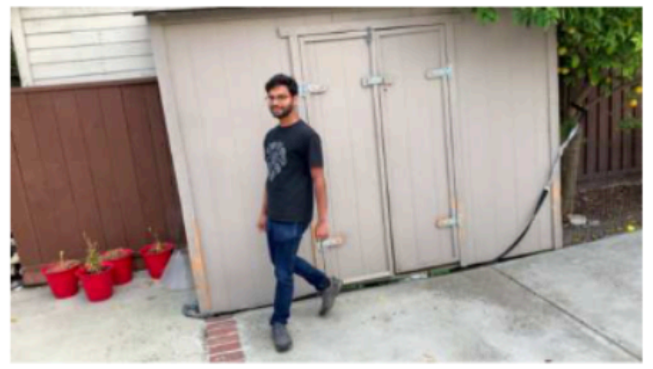
- Pose: skeleton configuration
- Shape: body shape variations (height, weight etc.)
- Canonical body mesh: T-pose
- Posed mesh: Linear Blend Skinning (LBS)



Method overview


captured frames
frame 0
camera pose 0
SMPL pose 0 $(\boldsymbol{\theta}_0, \boldsymbol{\beta})$
frame 1
camera pose 1
SMPL pose 1 $(\boldsymbol{\theta}_1, \boldsymbol{\beta})$
⋮
frame t
camera pose t
SMPL pose t $(\boldsymbol{\theta}_t, \boldsymbol{\beta})$
static scene Gaussians
in the world coord.

captured frames



frame 0

camera pose 0

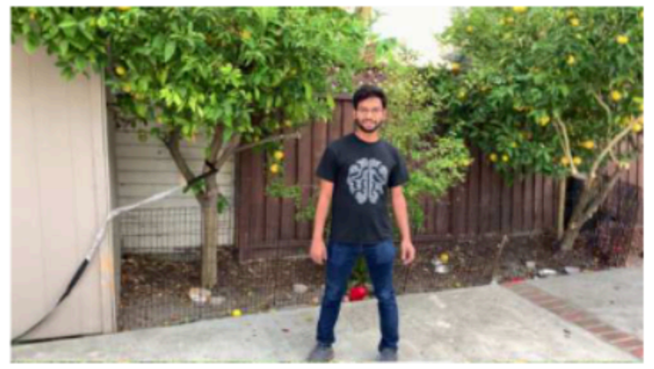
SMPL pose 0 $(\boldsymbol{\theta}_0, \boldsymbol{\beta})$

frame 1

camera pose 1

SMPL pose 1 $(\boldsymbol{\theta}_1, \boldsymbol{\beta})$

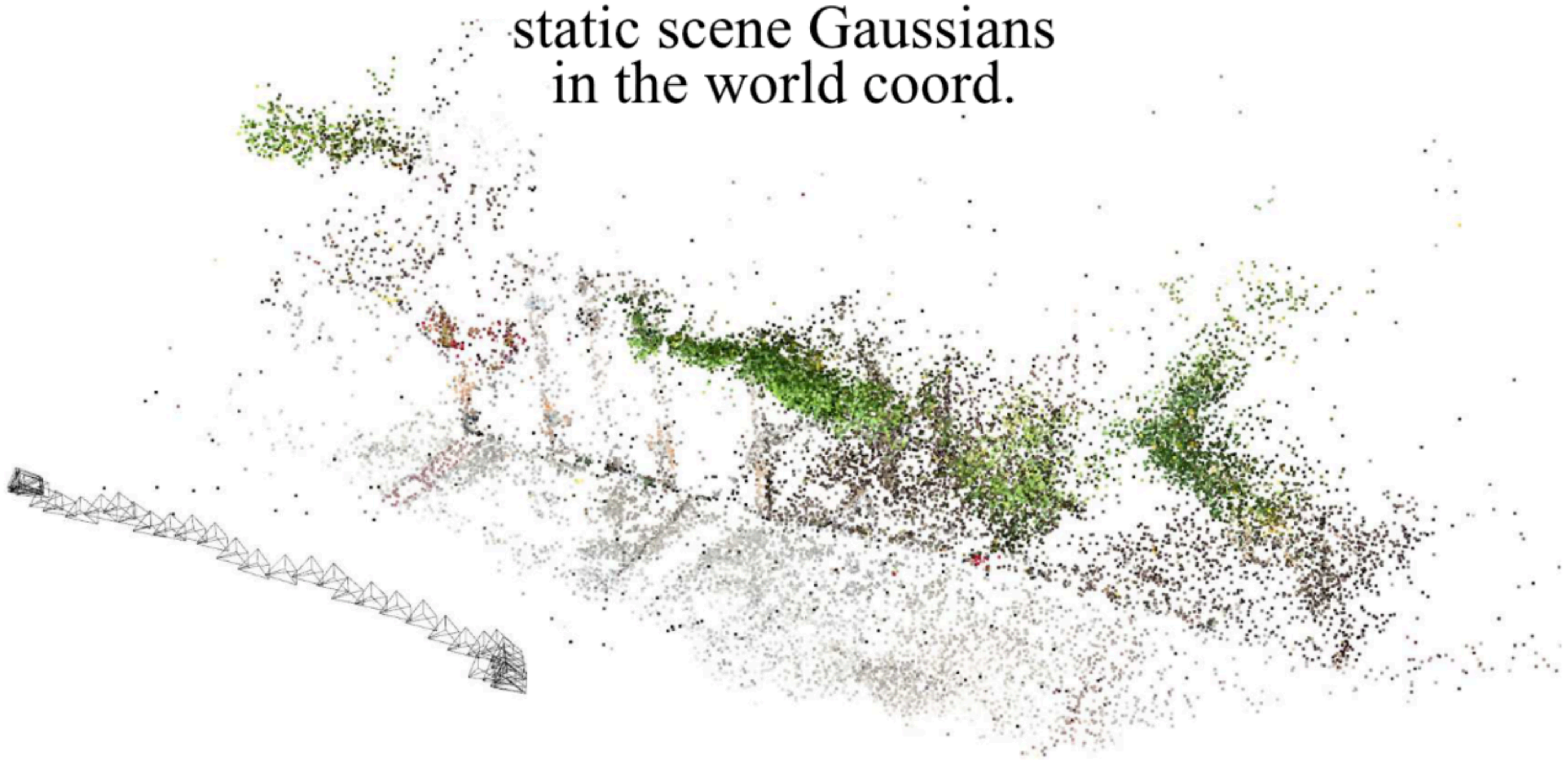
⋮



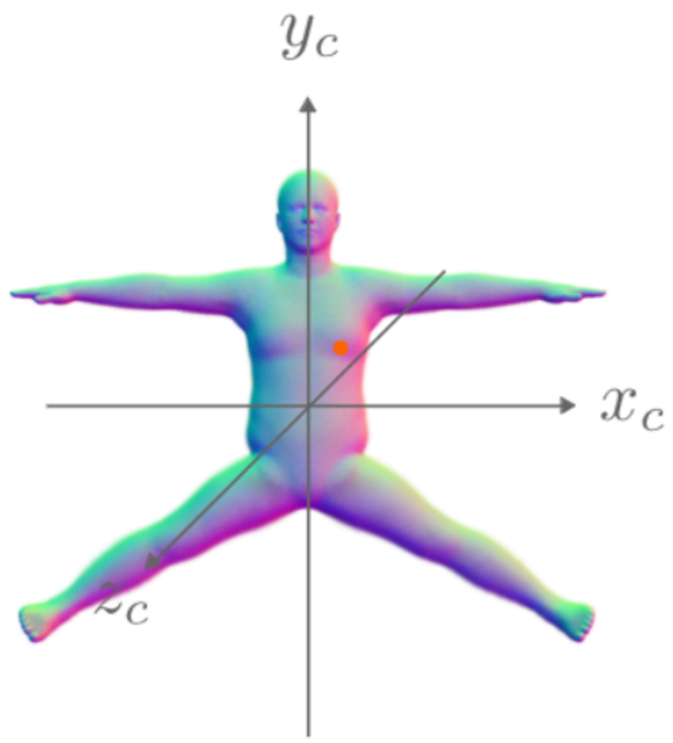
frame t

camera pose t

SMPL pose t $(\boldsymbol{\theta}_t, \boldsymbol{\beta})$



static scene Gaussians
in the world coord.



canonical space

captured frames

frame 0

camera pose 0

SMPL pose 0 $(\boldsymbol{\theta}_0, \boldsymbol{\beta})$

frame 1

camera pose 1

SMPL pose 1 $(\boldsymbol{\theta}_1, \boldsymbol{\beta})$

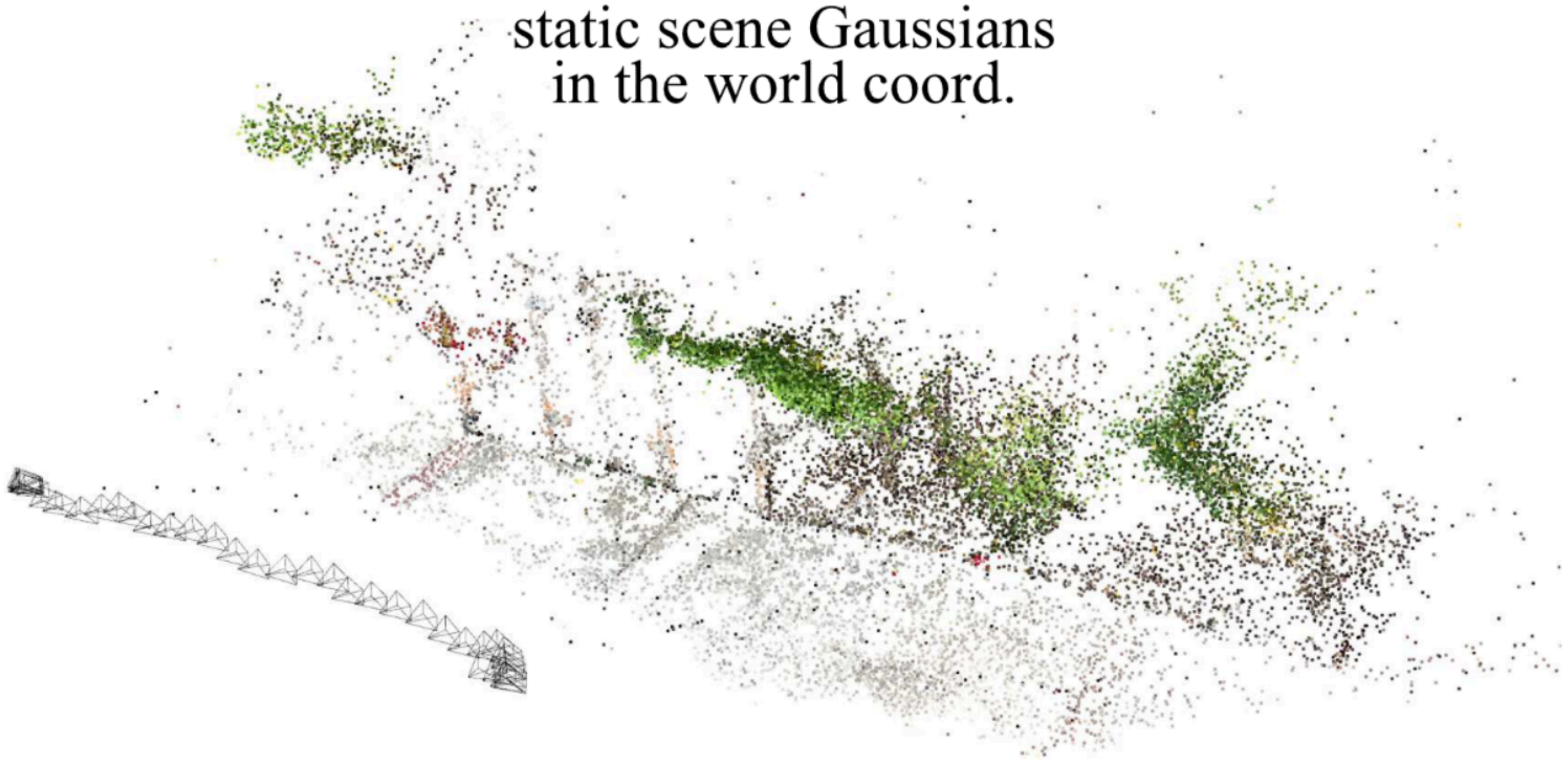
⋮

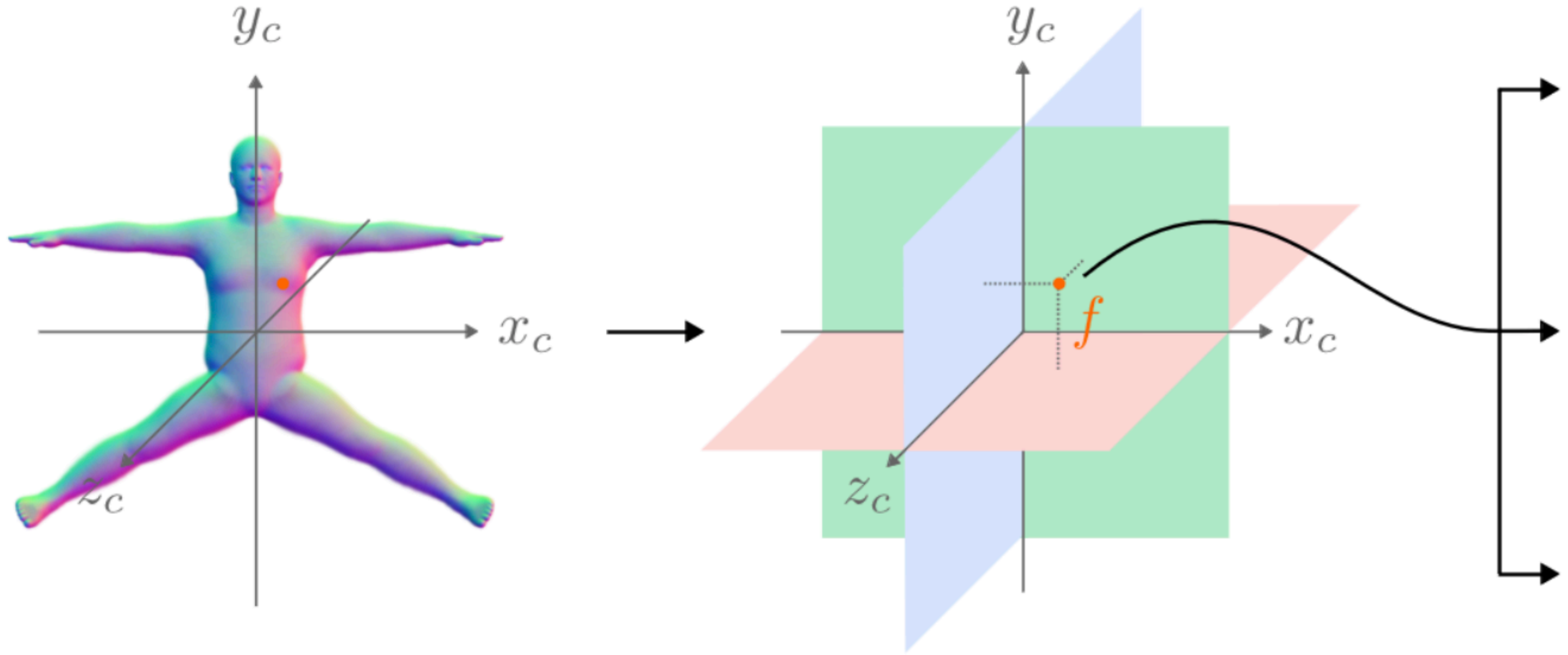
frame t

camera pose t

SMPL pose t $(\boldsymbol{\theta}_t, \boldsymbol{\beta})$

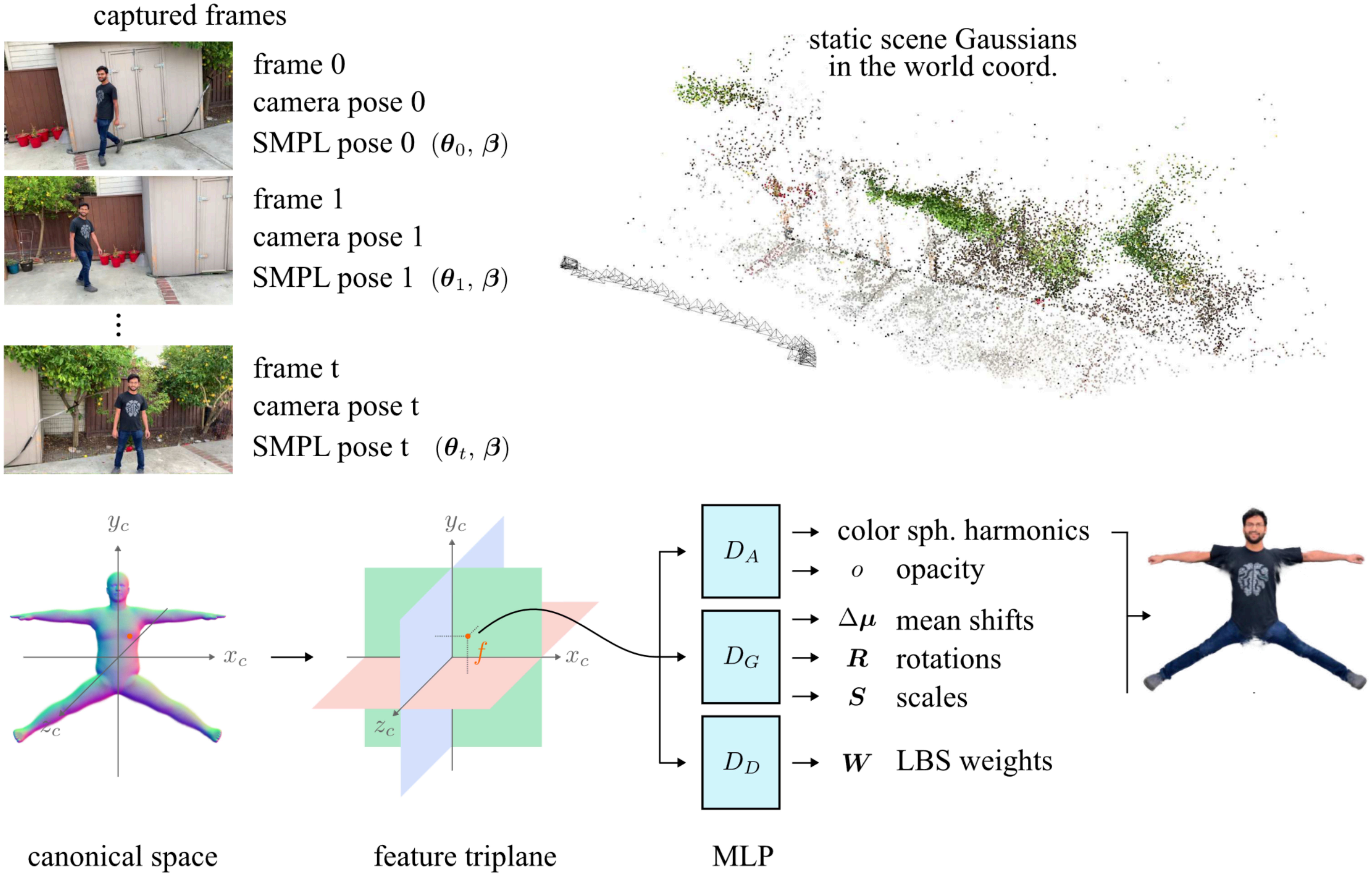
static scene Gaussians
in the world coord.





canonical space

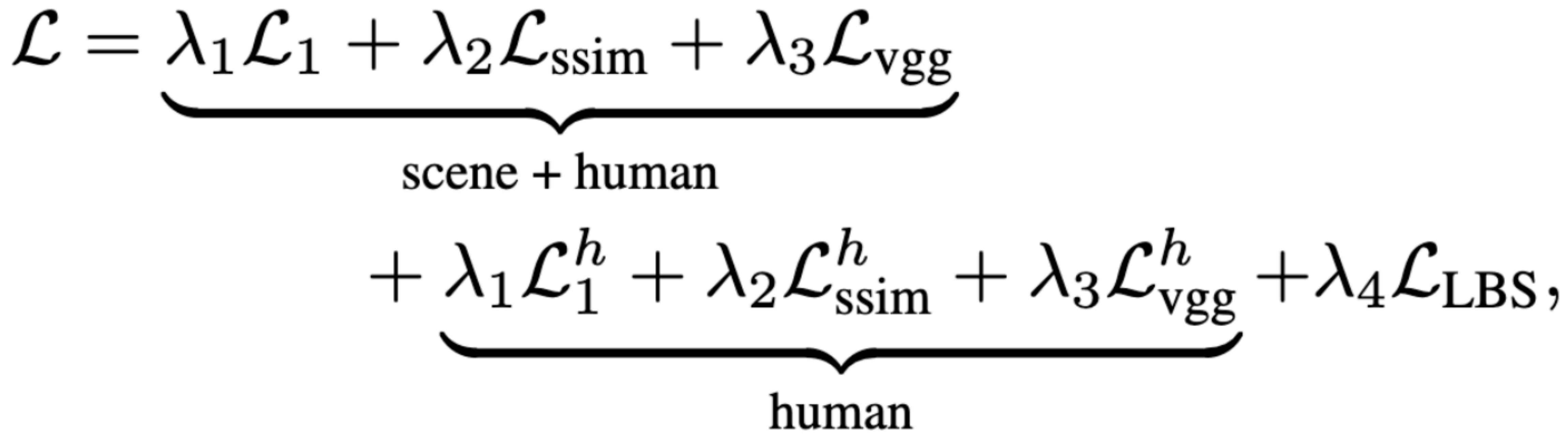
feature triplane


captured frames
frame 0
camera pose 0
SMPL pose 0 $(\boldsymbol{\theta}_0, \boldsymbol{\beta})$
frame 1
camera pose 1
SMPL pose 1 $(\boldsymbol{\theta}_1, \boldsymbol{\beta})$
⋮
frame t
camera pose t
SMPL pose t $(\boldsymbol{\theta}_t, \boldsymbol{\beta})$
static scene Gaussians
in the world coord.
y_c
x_c
z_c
canonical space
y_c
x_c
z_c
f
feature triplane
D_A
D_G
D_D
MLP
color sph. harmonics
o opacity
$\Delta\boldsymbol{\mu}$ mean shifts
$\boldsymbol{R}$ rotations
$\boldsymbol{S}$ scales
$\boldsymbol{W}$ LBS weights

captured frames
frame 0
camera pose 0
SMPL pose 0 $(\boldsymbol{\theta}_0, \boldsymbol{\beta})$
frame 1
camera pose 1
SMPL pose 1 $(\boldsymbol{\theta}_1, \boldsymbol{\beta})$
⋮
frame t
camera pose t
SMPL pose t $(\boldsymbol{\theta}_t, \boldsymbol{\beta})$
static scene Gaussians
in the world coord.
camera pose t
splat
rendered frame t
y_c
x_c
z_c
y_c
f
x_c
z_c
D_A
D_G
D_D
color sph. harmonics
o opacity
$\Delta\boldsymbol{\mu}$ mean shifts
$\boldsymbol{R}$ rotations
$\boldsymbol{S}$ scales
$\boldsymbol{W}$ LBS weights
canonical space
feature triplane
MLP
human Gaussians
in the world coord.

Loss function

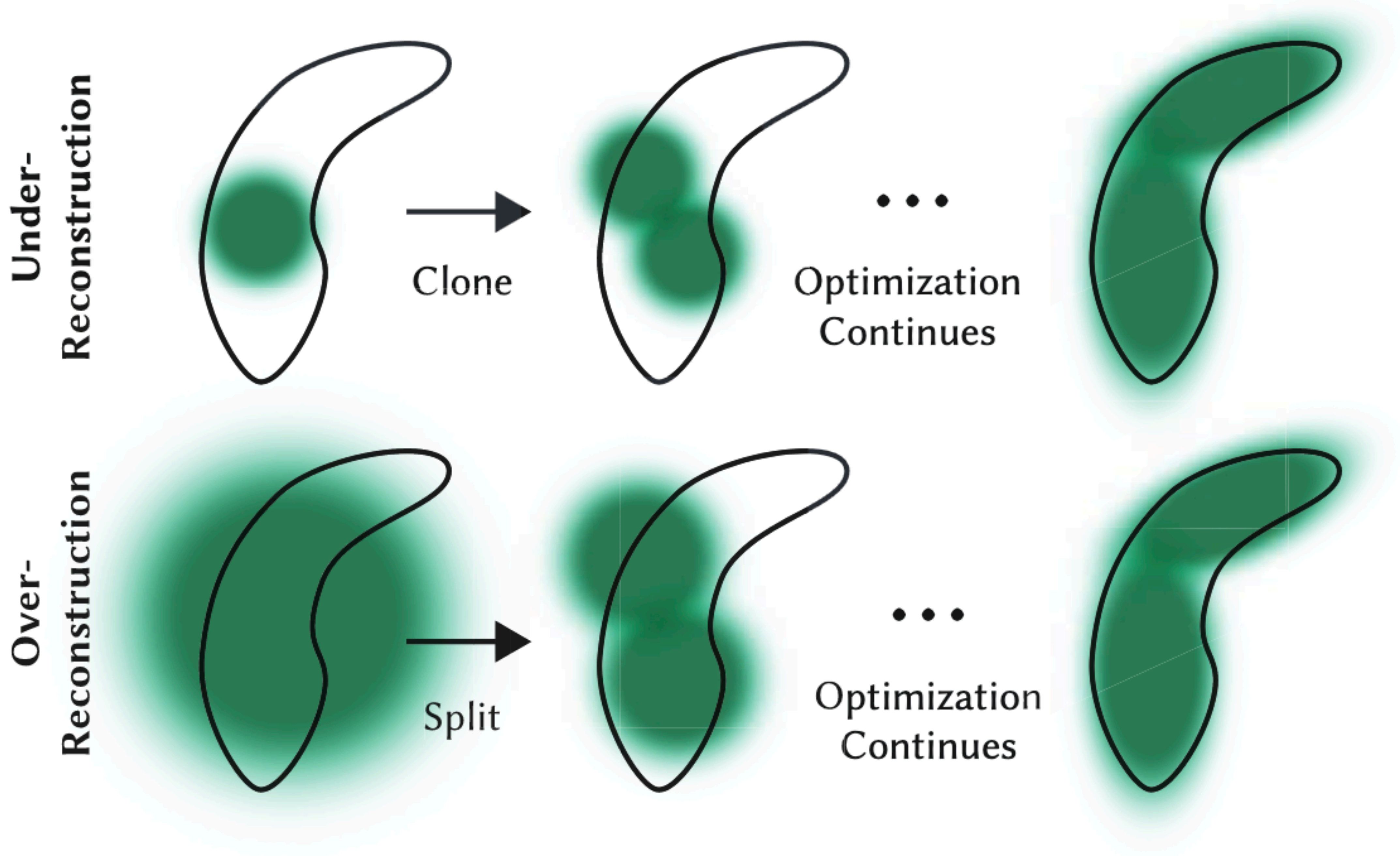
- $\mathscr{L}^h$: || human-only image - segmented GT image ||



$$\mathcal{L} = \underbrace{\lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{ssim}} + \lambda_3 \mathcal{L}_{\text{vgg}}}_{\text{scene + human}} + \underbrace{\lambda_1 \mathcal{L}_1^h + \lambda_2 \mathcal{L}_{\text{ssim}}^h + \lambda_3 \mathcal{L}_{\text{vgg}}^h}_{\text{human}} + \lambda_4 \mathcal{L}_{\text{LBS}},$$

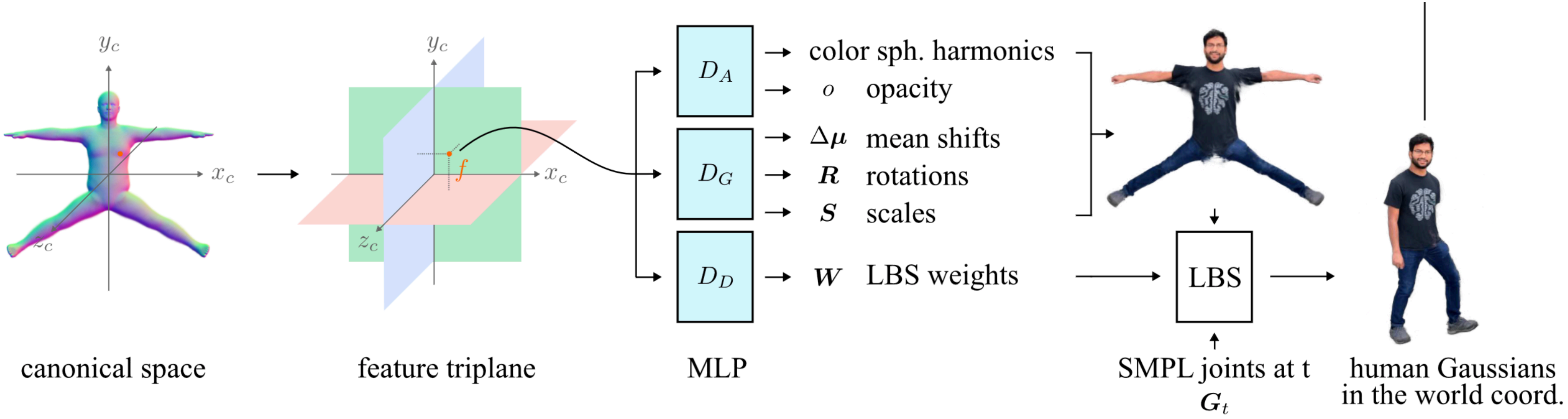
Adaptive Control of the number of Gaussians

- Clone, split, prune Gaussians based on screen-space positional gradients and opacity



Test time

- Rendering speed: 60 FPS.
- Evaluate Triplane+MLP for each subject once.
- Skinning and 3DGS rendering are the only operations



Results

Training video

Canonical avatar

Scene

Novel view and pose

Training video
2583
Canonical avatar
Scene
Novel view and pose

Training video

Canonical avatar

Scene

Novel view and pose

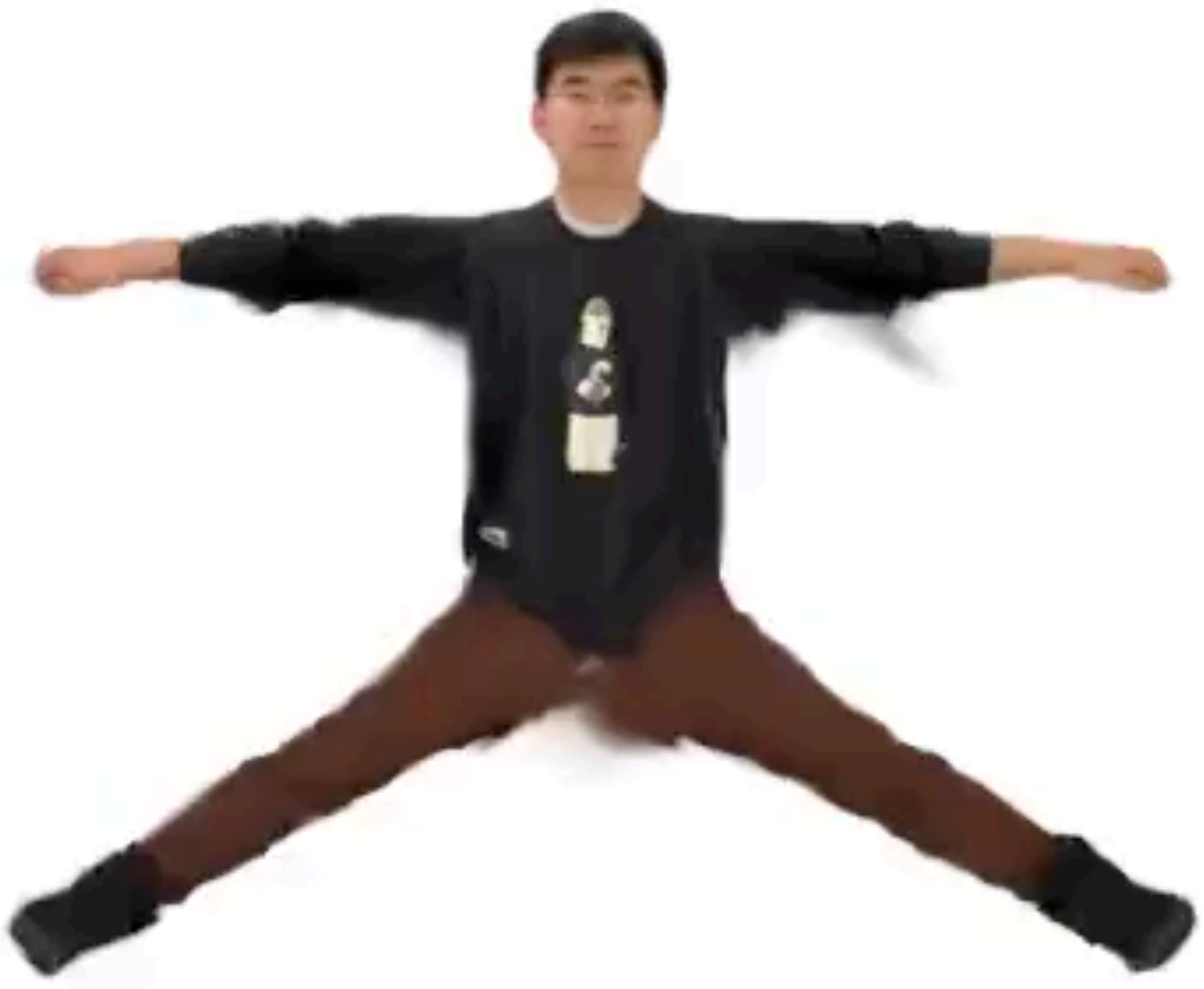
Training video
Canonical avatar
Scene
Novel view and pose

Novel scene + multiperson

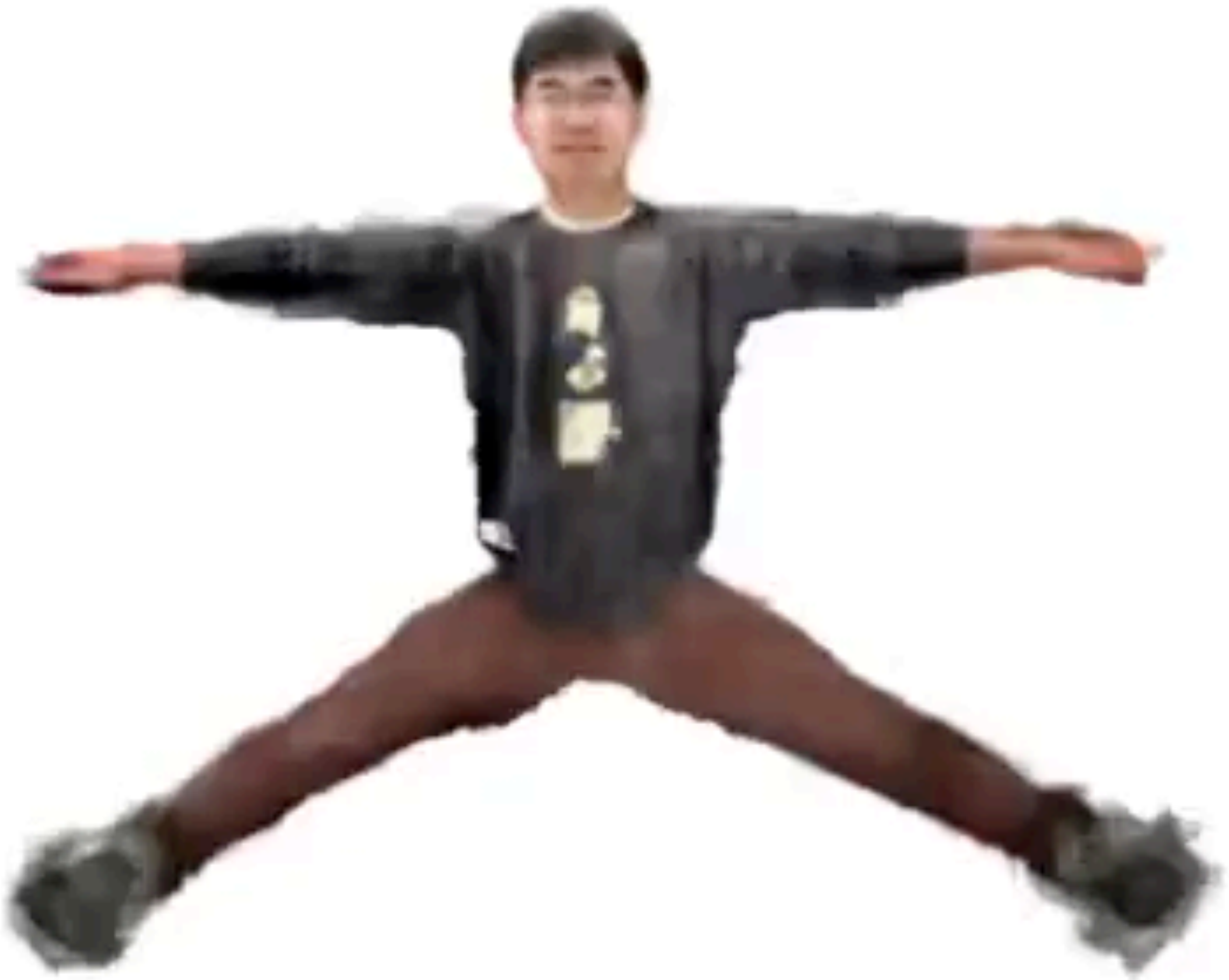
ÜRAL POR LEN

HUGS vs NeuMan

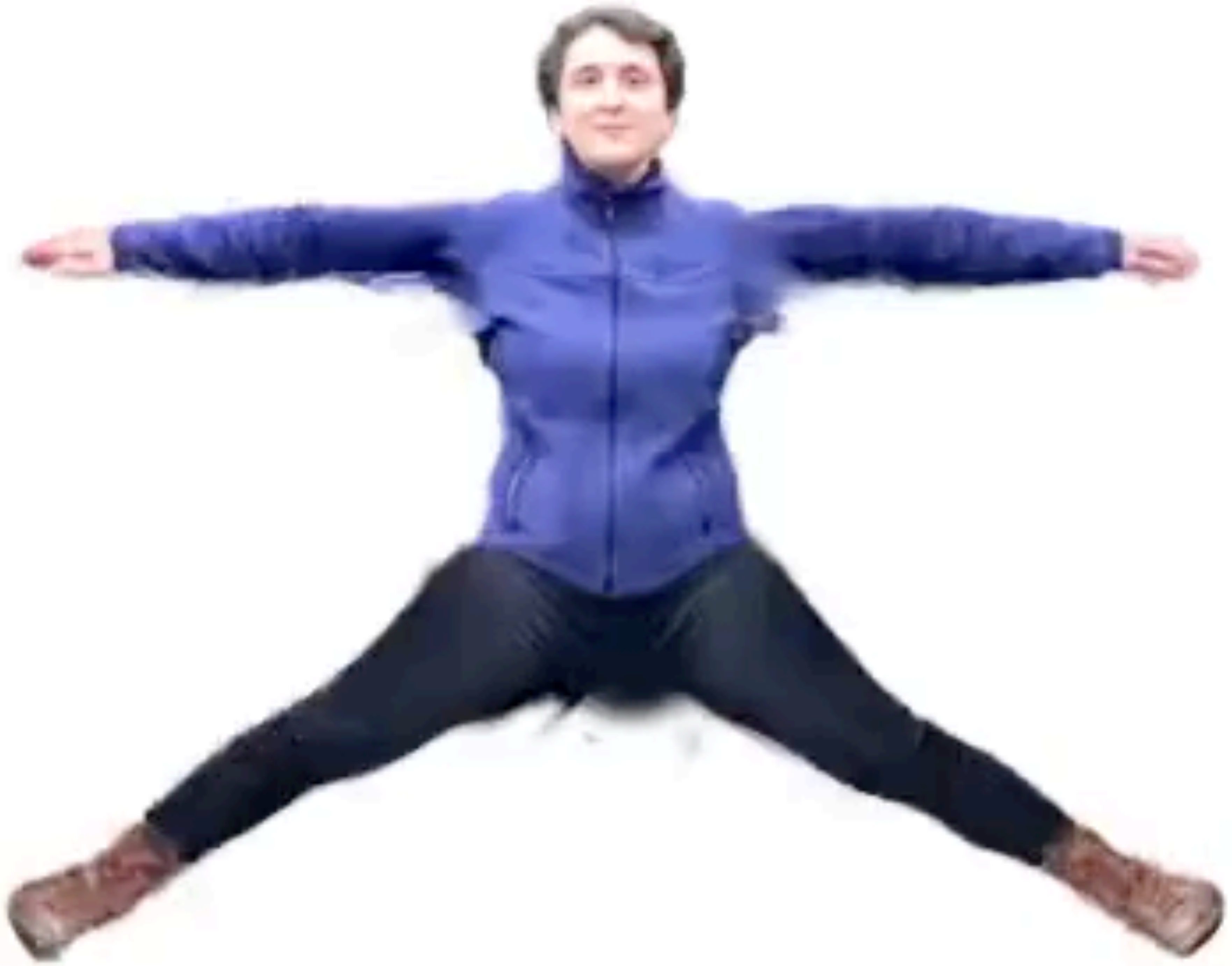
HUGS



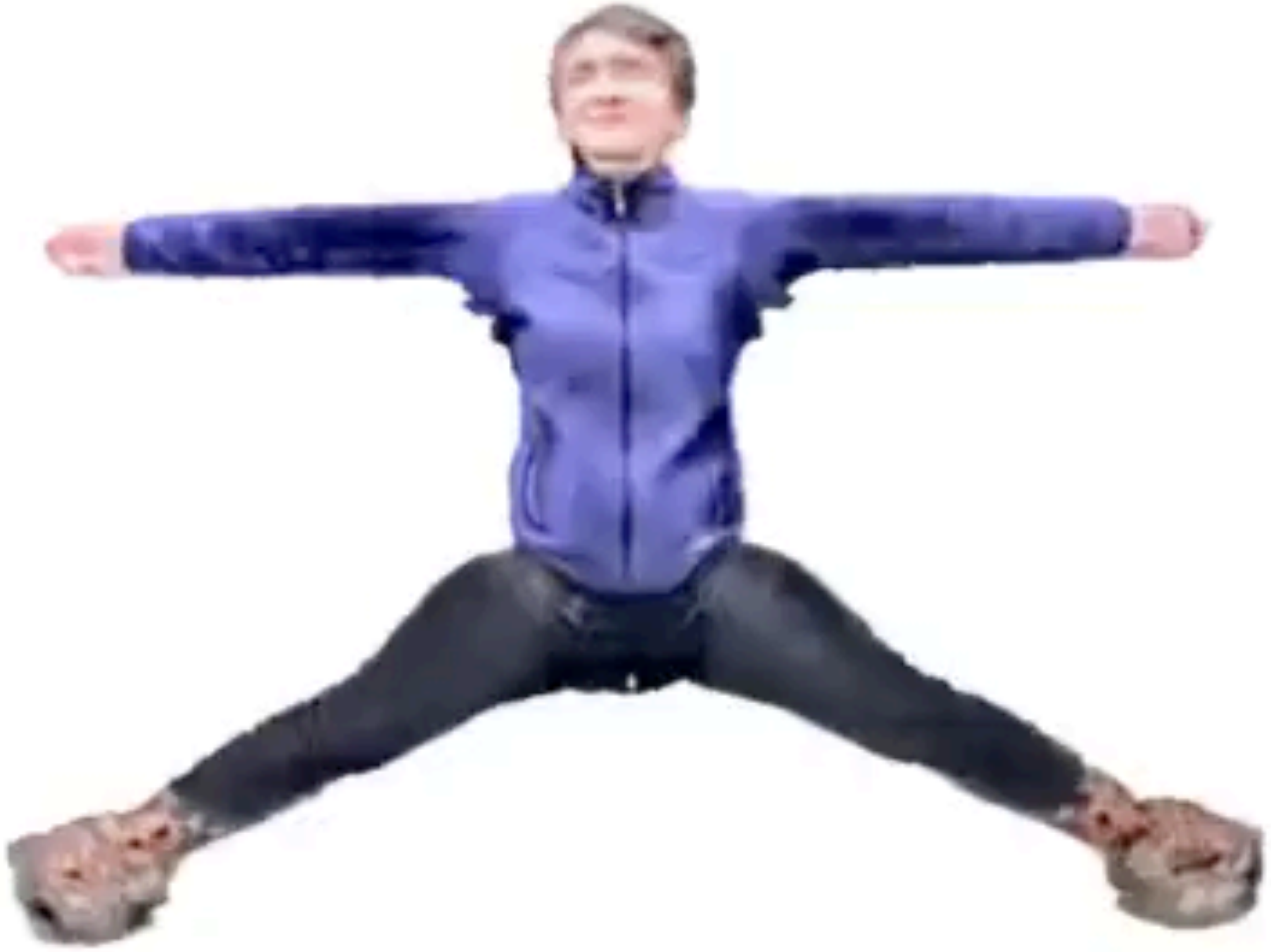
NeuMan



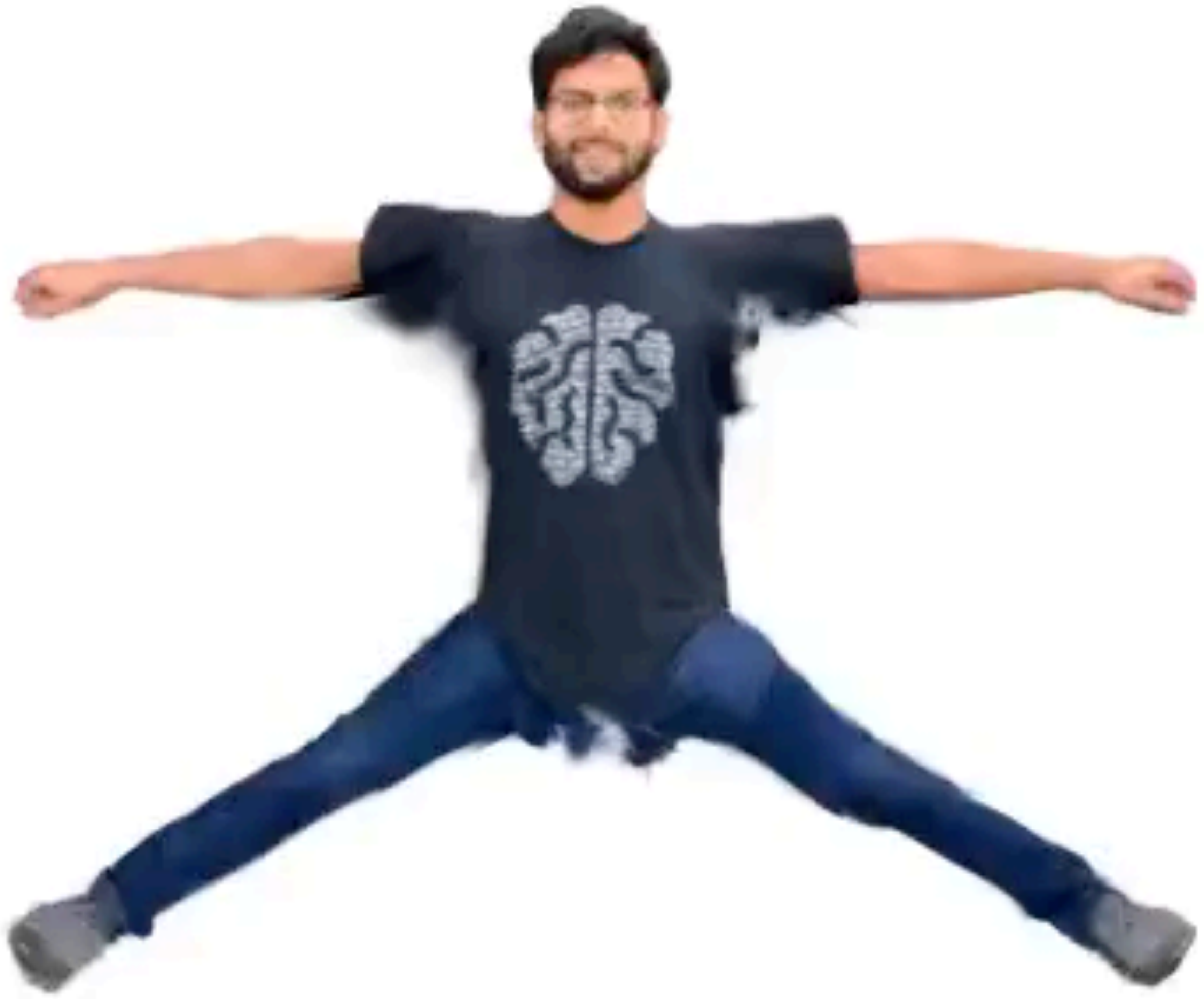
HUGS



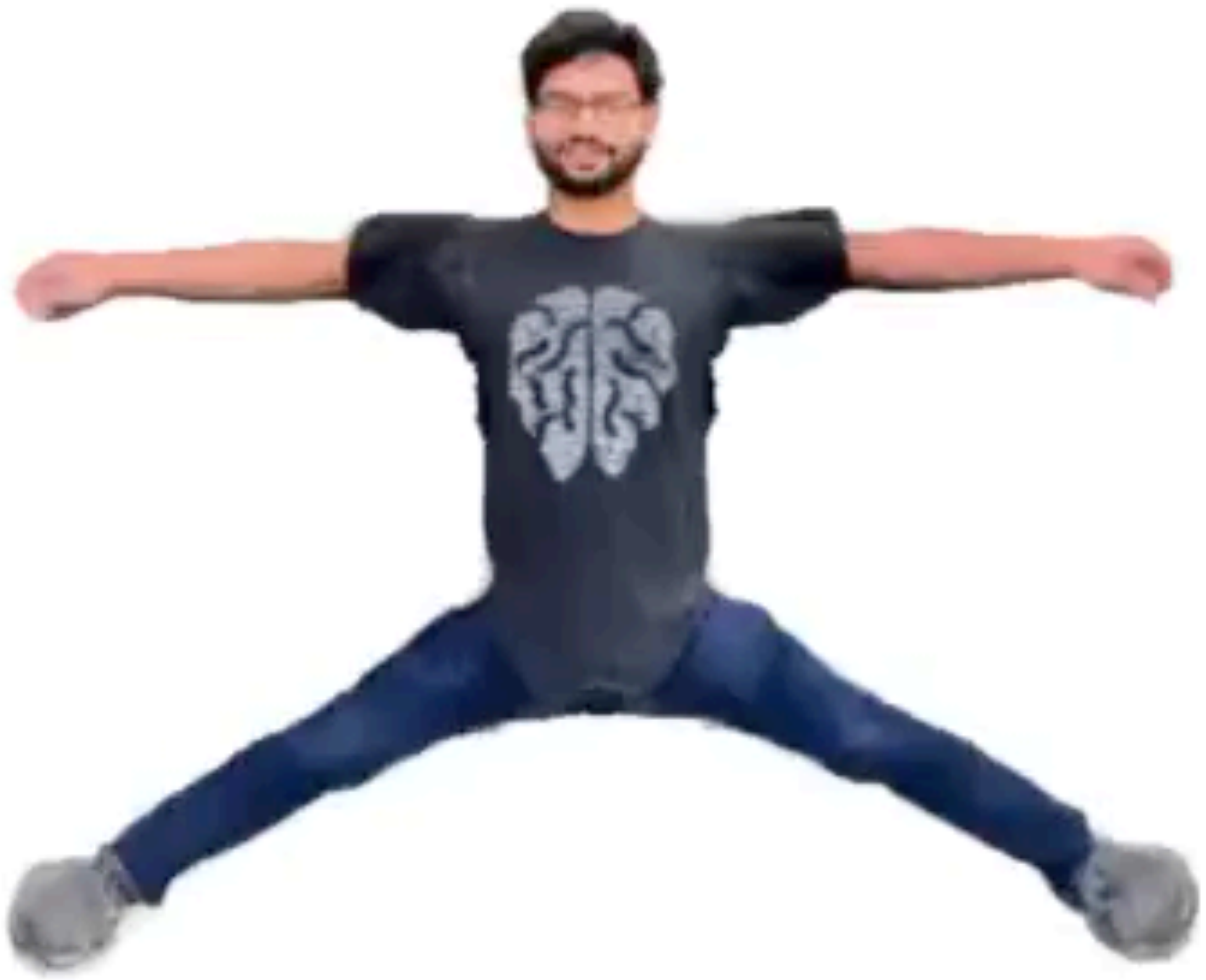
NeuMan



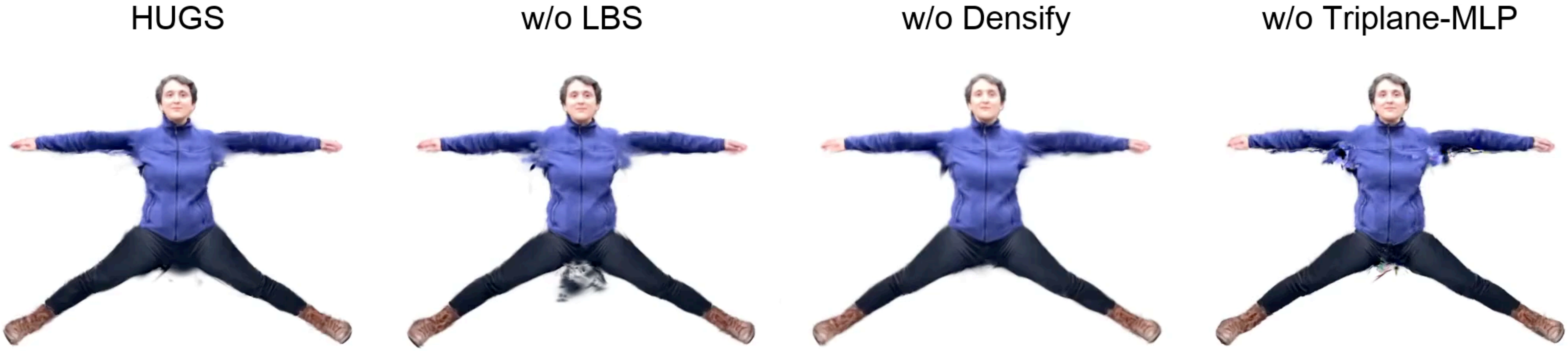
HUGS



NeuMan



Ablation experiments



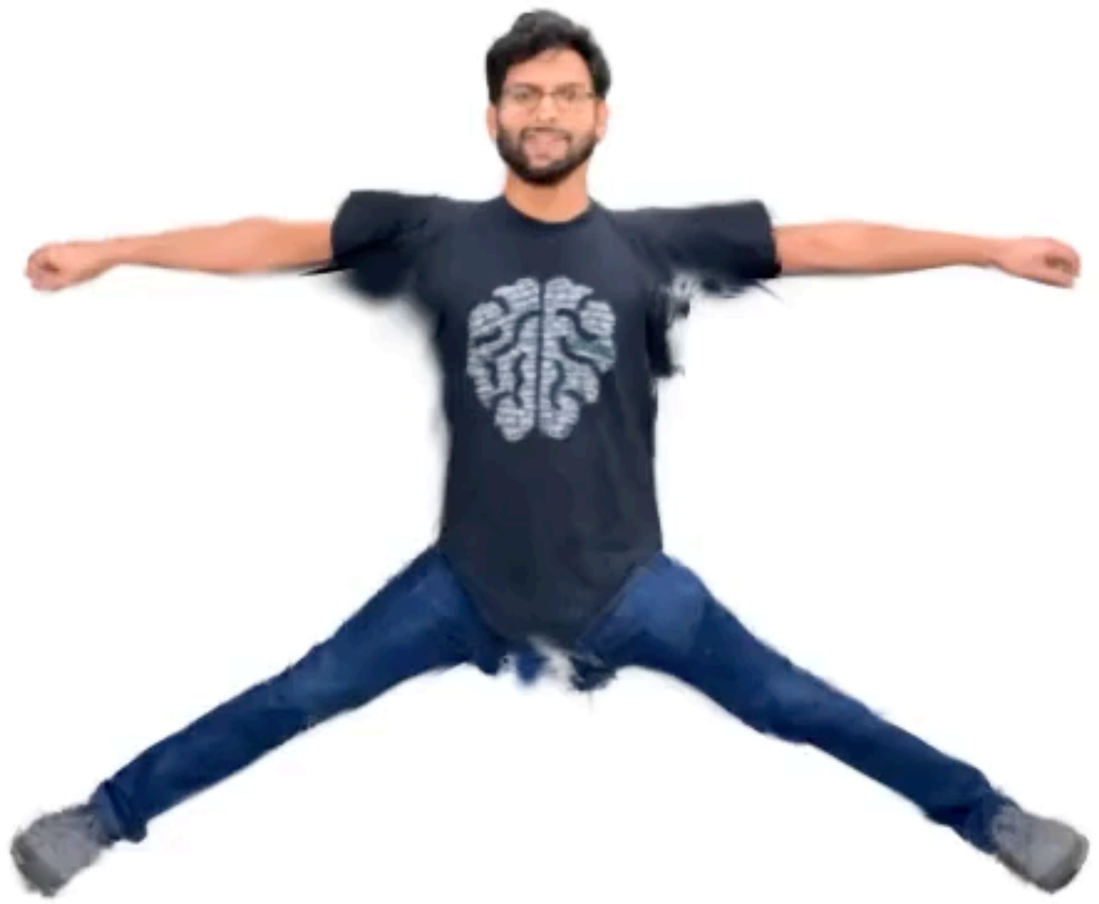
HUGS

w/o LBS

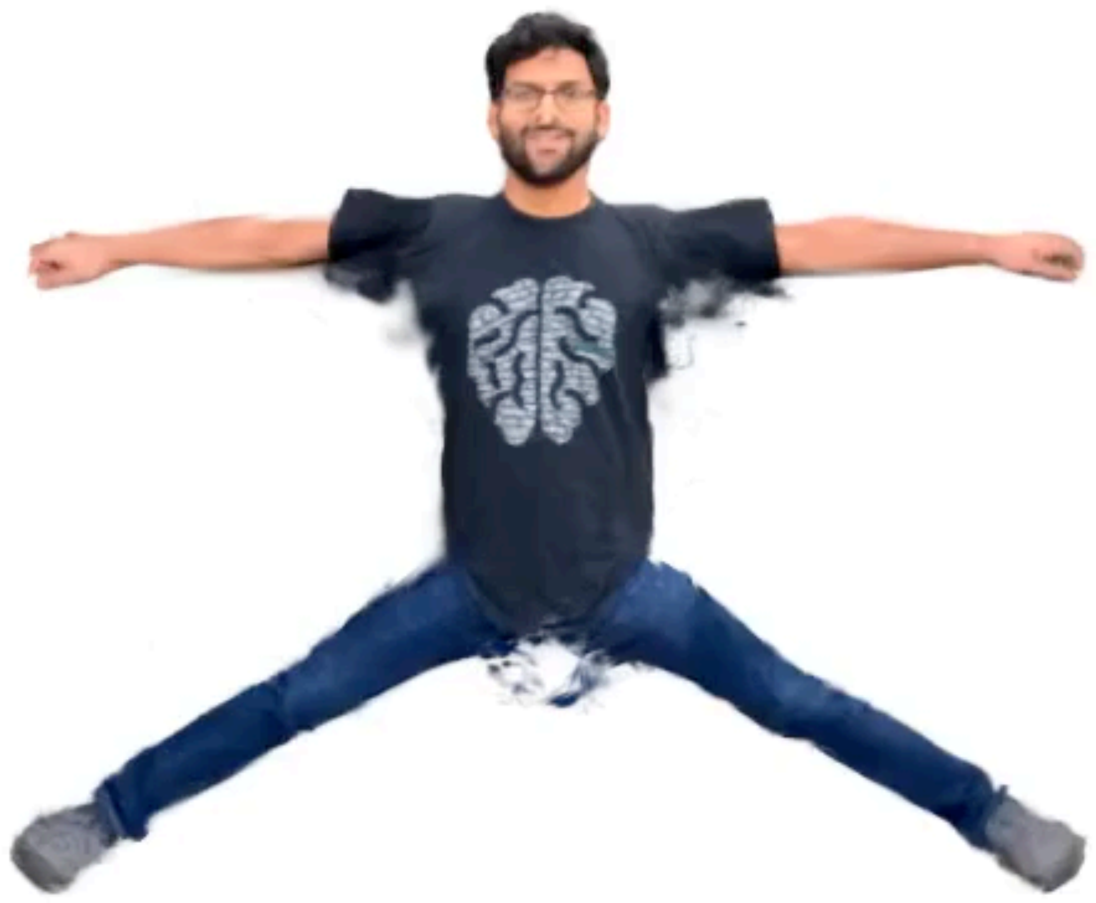
w/o Densify

w/o Triplane-MLP

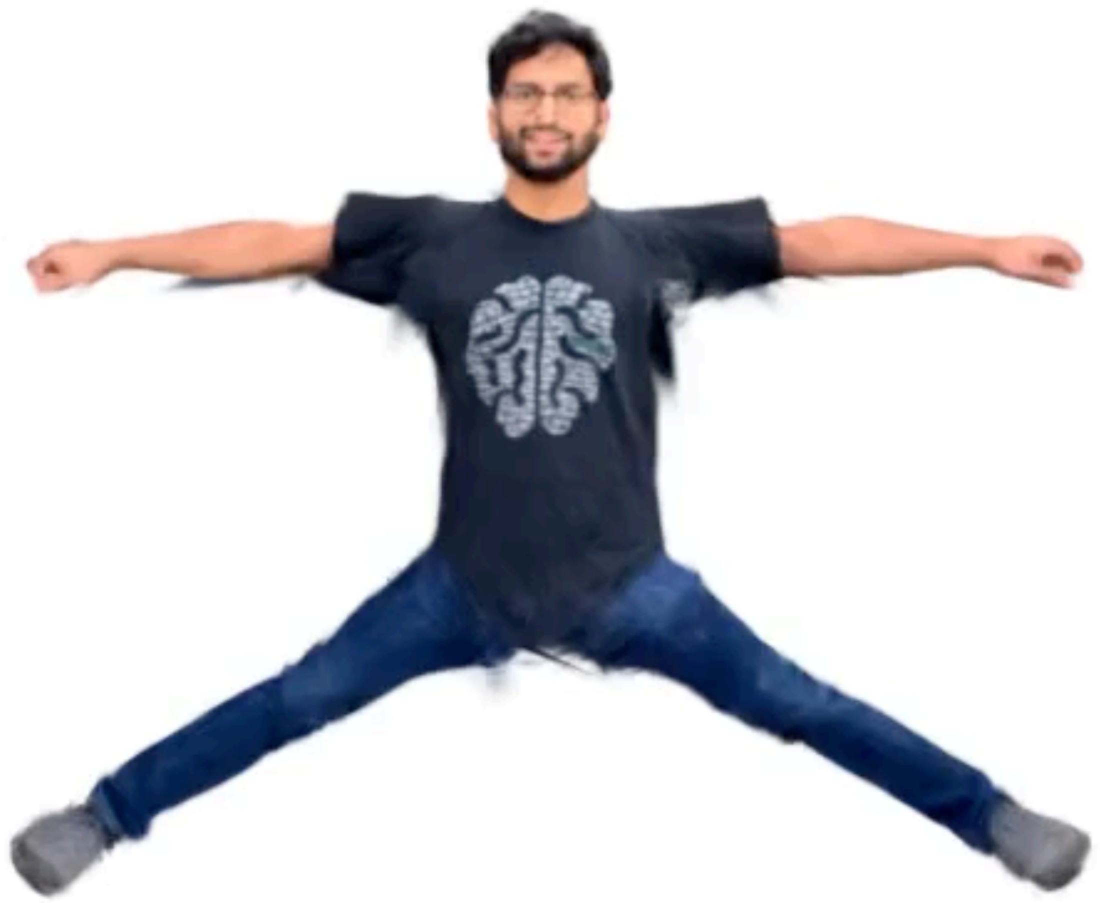
HUGS



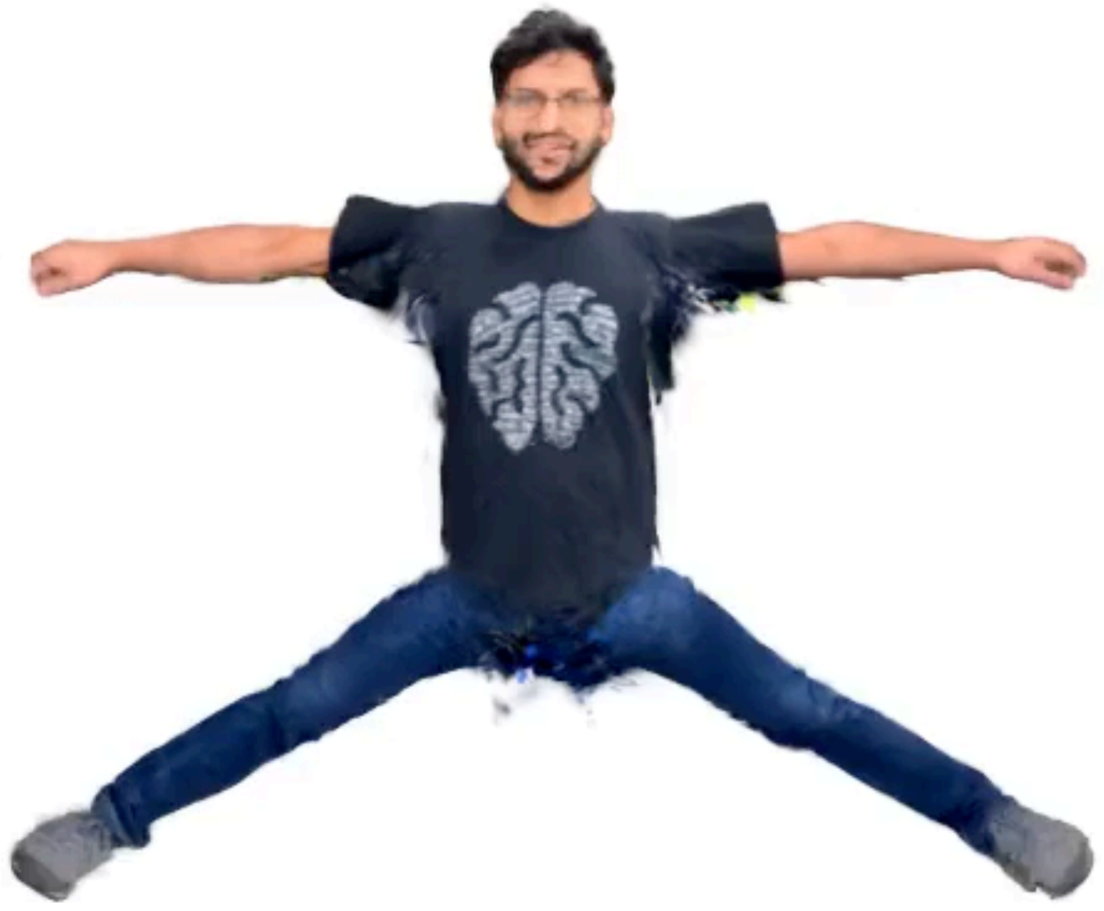
w/o LBS



w/o Densify



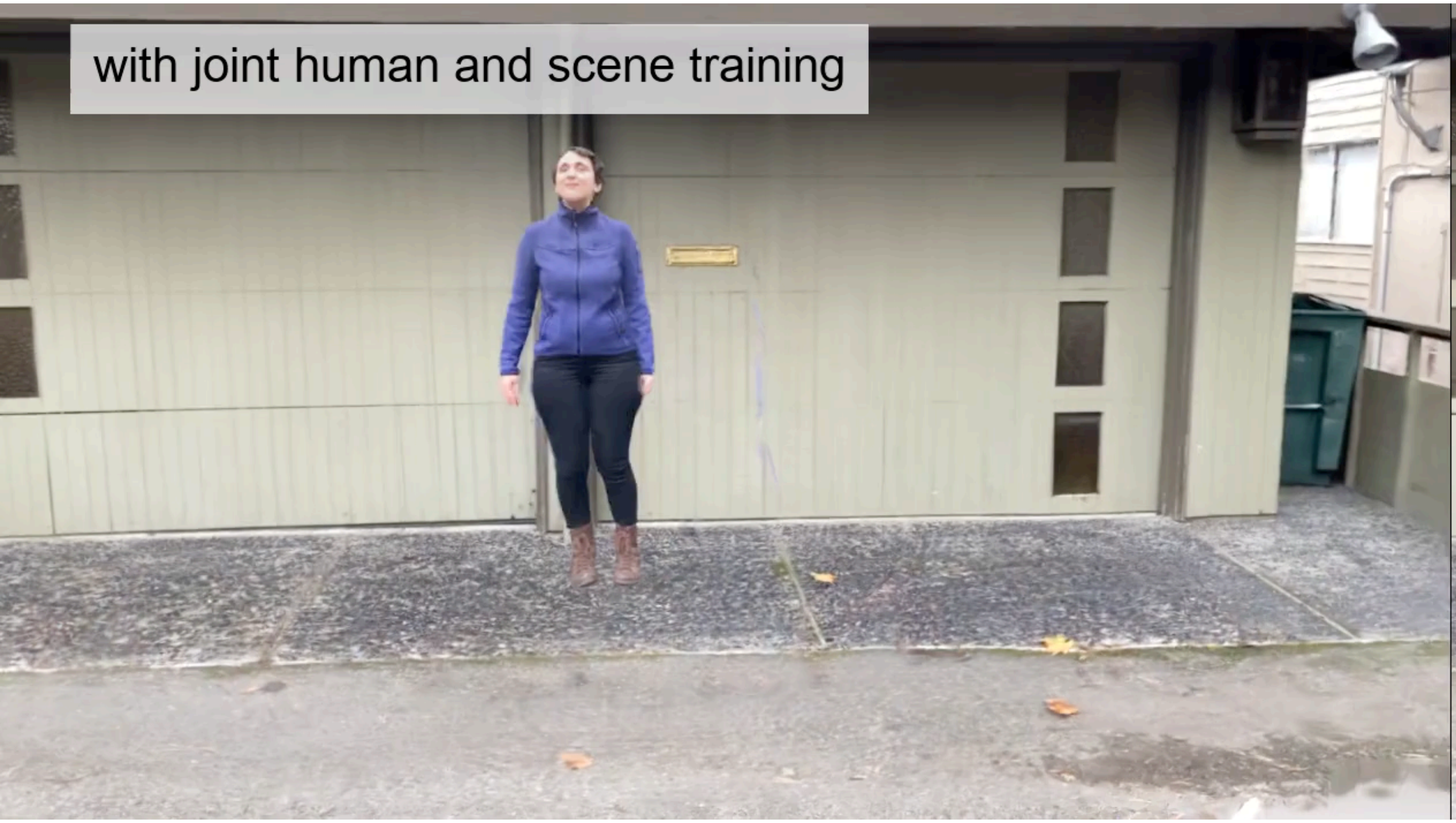
w/o Triplane-MLP



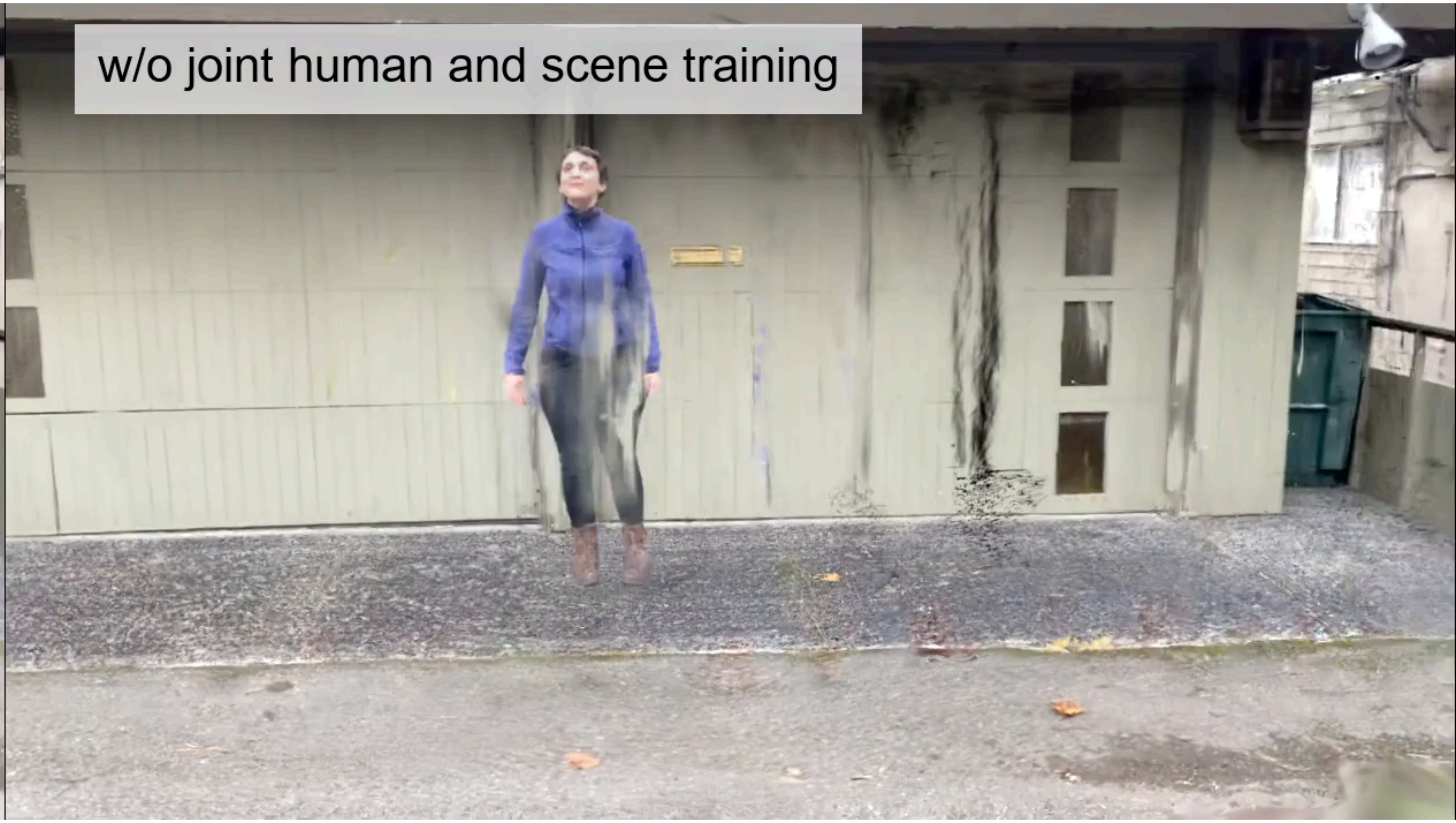
a. HUGS

b. Point features+MLP

c. HUGS--TM



with joint human and scene training



w/o joint human and scene training

Conclusion

HUGS: Human Gaussian Splats

Project page: https://machinelearning.apple.com/research/hugs

Code: https://github.com/apple/ml-hugs