Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang

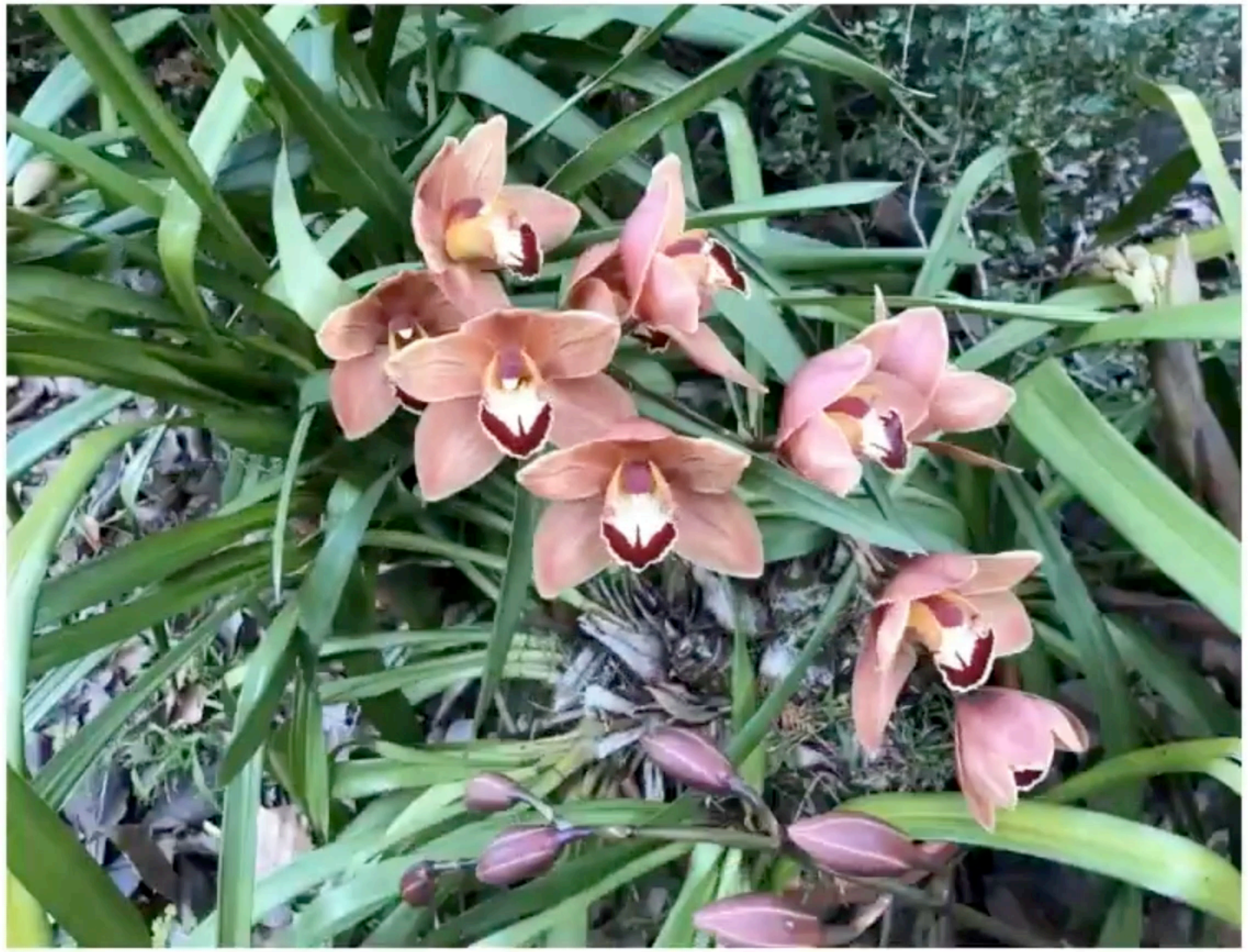
Qing Shuai, Hujun Bao, Xiaowei Zhou

浙江大學 ZHEJIANG UNIVERSITY

香港中文大學 The Chinese University of Hong Kong

Cornell University

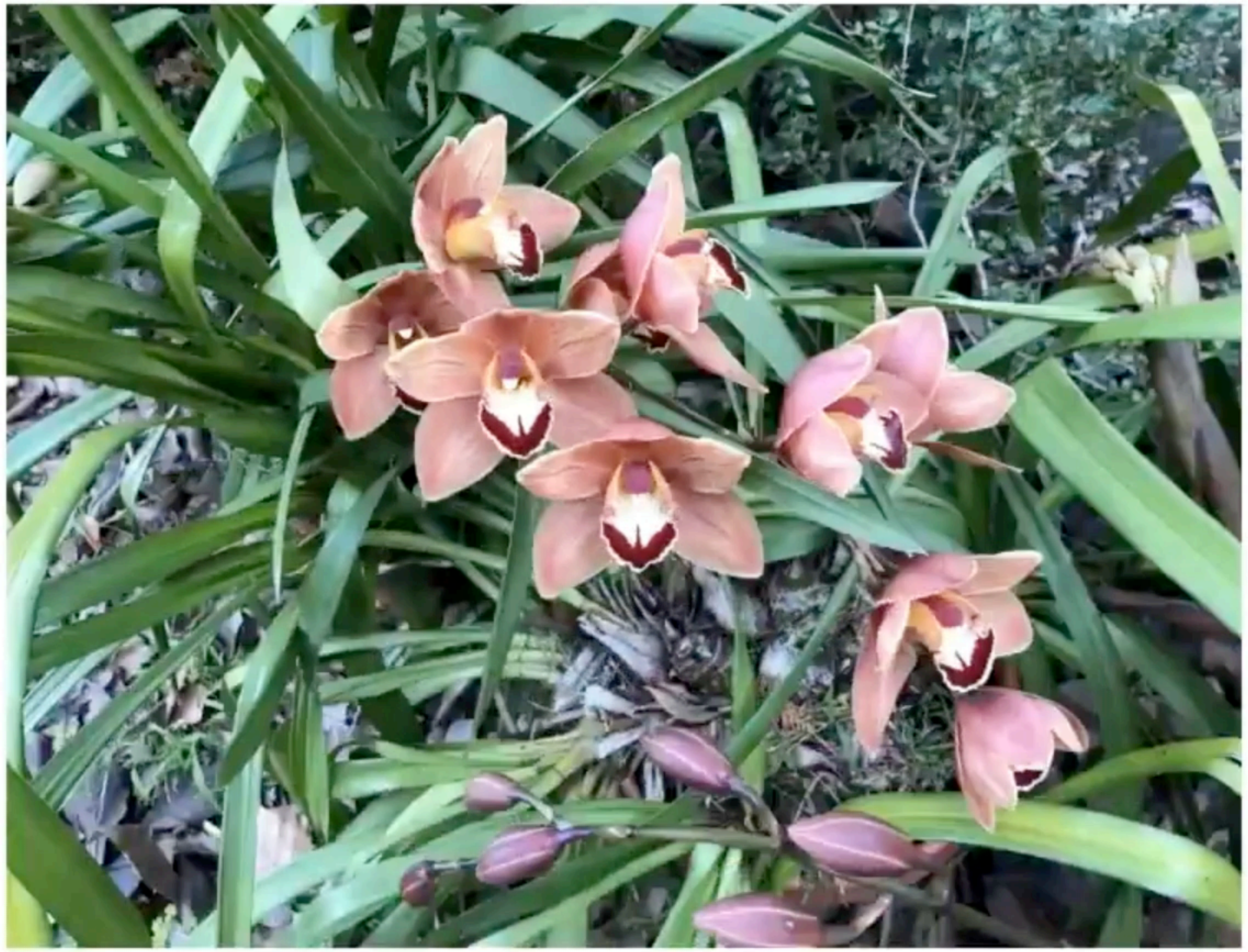
Problem statement: what is novel view synthesis



Input views

Mildenhall, Ben, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

Problem statement: what is novel view synthesis

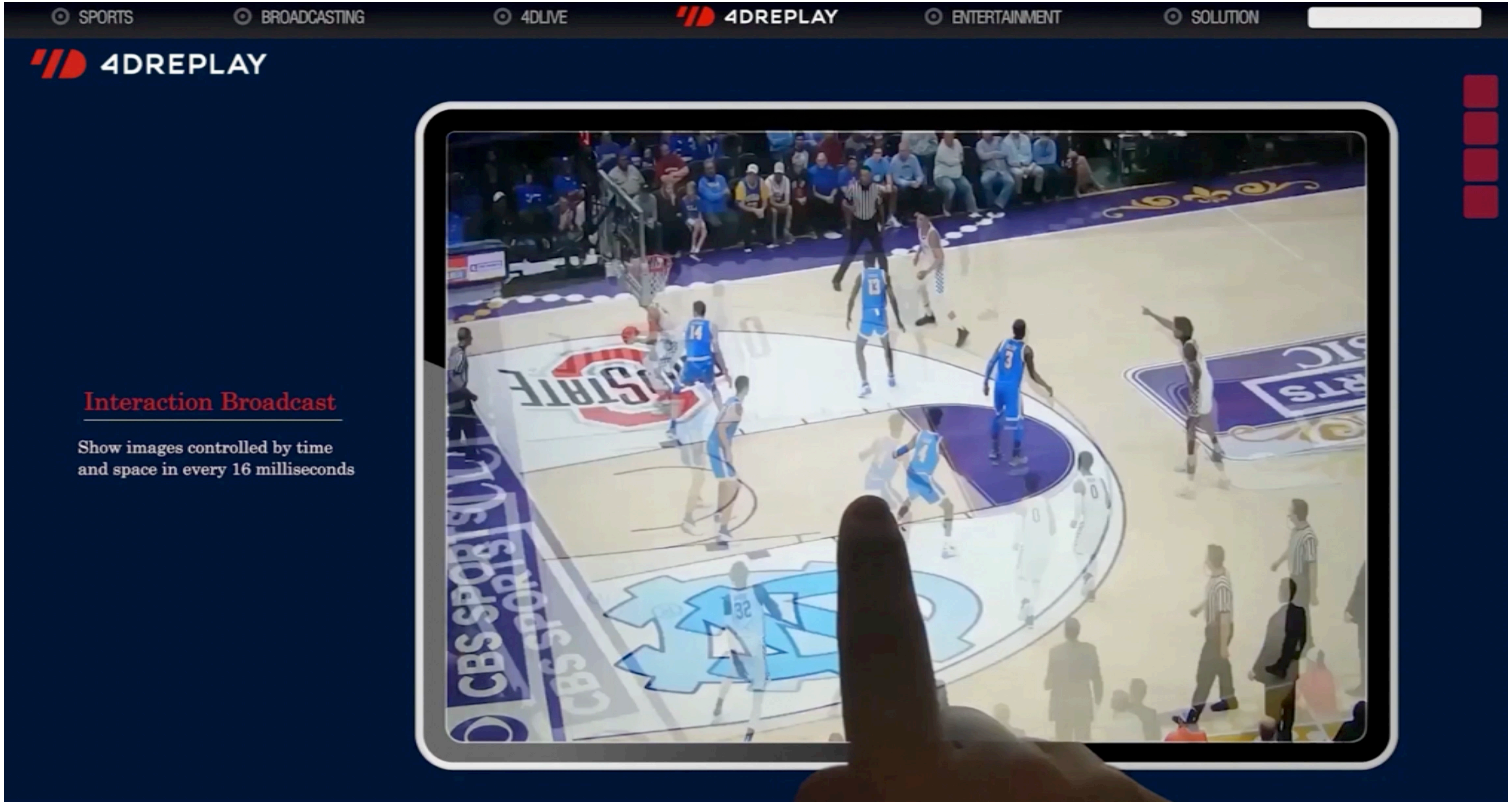


Input views

Novel view synthesis

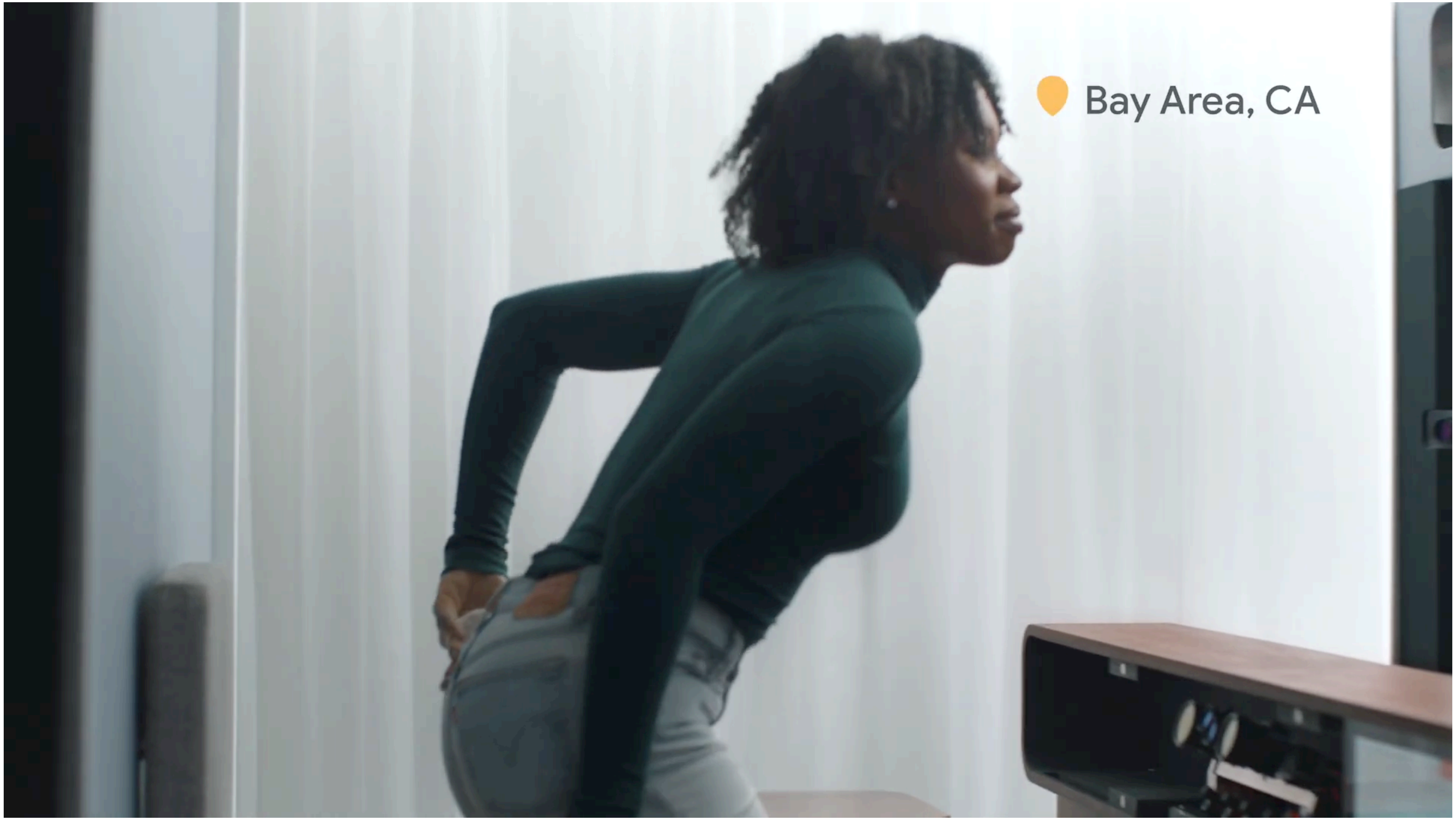
Mildenhall, Ben, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

Application： Sports broadcasting



4DREPLAY. https://www.4dreplay.com/

Application: Telepresence



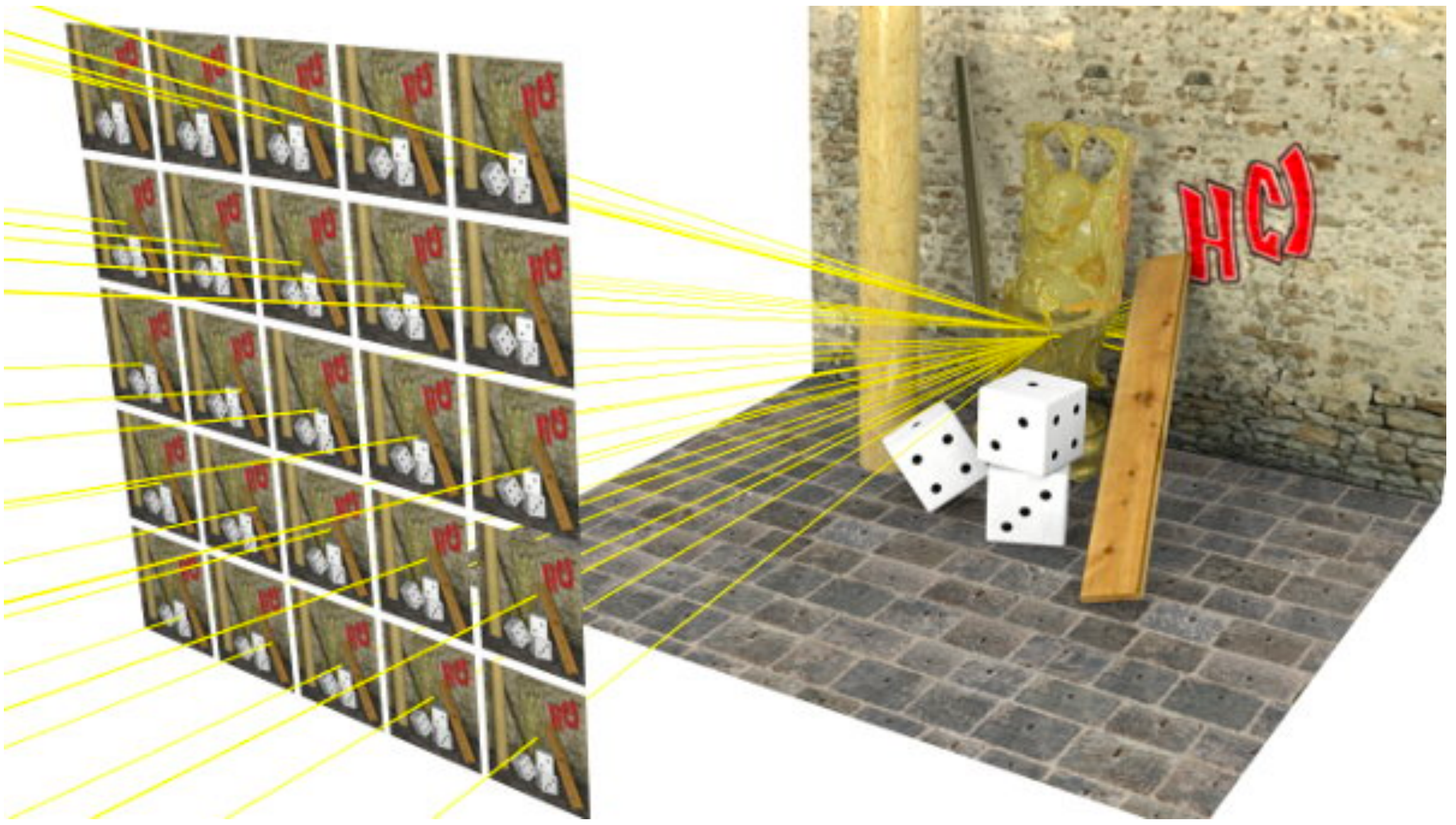
https://www.youtube.com/watch?v=Q13CishCKXY

Application：Telepresence

https://www.youtube.com/watch?v=Q13CishCKXY

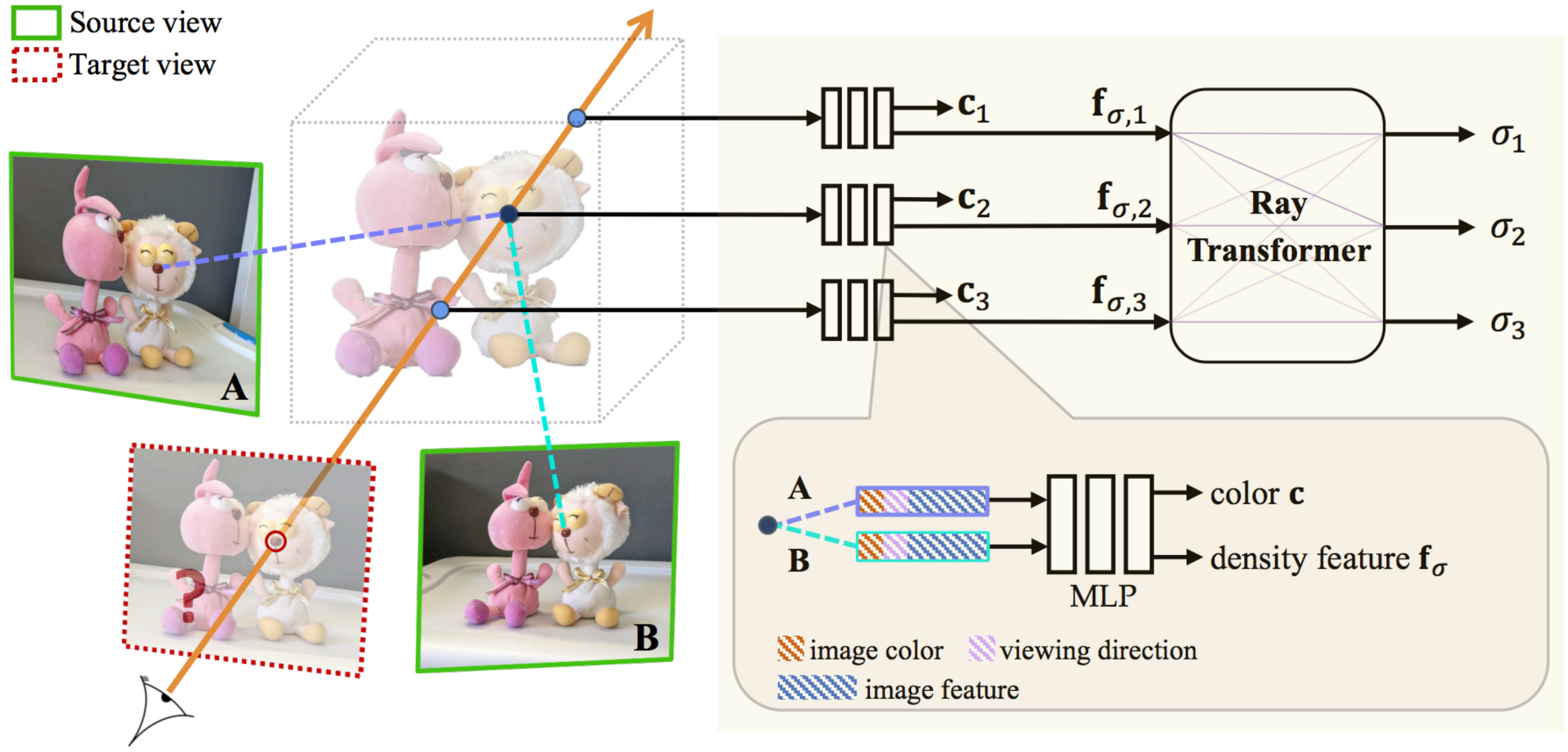
Related work

Light field interpolation



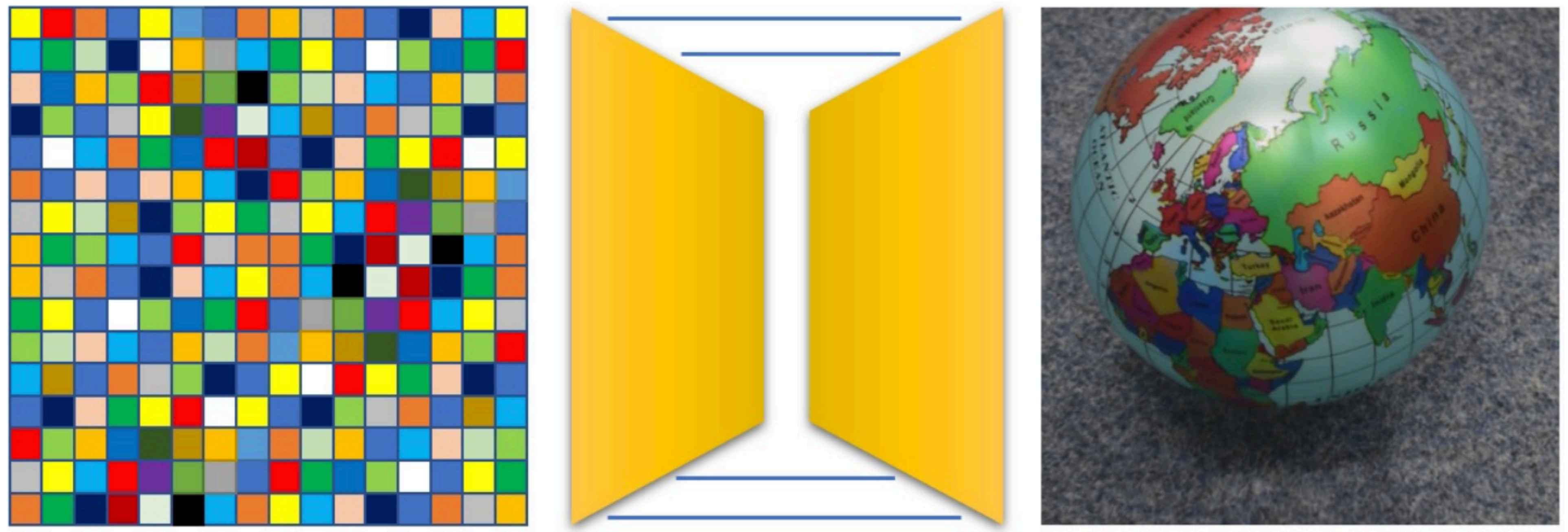
Gortler, Davis, Levoy, Hanrahan, et al.

Image-based rendering



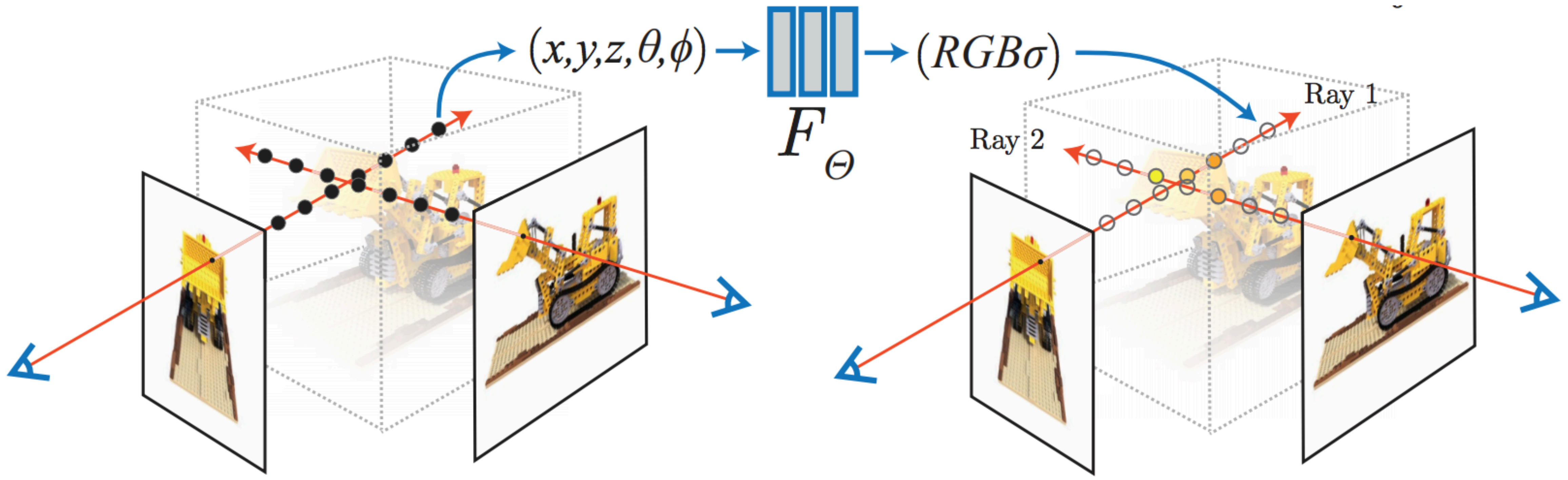
Kalantari, Hedman, Choi, Wang, et al.

Neural 3D representation



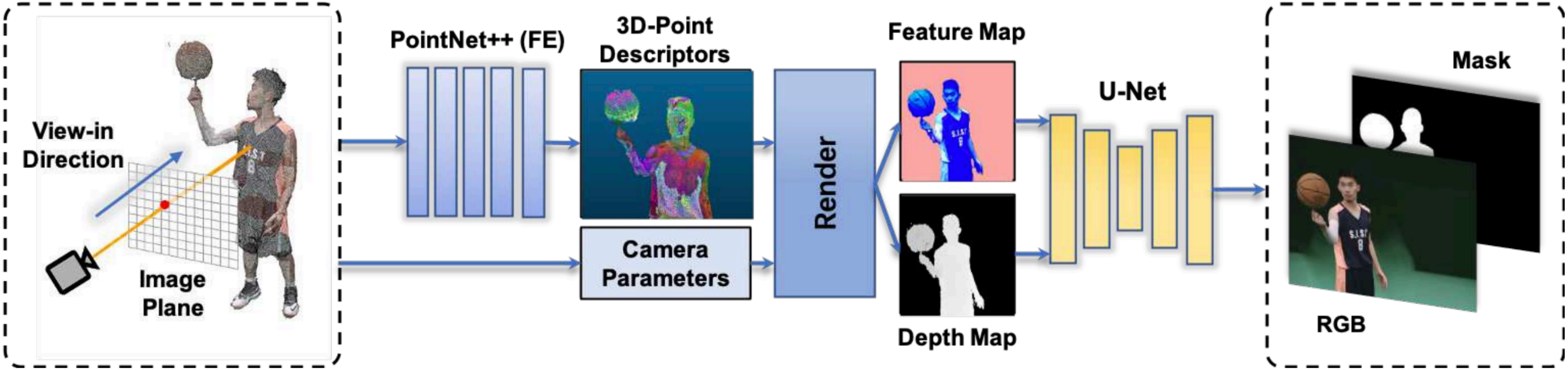
Sitzmann, Lombardi, Wu, Aliev, Thies, et al.

NeRF-like works



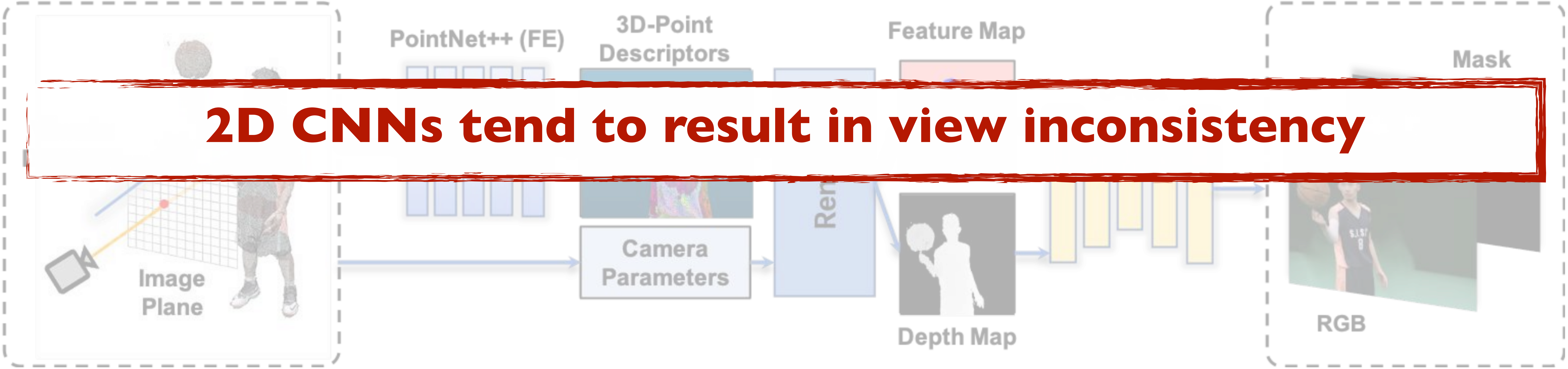
Mildenhall, Yu, Trevithick, Liu, Reiser, et al.

Related work: 2D CNN-based rendering



Multi-view Neural Human Rendering. In CVPR, 2020.

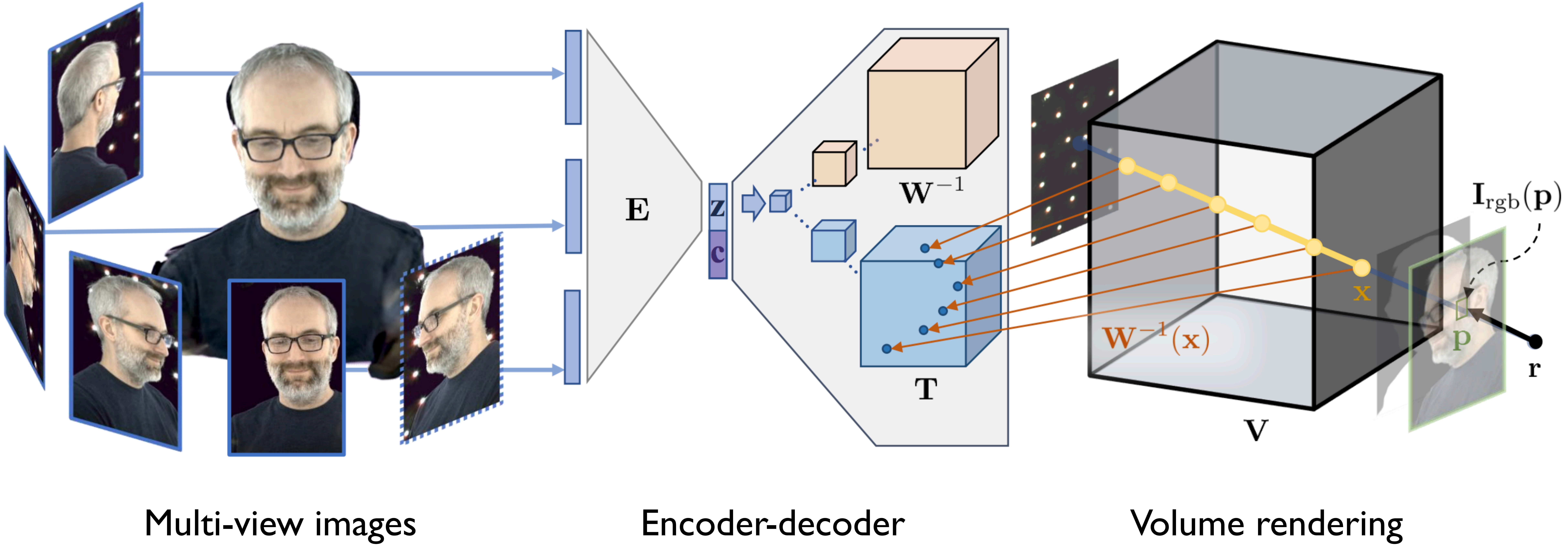
Related work: 2D CNN-based rendering



2D CNNs tend to result in view inconsistency

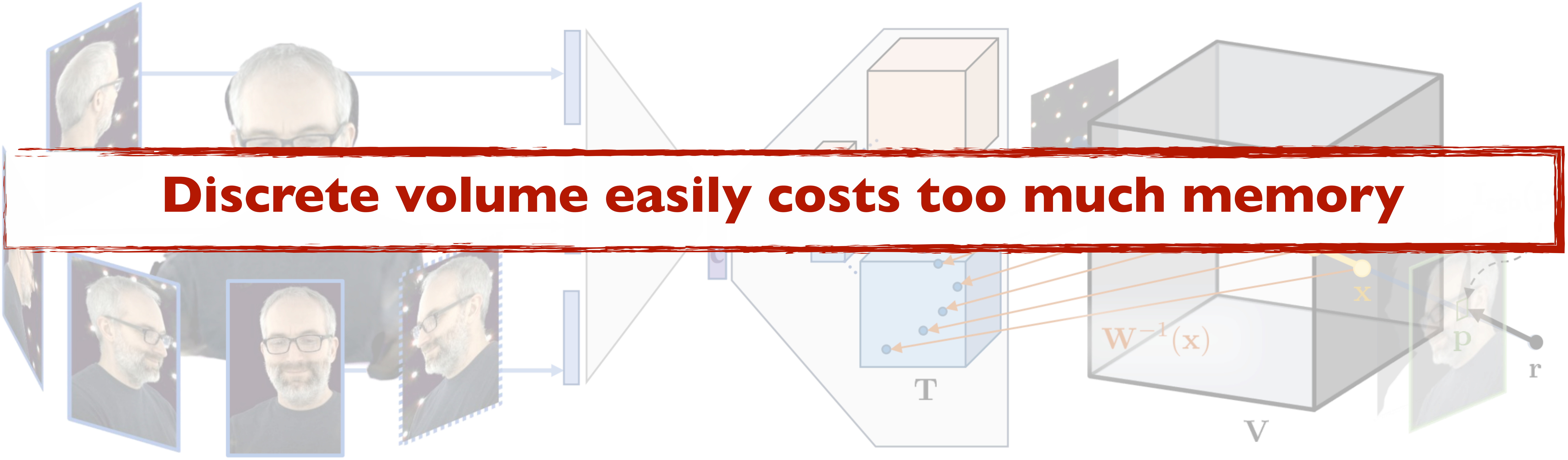
Multi-view Neural Human Rendering. In CVPR, 2020.

Related work: RGB-alpha volume



Neural Volumes: Learning Dynamic Renderable Volumes from Images. In SIGGRAPH, 2019.

Related work: RGB-alpha volume

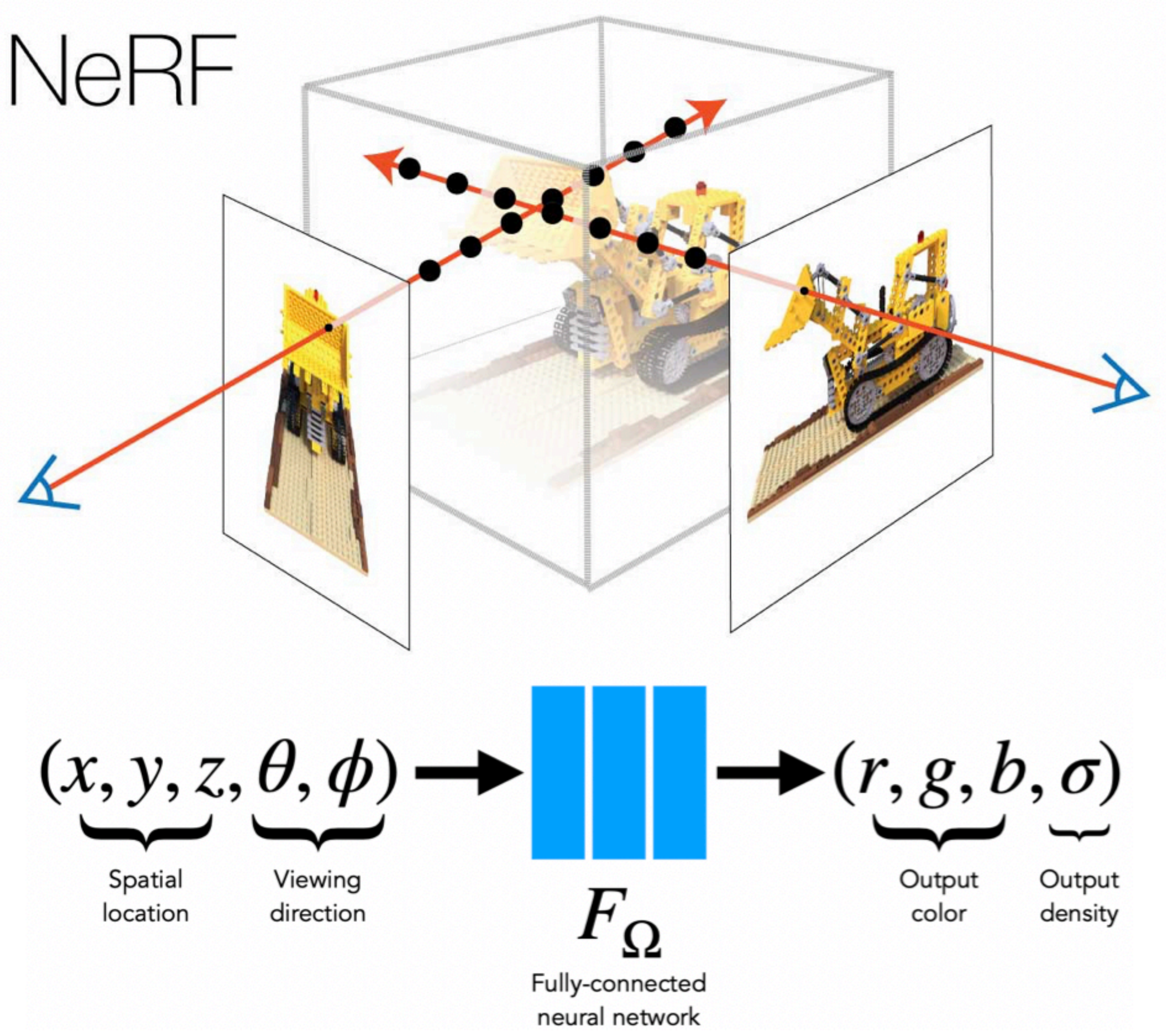


Discrete volume easily costs too much memory

Multi-view images Encoder-decoder Volume rendering

Neural Volumes: Learning Dynamic Renderable Volumes from Images. In SIGGRAPH, 2019.

Related work: Neural radiance field



Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

Challenges for NeRF

- Cannot handle dynamic scenes.

- Require dense input views.

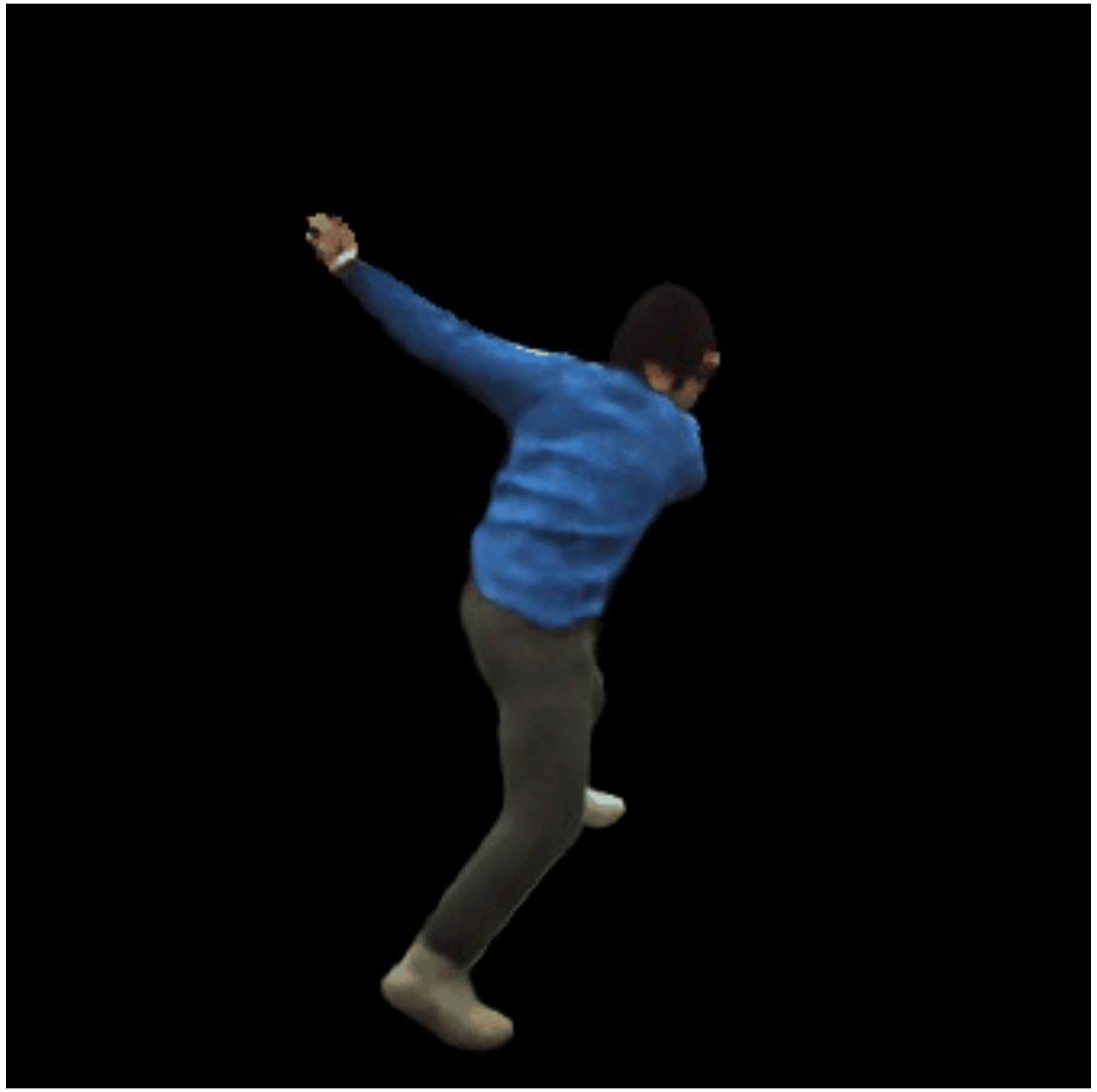
Challenges for NeRF

- Cannot handle dynamic scenes.

- Require dense input views.

Our task: Produce free-viewpoint videos from sparse multi-view videos

Input: 4-view video



Output: free viewpoint video

Our task: Produce free-viewpoint videos from sparse multi-view videos

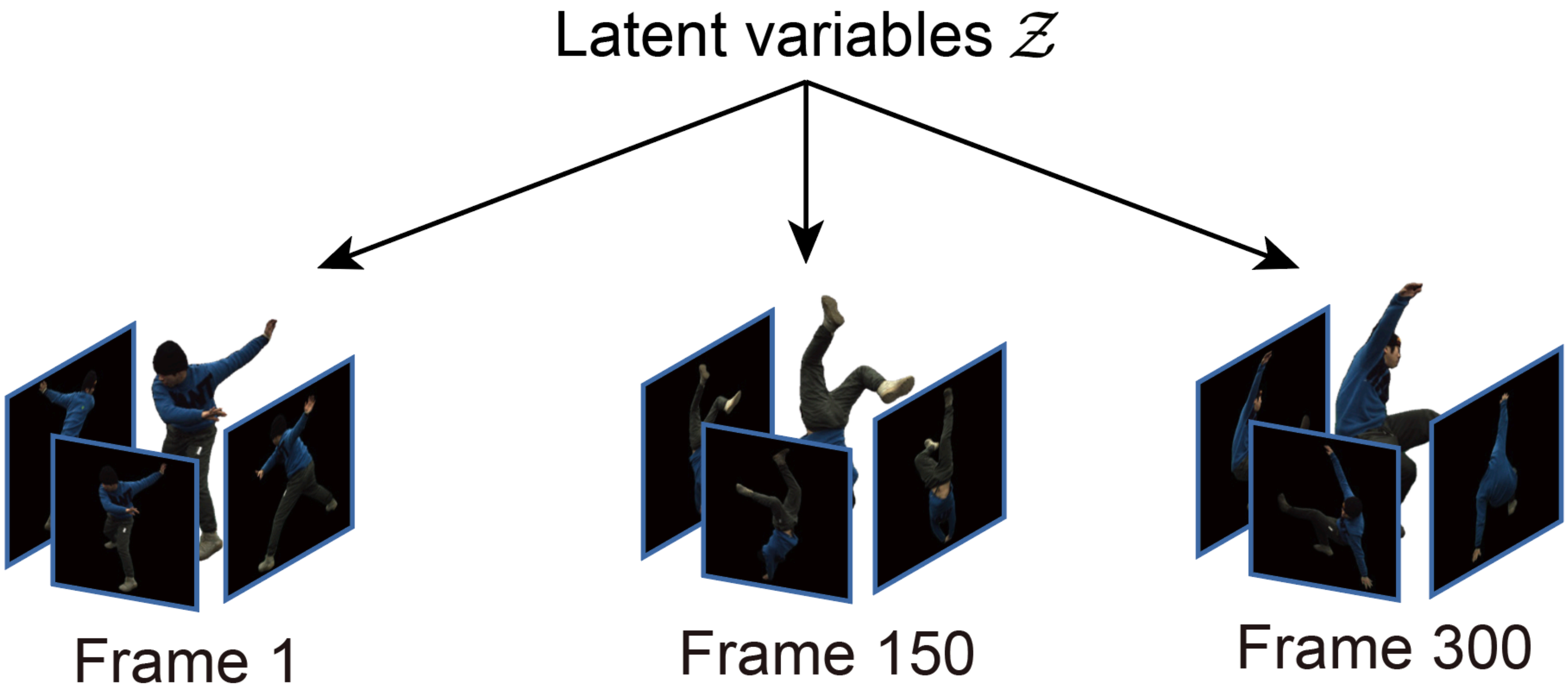
Motivation: Integrate temporal information for more observations

Input: 4-view video

Output: free viewpoint video

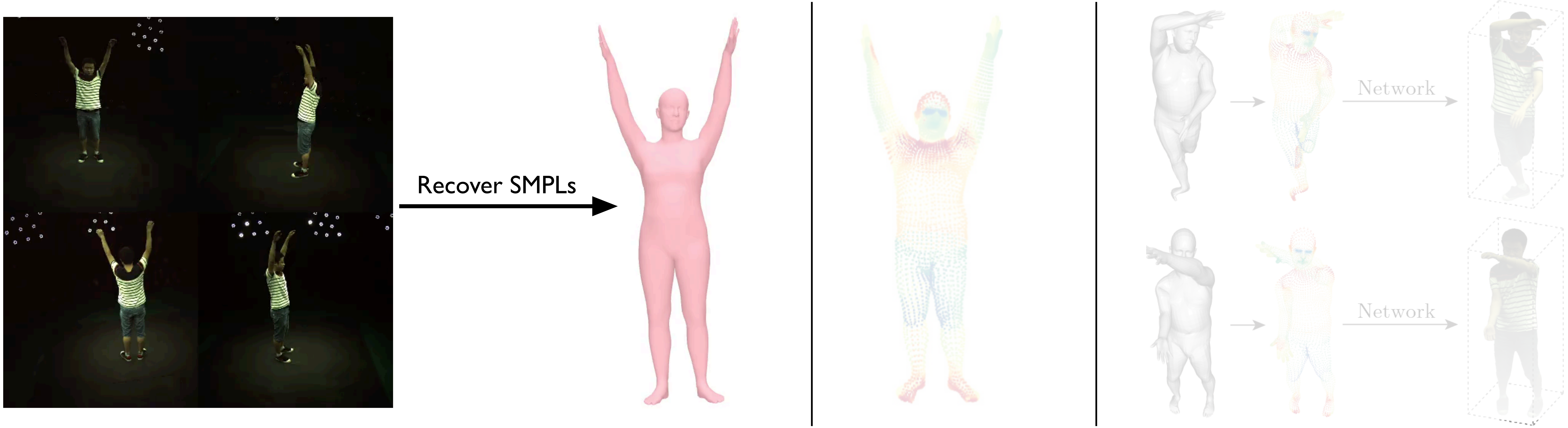
Key idea: Integrate temporal information with latent variable model

Generate scenes at different video frames from the same set of latent variables



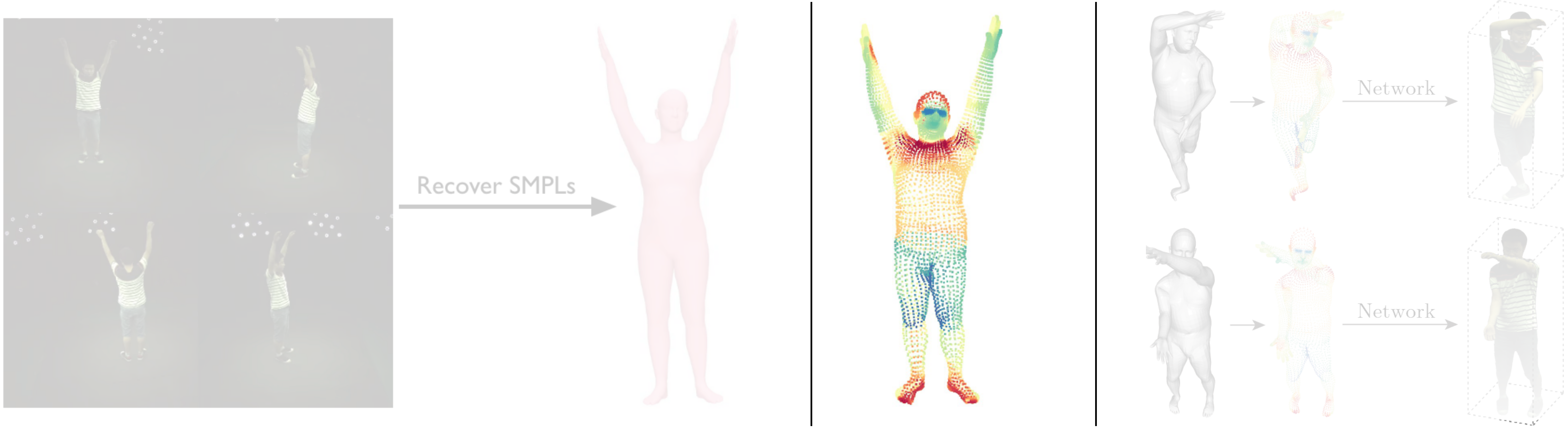
Overview of our method

- Human motion capture from multi-view videos.
- Structured latent codes.
- Generate neural radiance fields from structured latent codes.



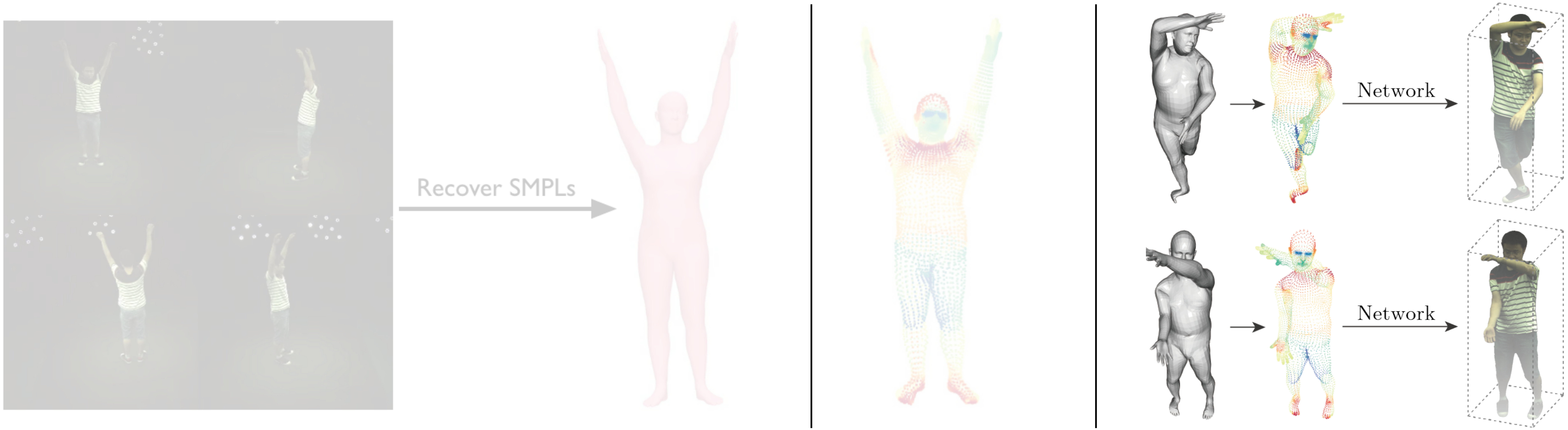
Overview of our method

- Human motion capture from multi-view videos.
- Structured latent codes.
- Generate neural radiance fields from structured latent codes.



Overview of our method

- Human motion capture from multi-view videos.
- Structured latent codes.
- Generate neural radiance fields from structured latent codes.



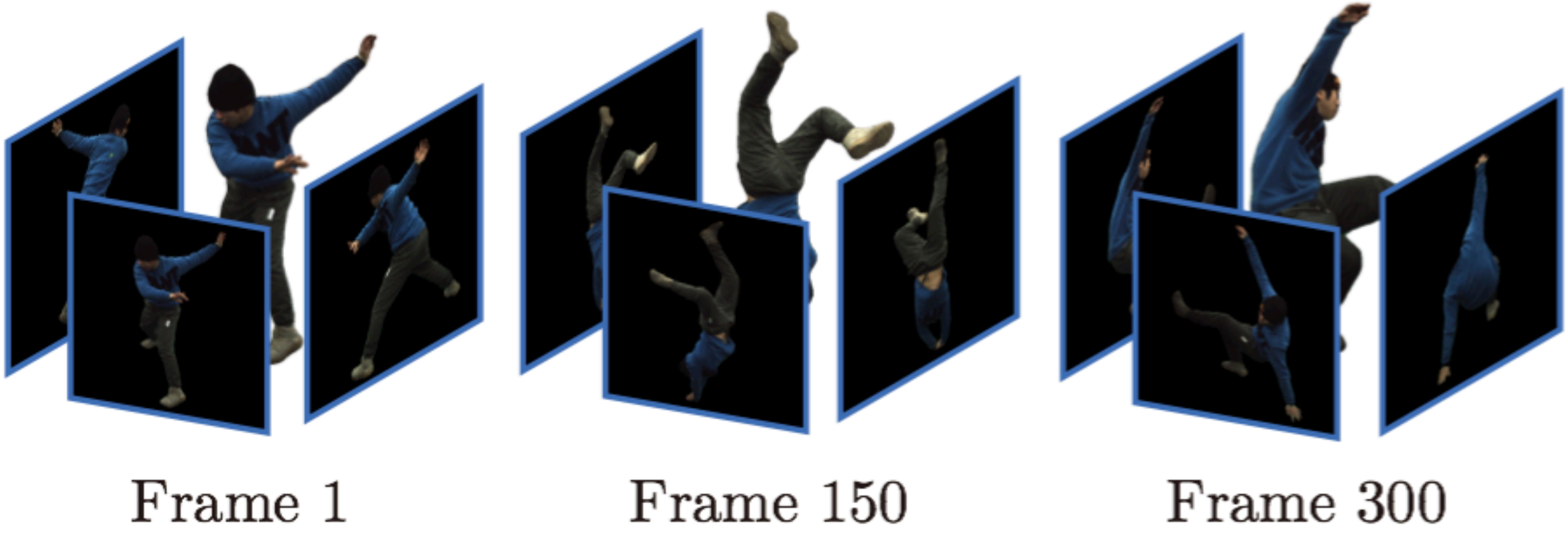
Method: 1) Human motion capture

Integrating temporal information requires us to associate different video frames

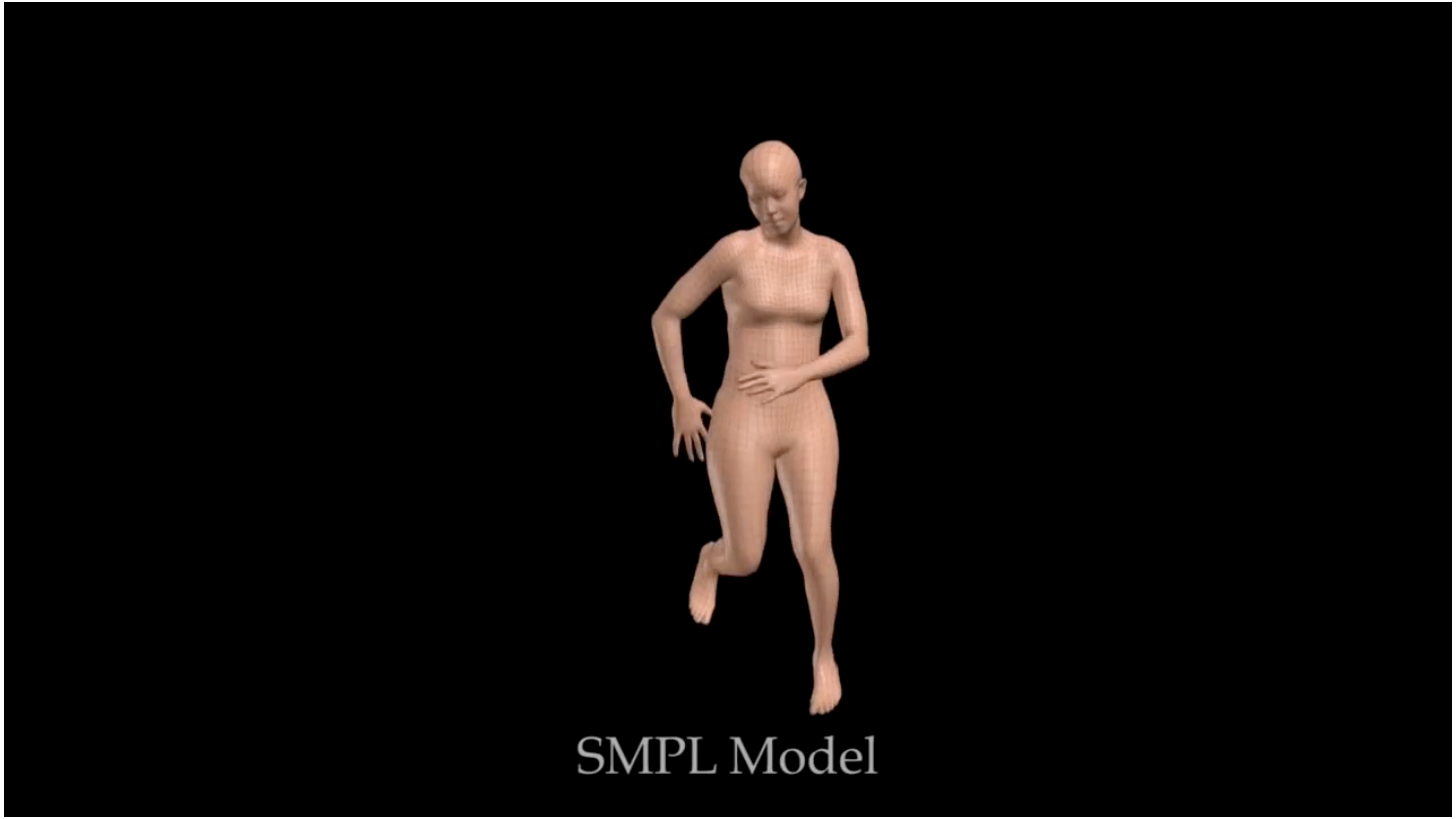
➔ need correspondences

➔ need proxy geometry

➔ SMPL model !



SMPL can be accurately recovered from sparse multi-view videos



https://www.youtube.com/watch?v=kuBlUyHeV5U

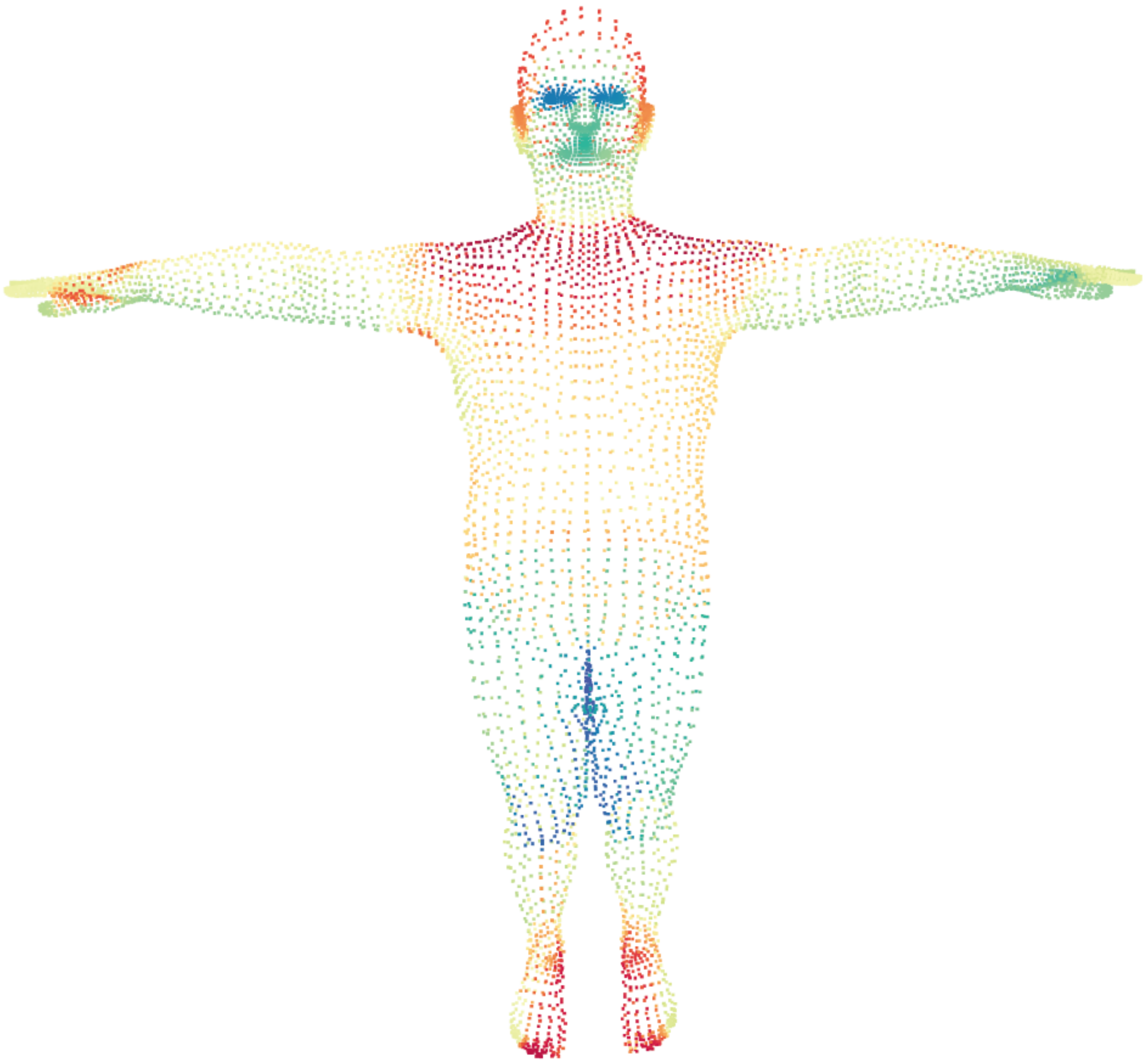
Method: 1) Human motion capture

Capture human motion using https://github.com/zju3dv/EasyMocap

Recover SMPLs

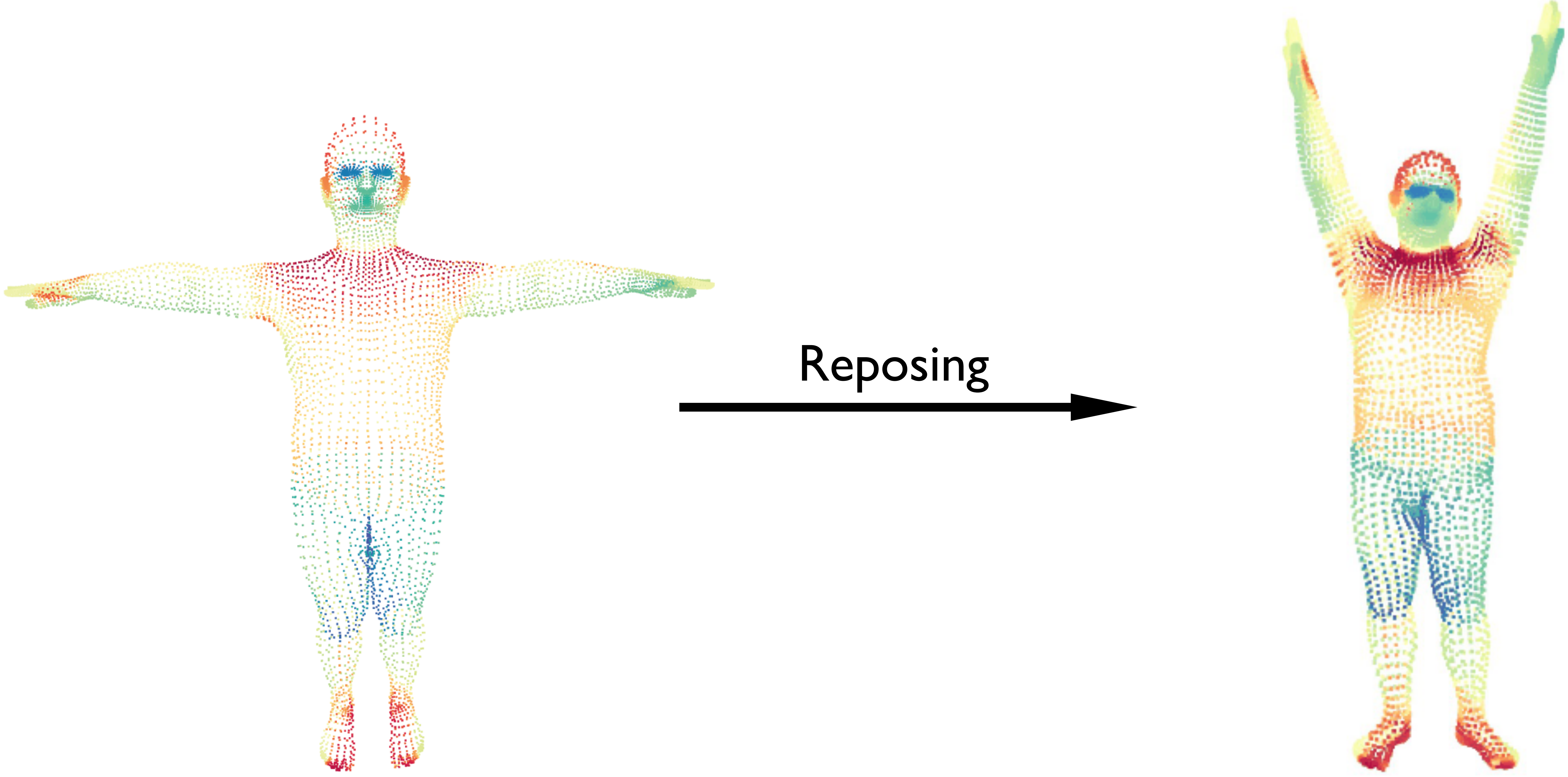
Method: 2) Define structured latent codes on SMPL

For each SMPL vertex, we assign a learnable latent code

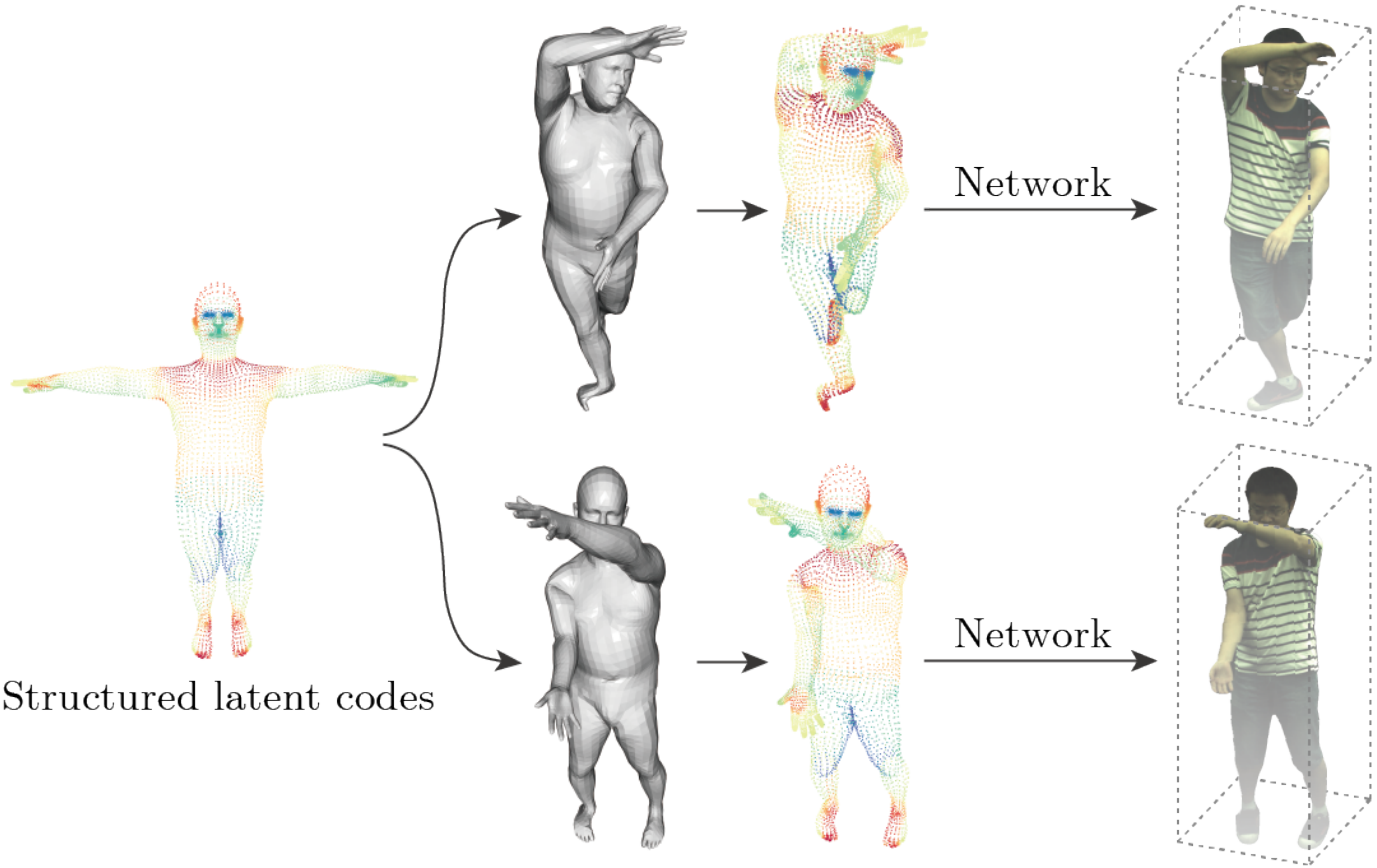


Method: 2) Define structured latent codes on SMPL

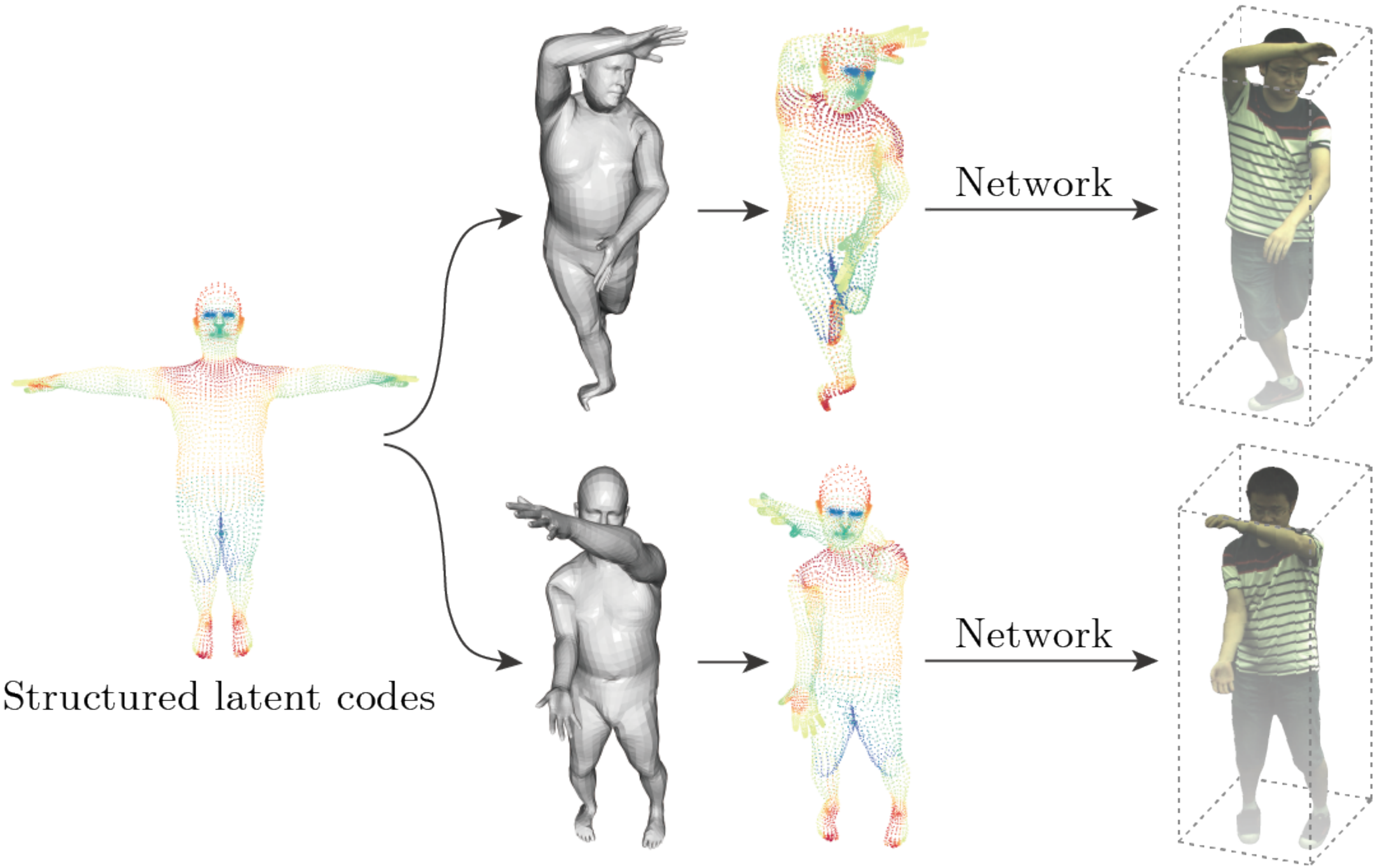
Set the code locations according to the SMPL pose



Method: 3) Generate scenes from structured latent codes

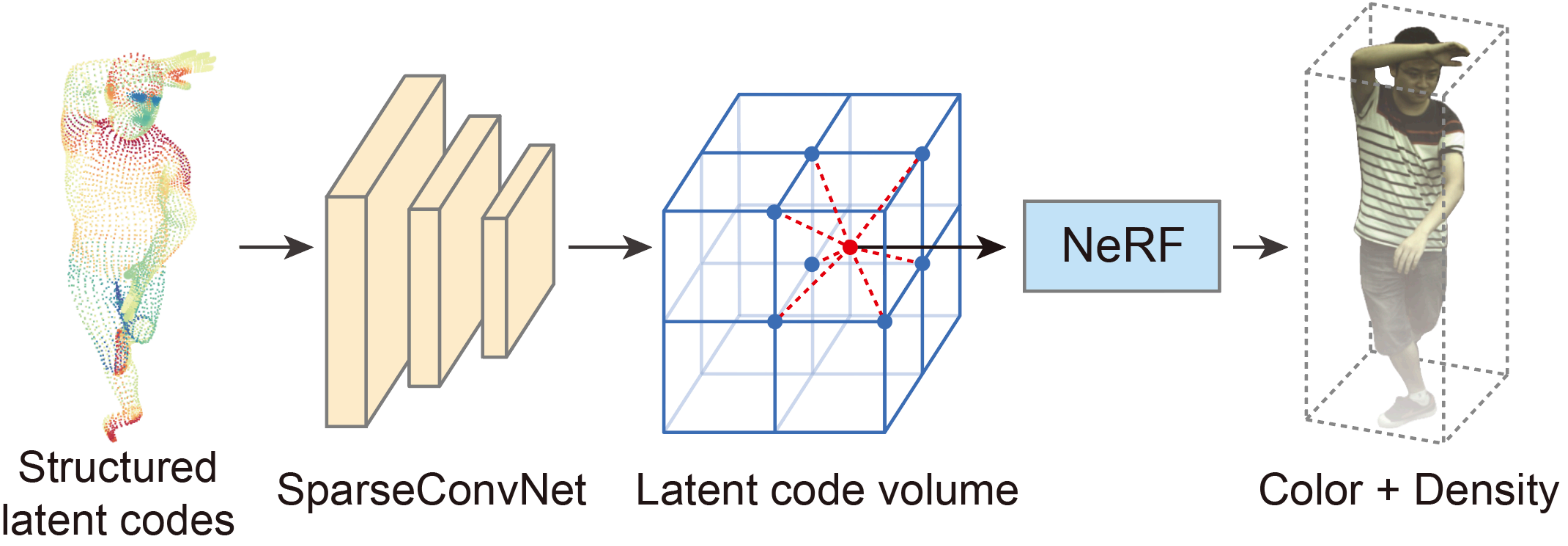


How to generate continuous scenes from discrete latent codes



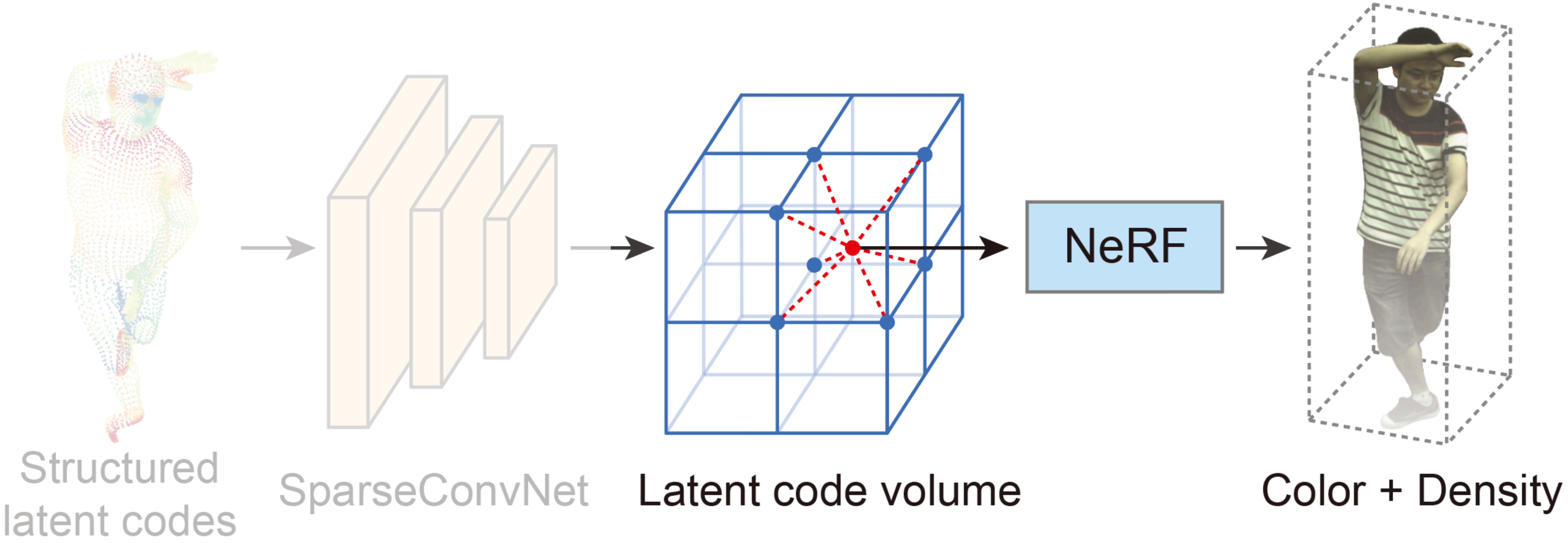
Method: 3) Generate scenes from structured latent codes

The network pipeline



Method: 3) Generate scenes from structured latent codes

The network pipeline



Method: 3) Generate scenes from structured latent codes

The network pipeline

Results on ZJU-MoCap dataset

training on 4-view videos

Novel view synthesis of frame 1

NeRF [1]

[1] Mildenhall, Ben, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

[2] Lombardi, Stephen, et al. Neural volumes: Learning dynamic renderable volumes from images. In SIGGRAPH, 2019.

[3] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.

Novel view synthesis of frame 1

NeRF [1] Neural Volumes [2]

[1] Mildenhall, Ben, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

[2] Lombardi, Stephen, et al. Neural volumes: Learning dynamic renderable volumes from images. In SIGGRAPH, 2019.

[3] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.

Novel view synthesis of frame 1

NeRF [1] Neural Volumes [2] NHR [3]

[1] Mildenhall, Ben, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

[2] Lombardi, Stephen, et al. Neural volumes: Learning dynamic renderable volumes from images. In SIGGRAPH, 2019.

[3] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.

Novel view synthesis of frame 1

NeRF [1] Neural Volumes [2] NHR [3] OURS

[1] Mildenhall, Ben, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

[2] Lombardi, Stephen, et al. Neural volumes: Learning dynamic renderable volumes from images. In SIGGRAPH, 2019.

[3] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.

Novel view synthesis of frame 1

NeRF [1] Neural Volumes [2] NHR [3] OURS

[1] Mildenhall, Ben, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

[2] Lombardi, Stephen, et al. Neural volumes: Learning dynamic renderable volumes from images. In SIGGRAPH, 2019.

[3] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.

Novel view synthesis of dynamic human

Neural Volumes [1] NHR [2] OURS

[1] Lombardi, Stephen, et al. Neural volumes: Learning dynamic renderable volumes from images. In SIGGRAPH, 2019.

[2] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.

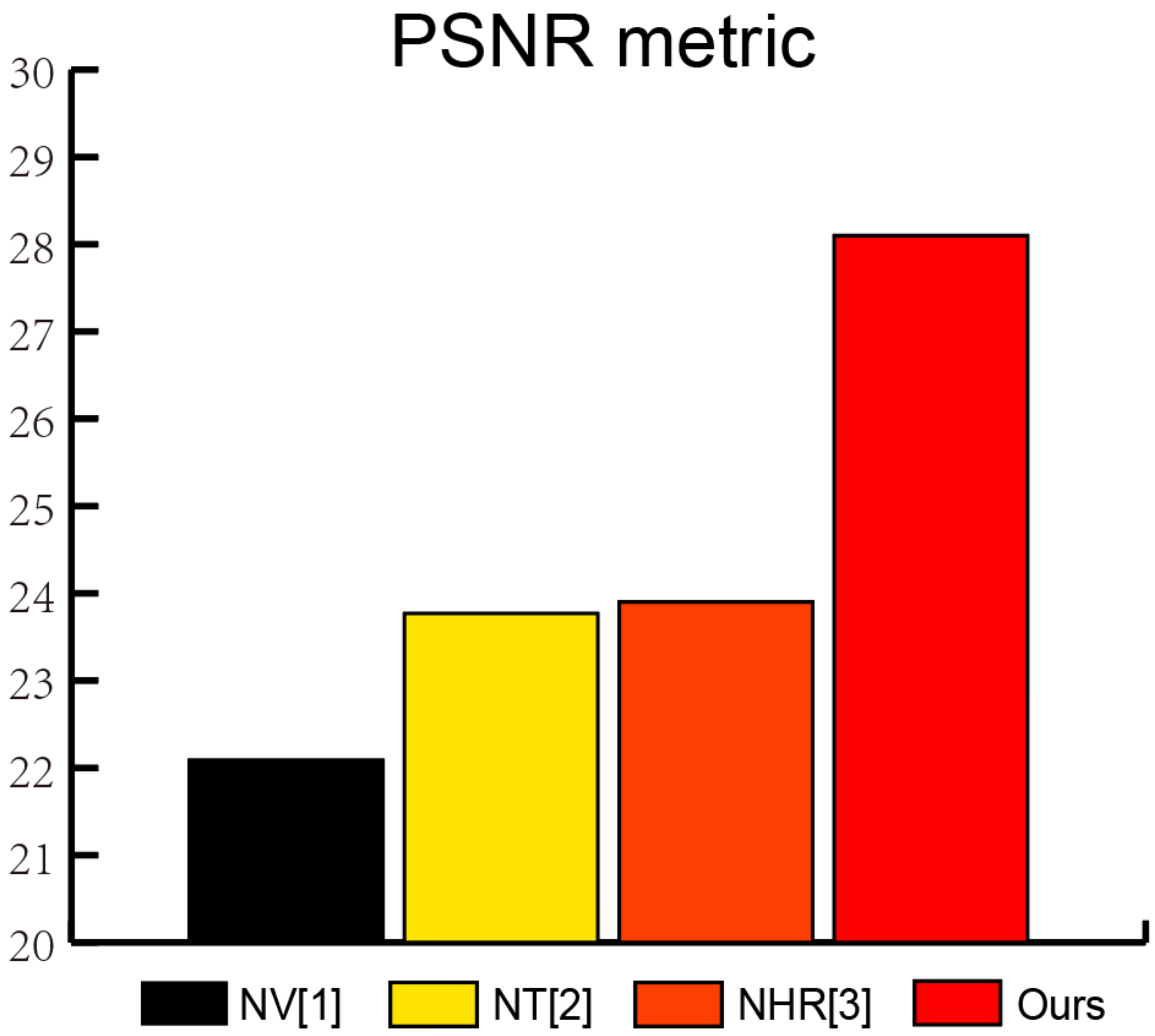
Novel view synthesis of dynamic human

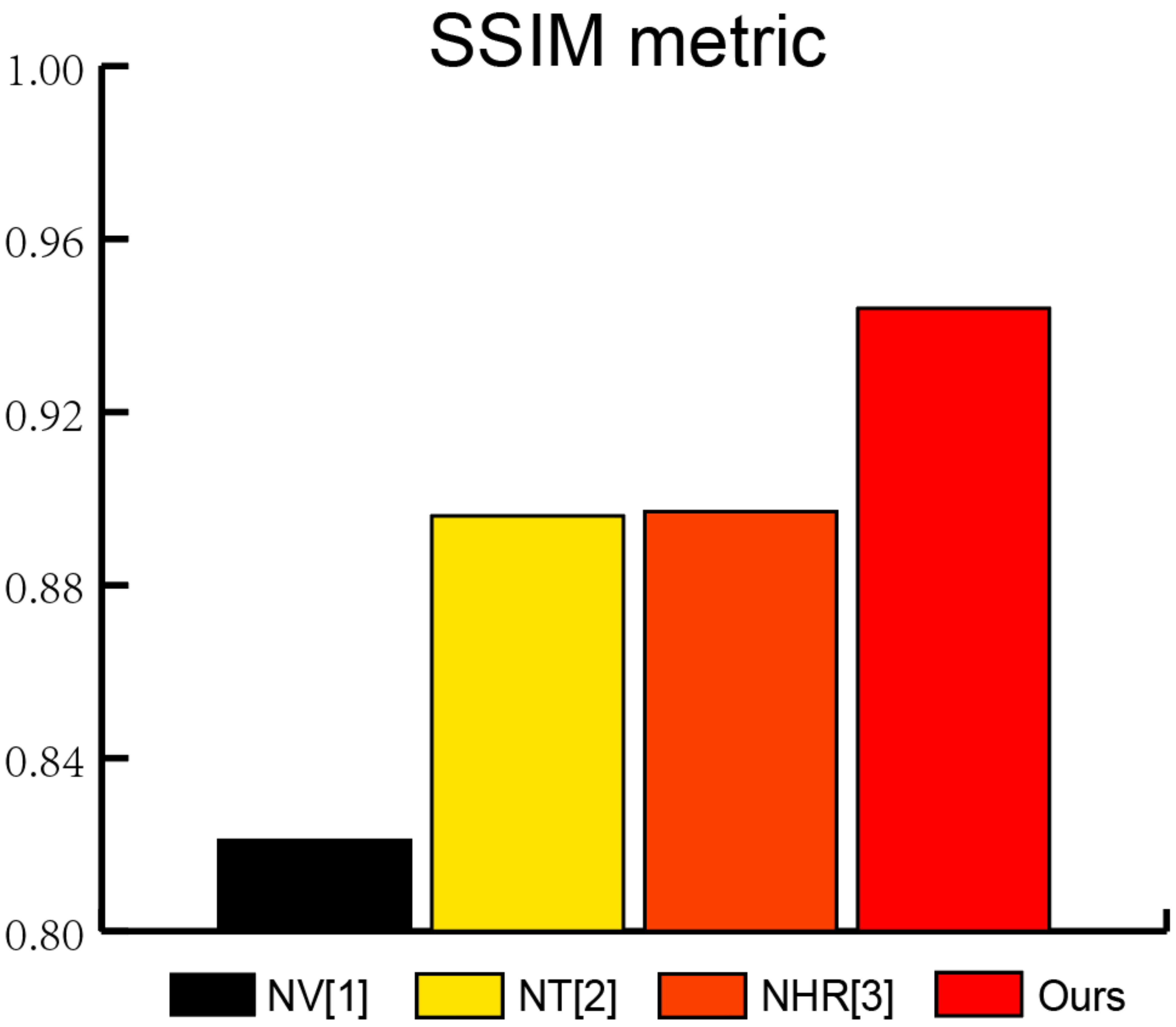
Neural Volumes [1] NHR [2] OURS

[1] Lombardi, Stephen, et al. Neural volumes: Learning dynamic renderable volumes from images. In SIGGRAPH, 2019.

[2] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.

Quantitative comparison





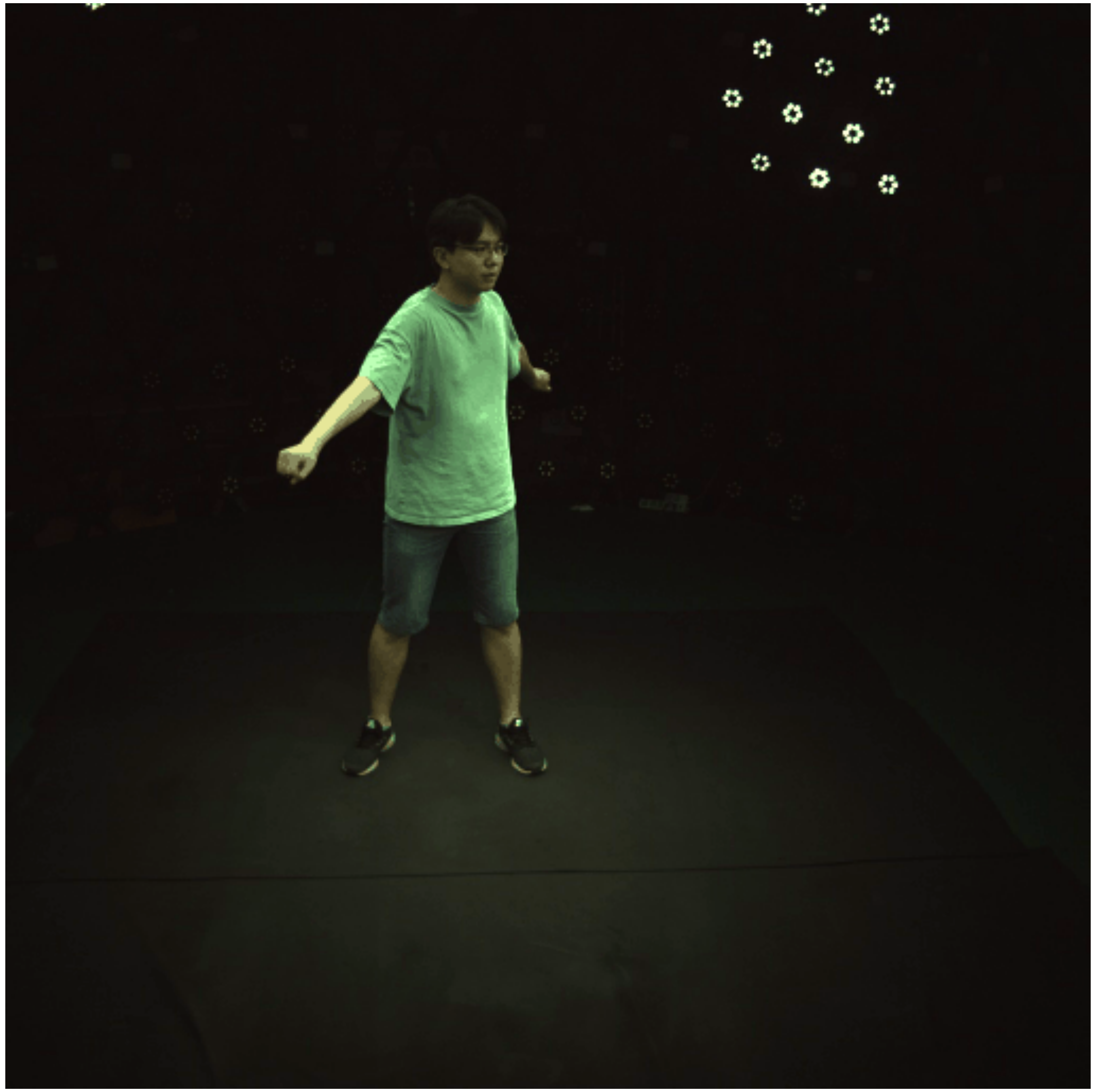
[1] Lombardi, Stephen, et al. Neural volumes: Learning dynamic renderable volumes from images. In SIGGRAPH, 2019.
[2] Thies, Justus, et al. Deferred neural rendering: Image synthesis using neural textures. In ACM TOG, 2019.
[3] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.

Ablation studies: video length

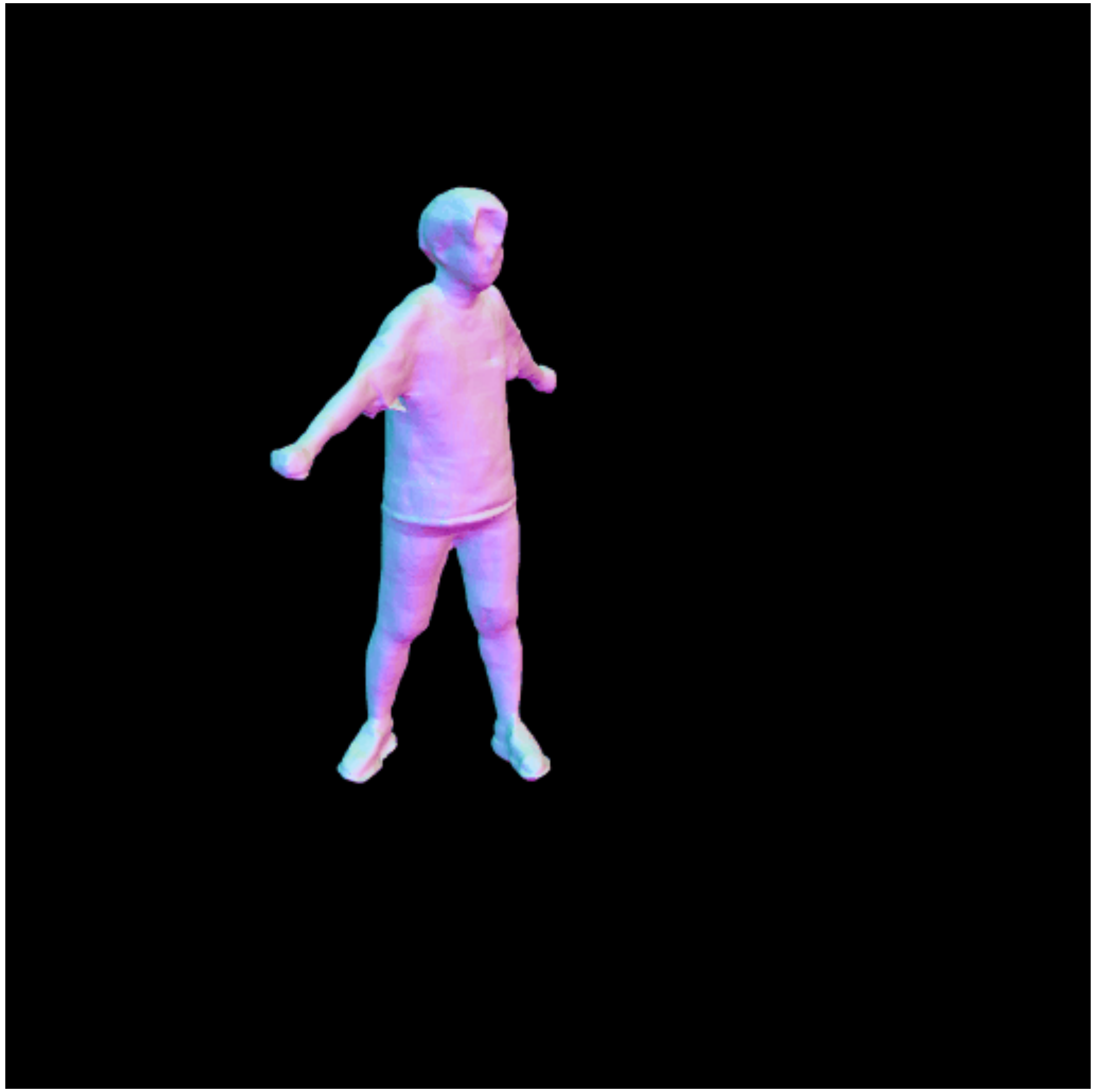
Frames	1	60	300	600	1200
PSNR	25.64	30.14	**30.66**	30.59	29.97
SSIM	0.940	0.970	**0.971**	0.970	0.970

Table 4: **Results of models trained with different numbers of training frames.** We train models on 1, 60, 300, 600, and 1200 frames and test on the first frame of "Twirl".

3D Reconstruction



Input video



Reconstructed geometry

3D Reconstruction

PIFuHD

Saito, Shunsuke, et al. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In CVPR, 2020.

3D Reconstruction

PIFuHD

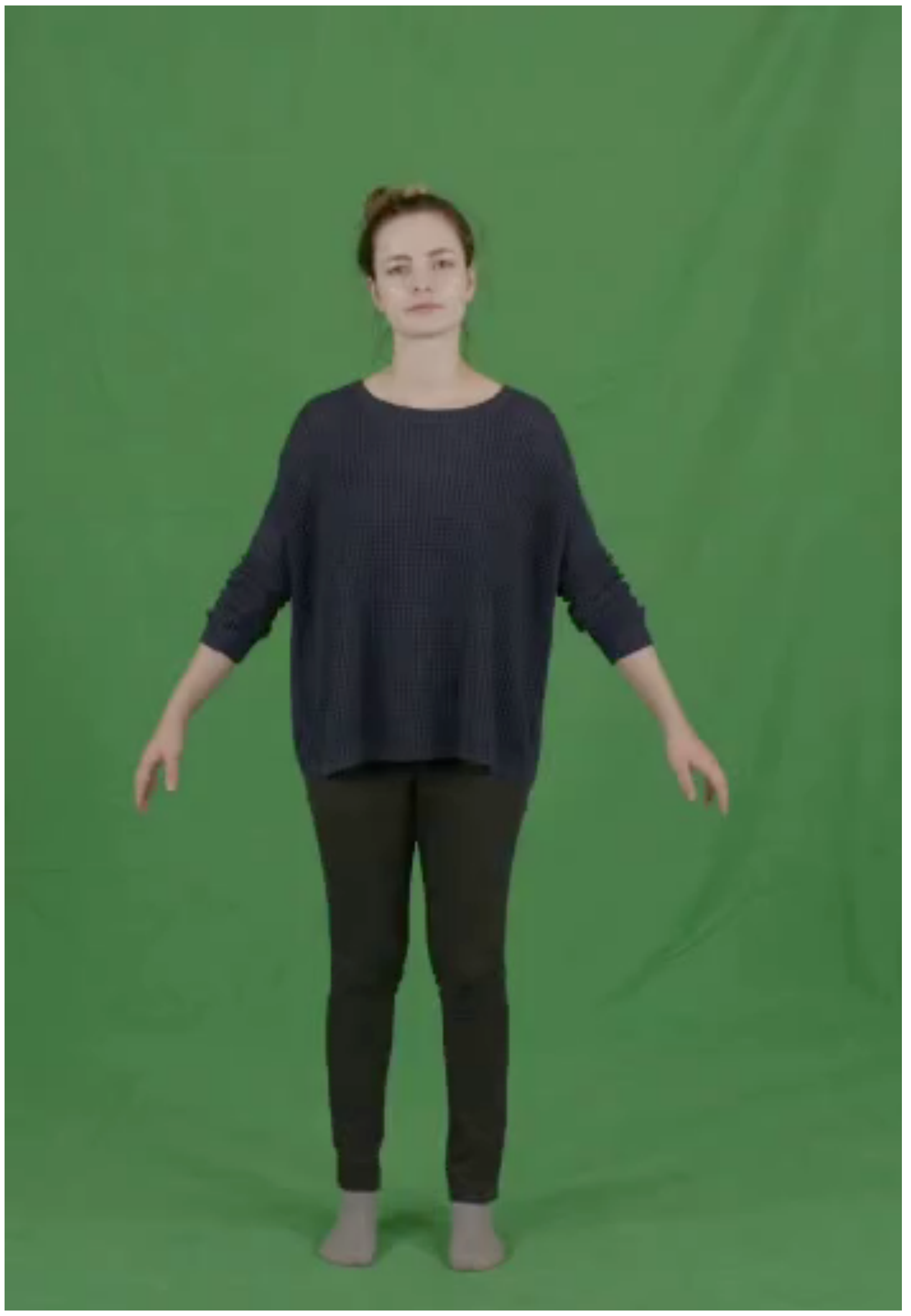
OURS

Saito, Shunsuke, et al. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In CVPR, 2020.

Results on People-Snapshot dataset

training on monocular videos

Results of reconstruction and view synthesis



People-Snapshot [1]   Ours   People-Snapshot [1]   Ours

[1] Alldieck, Thiemo, et al. Video based reconstruction of 3d people models. In CVPR, 2018.

Results of reconstruction and view synthesis

People-Snapshot [1] Ours People-Snapshot [1] Ours

[1] Alldieck, Thiemo, et al. Video based reconstruction of 3d people models. In CVPR, 2018.

Summary

- We propose structured latent codes, which combines SMPL model and NeRF and enables us to represent dynamic humans.

Summary

- We propose structured latent codes, which combines SMPL model and NeRF and enables us to represent dynamic humans.
- As a latent variable model, our method naturally integrates temporal information across video frames.

Summary

- We propose structured latent codes, which combines SMPL model and NeRF and enables us to represent dynamic humans.
- As a latent variable model, our method naturally integrates temporal information across video frames.
- Neural Body can reconstruct high-quality 3D human models from very sparse multi-view videos.

Limitations

- Since our model is built on the SMPL model, we have difficulty in handling performers with loose clothes.

Limitations

- Since our model is built on the SMPL model, we have difficulty in handling performers with loose clothes.
- Neural Body trains a network for each human subject, which takes about 12 hours and costs a lot of time.

Limitations

- Since our model is built on the SMPL model, we have difficulty in handling performers with loose clothes.
- Neural Body trains a network for each human subject, which takes about 12 hours and costs a lot of time.
- Our method has difficulty in generating high-quality novel views for unseen human poses.

Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies

Sida Peng*, Junting Dong*, Qianqian Wang, Shangzhan Zhang

Qing Shuai, Hujun Bao, Xiaowei Zhou

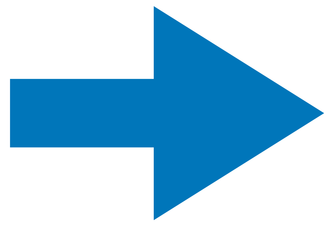
浙江大學 ZHEJIANG UNIVERSITY

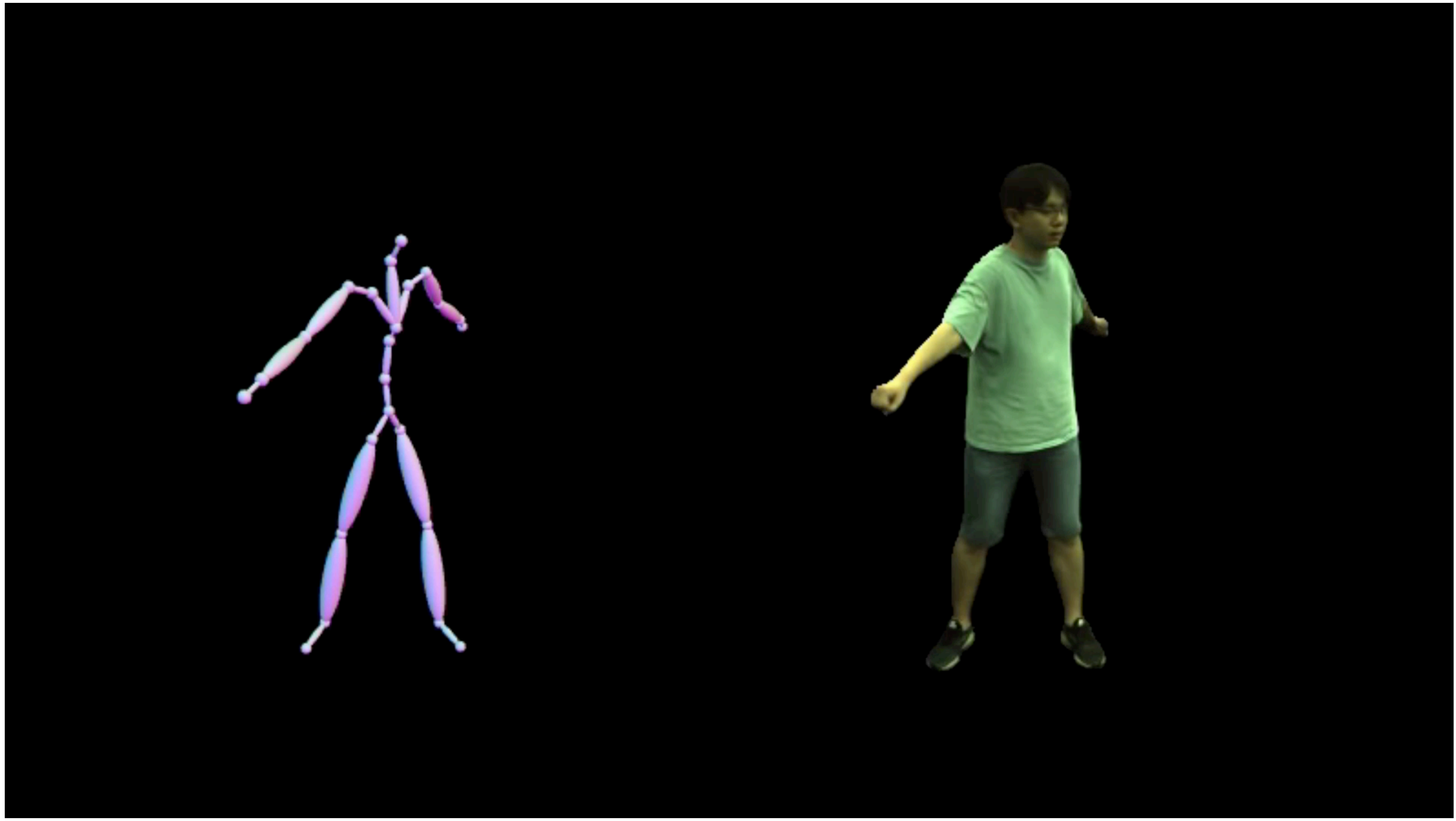
Cornell University

Problem statement

Input: sparse-view videos

Output: animatable human models





Related work: Neural Body

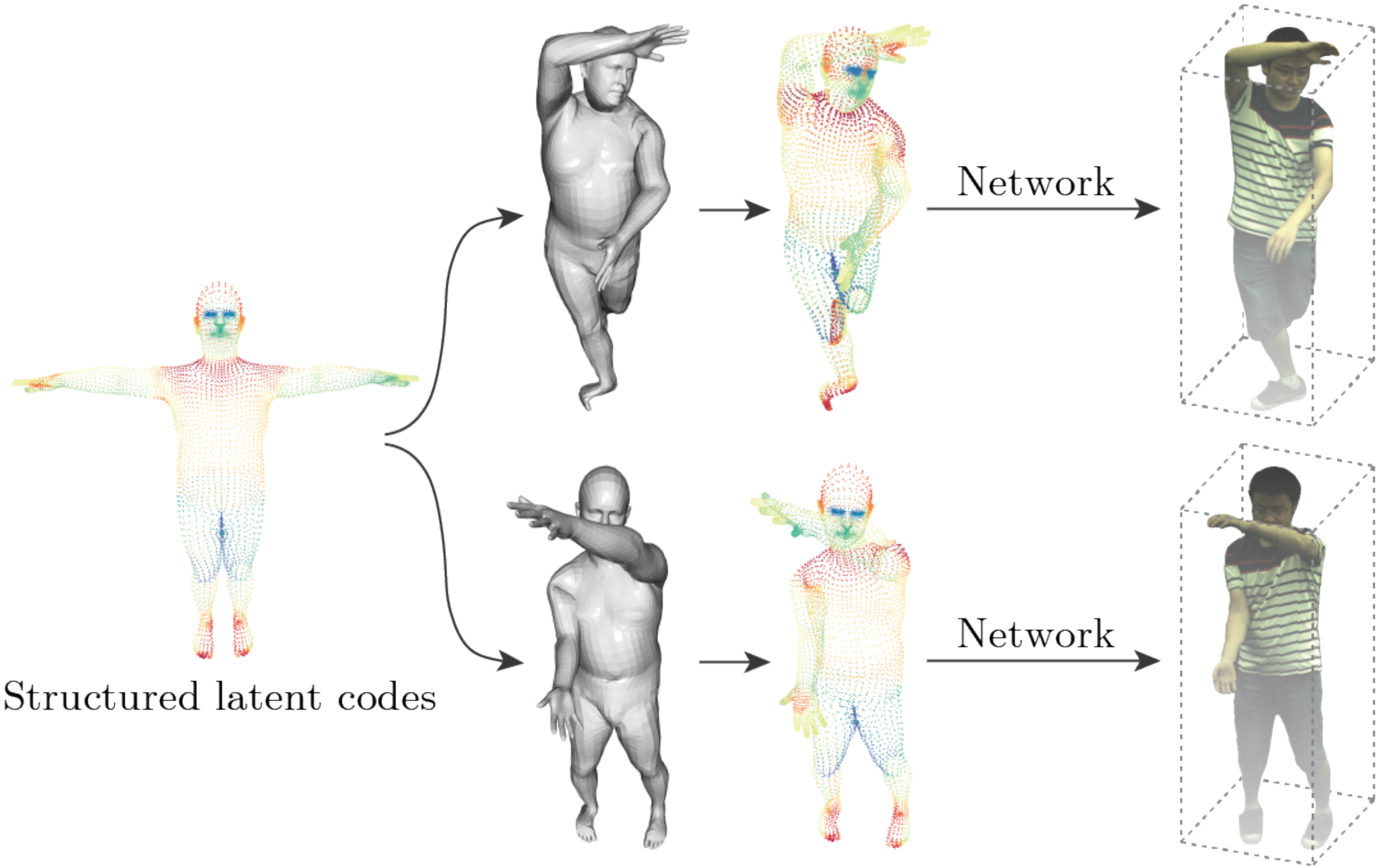


Image credit: Peng, et al. CVPR 2021.

Peng, Sida, et al. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In CVPR, 2021.

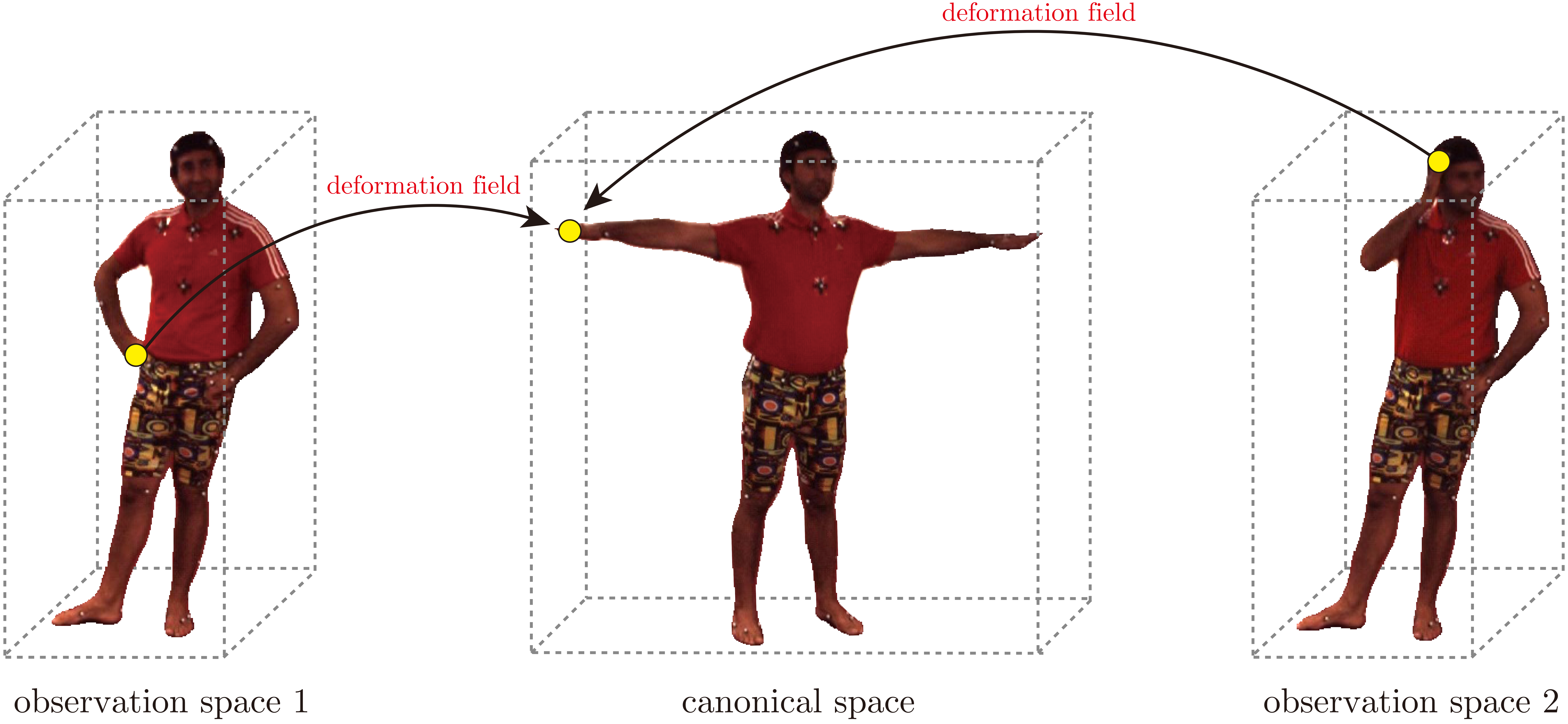
Related work: Neural Body

Limitation

Cannot generalize to unseen human poses

Peng, Sida, et al. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In CVPR, 2021.

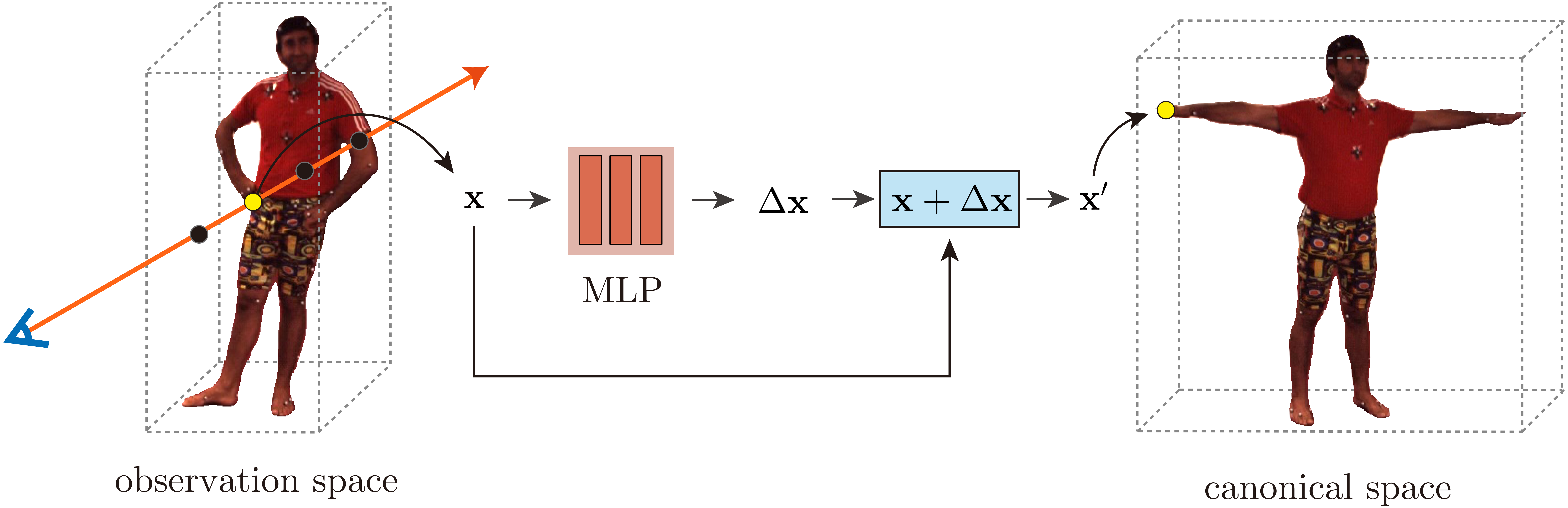
Related work: D-NeRF



Pumarola, Albert, et al. D-nerf: Neural radiance fields for dynamic scenes. In CVPR, 2021.

Related work: D-NeRF

Represent deformation fields as translational vector fields



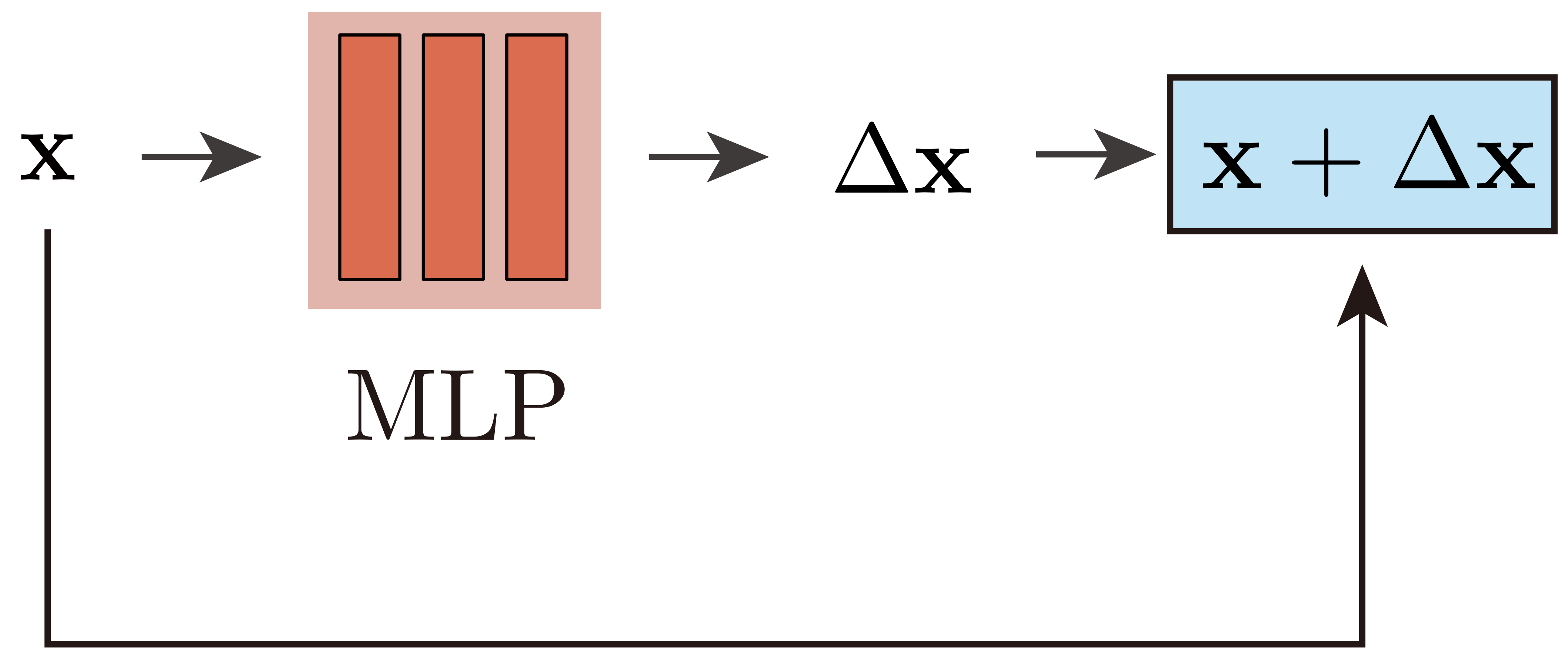
Pumarola, Albert, et al. D-nerf: Neural radiance fields for dynamic scenes. In CVPR, 2021.

Related work: D-NeRF

Limitations

1. Use networks to predict translational vectors, which cannot easily generalize to novel poses.

Translational vector fields



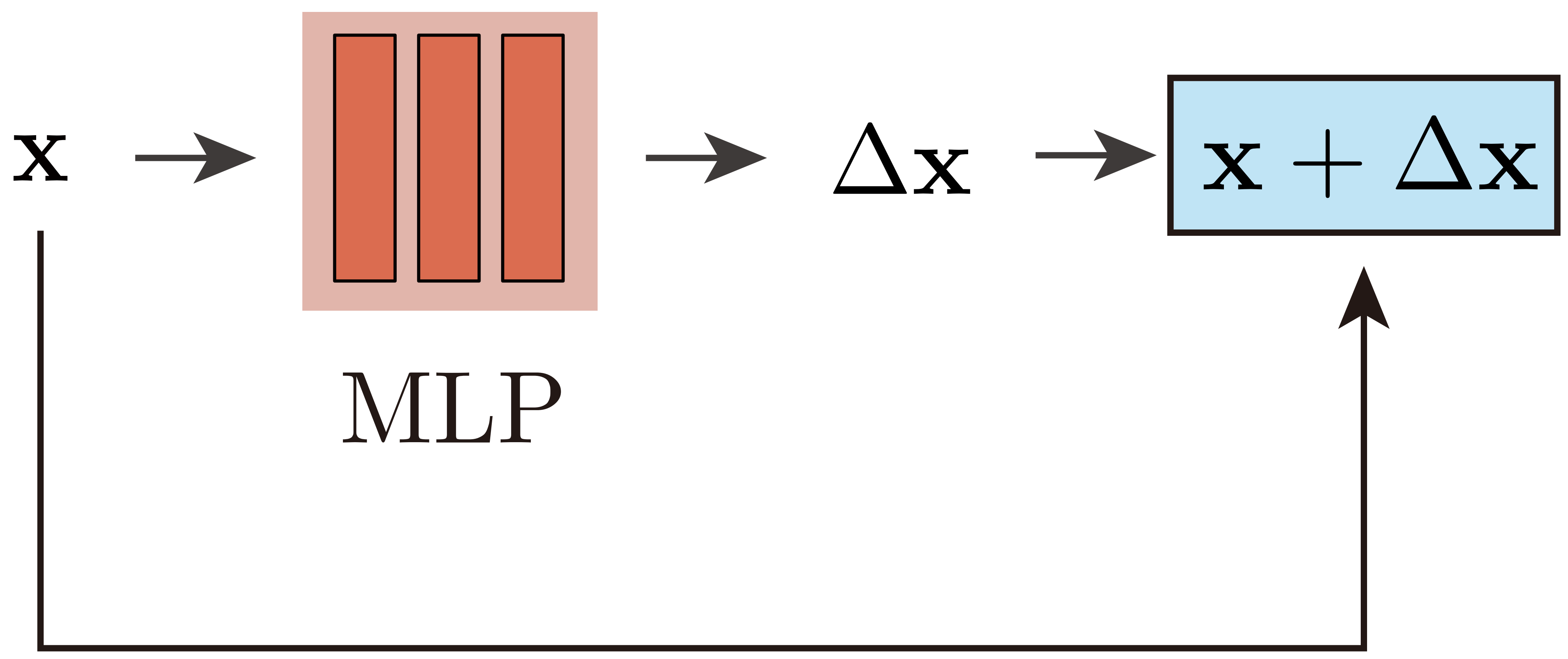
Pumarola, Albert, et al. D-nerf: Neural radiance fields for dynamic scenes. In CVPR, 2021.

Related work: D-NeRF

Limitations

1. Use networks to predict translational vectors, which cannot generalize to novel poses.
2. Optimizing neural radiance fields with vector fields is highly under-constrained.

Translational vector fields



Pumarola, Albert, et al. D-nerf: Neural radiance fields for dynamic scenes. In CVPR, 2021.

Represent deformation fields with LBS models

The blend weight fields are combined with human skeletons to output the deformation fields.

What are linear blend skinning models

Given T1 and T2, how do we transform the red point?

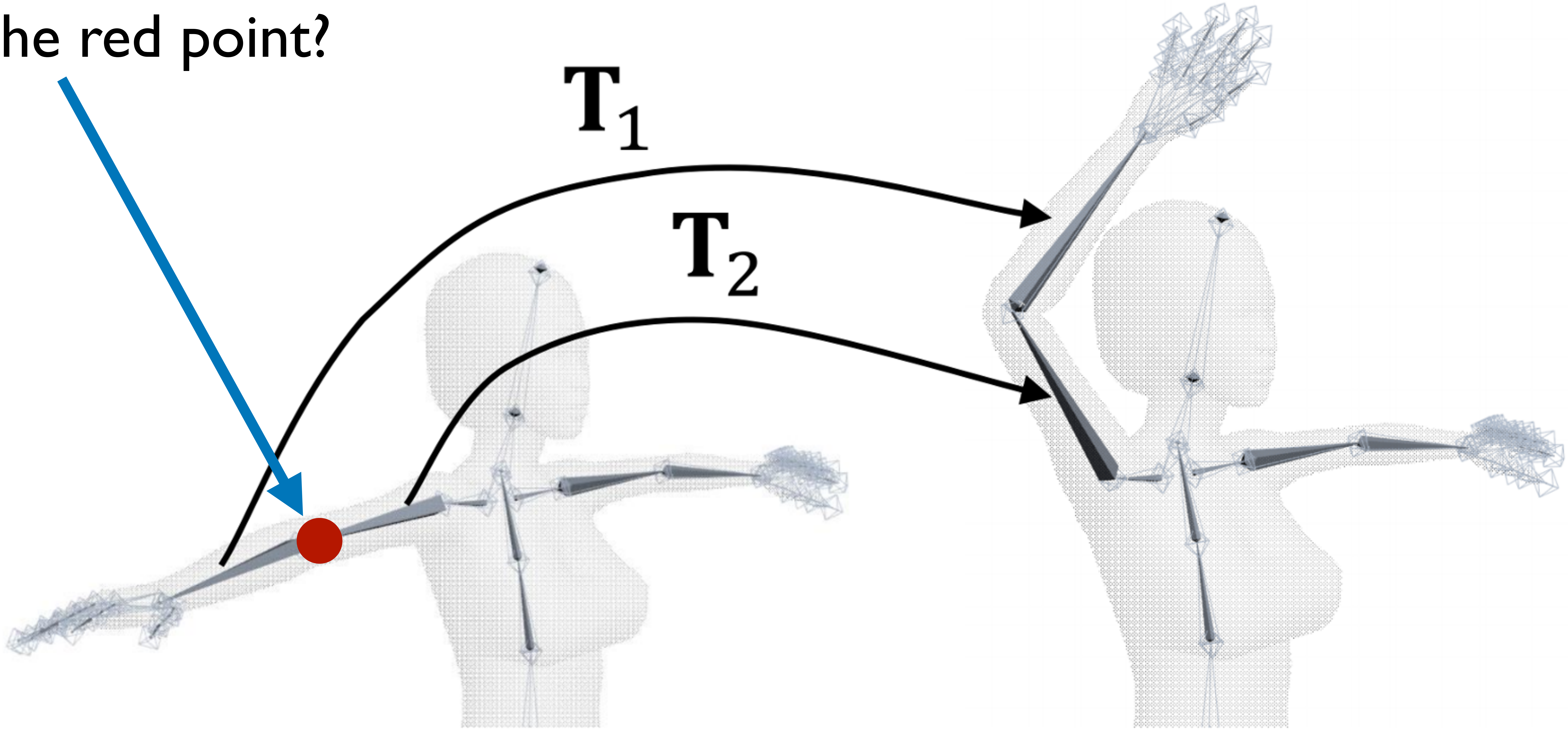


Image credit: https://skinning.org/direct-methods-slides.pdf

What are linear blend skinning models

Given T1 and T2, how do we transform the red point?

$$\left(\sum_i w_i T_i\right) v$$

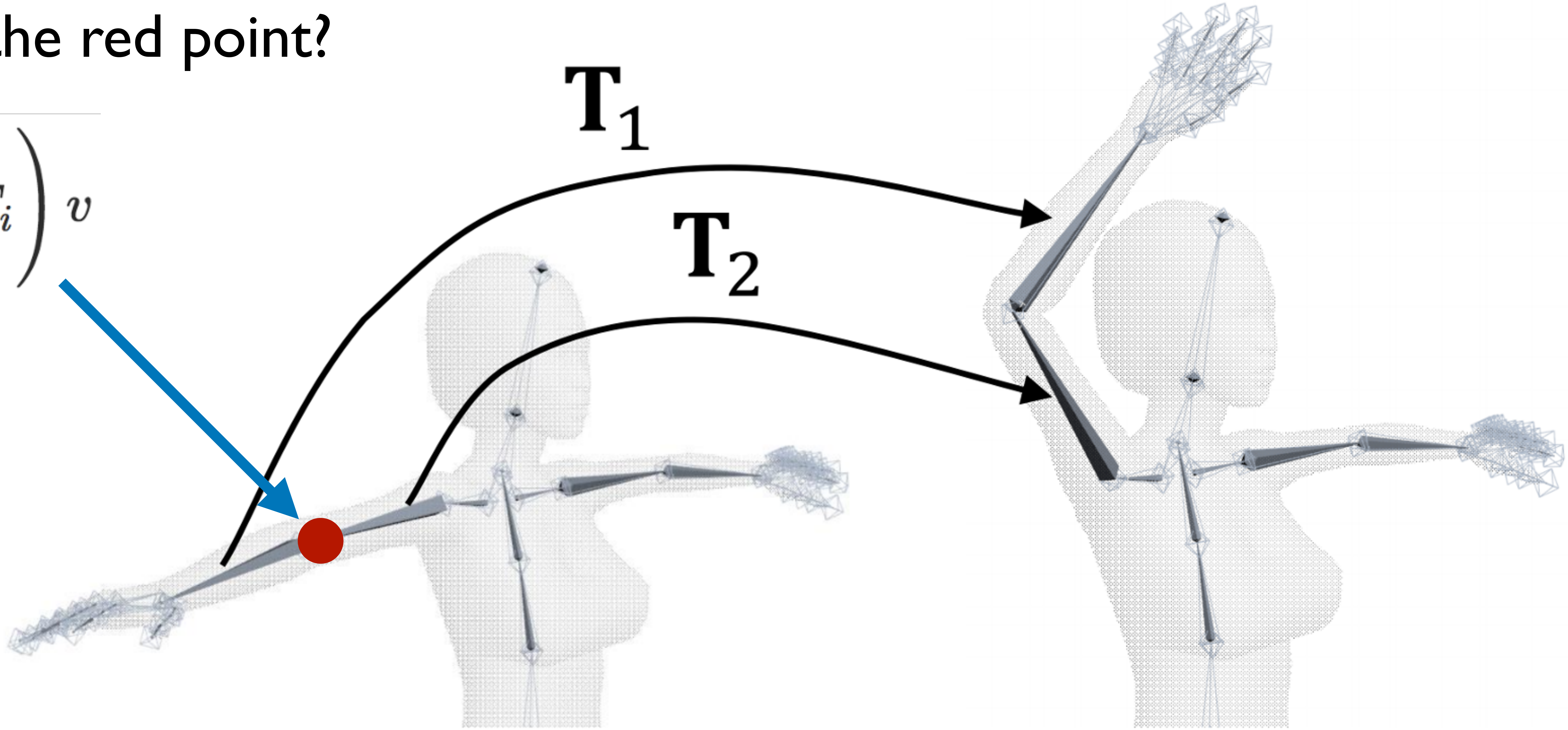
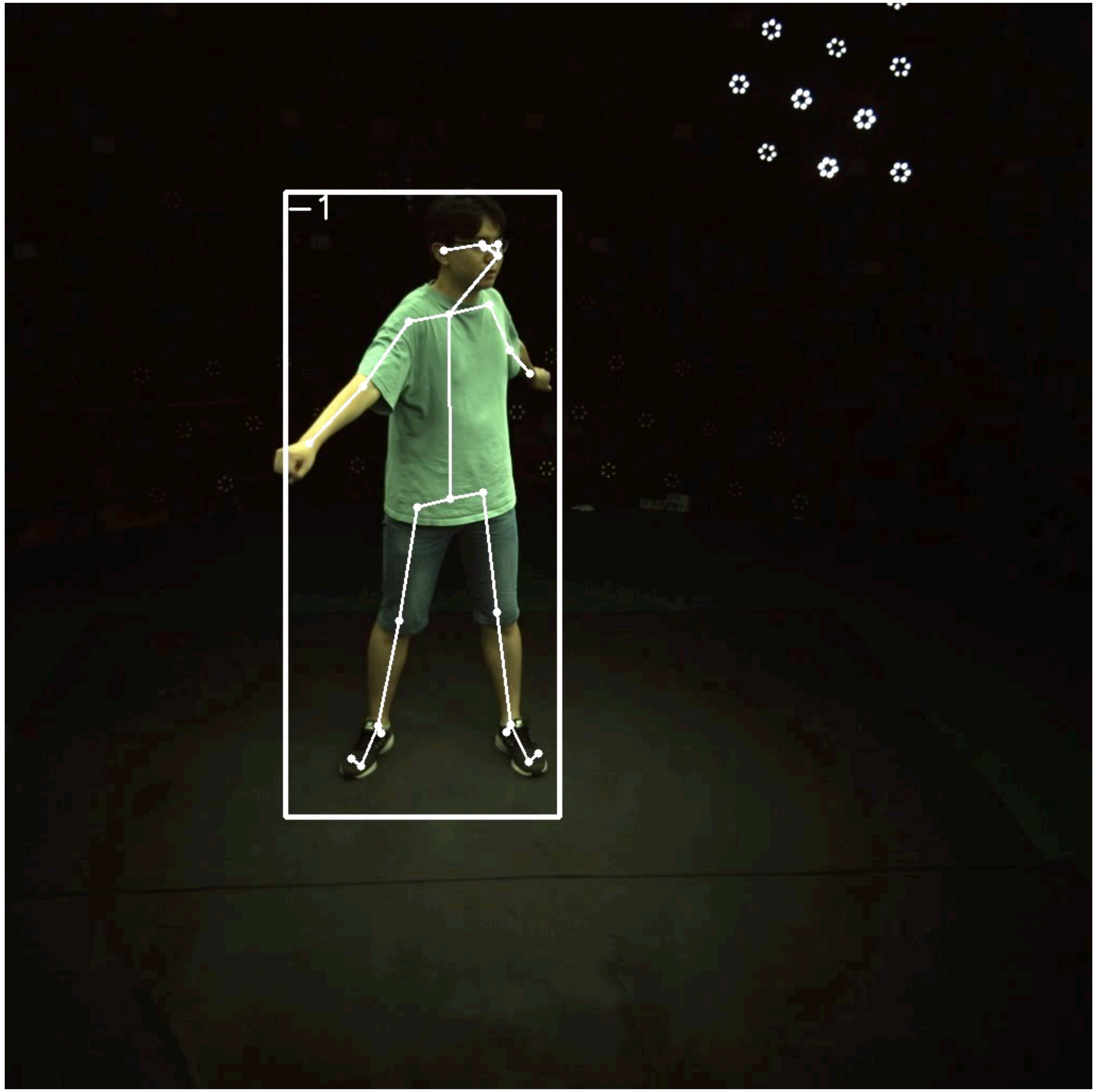


Image credit: https://skinning.org/direct-methods-slides.pdf

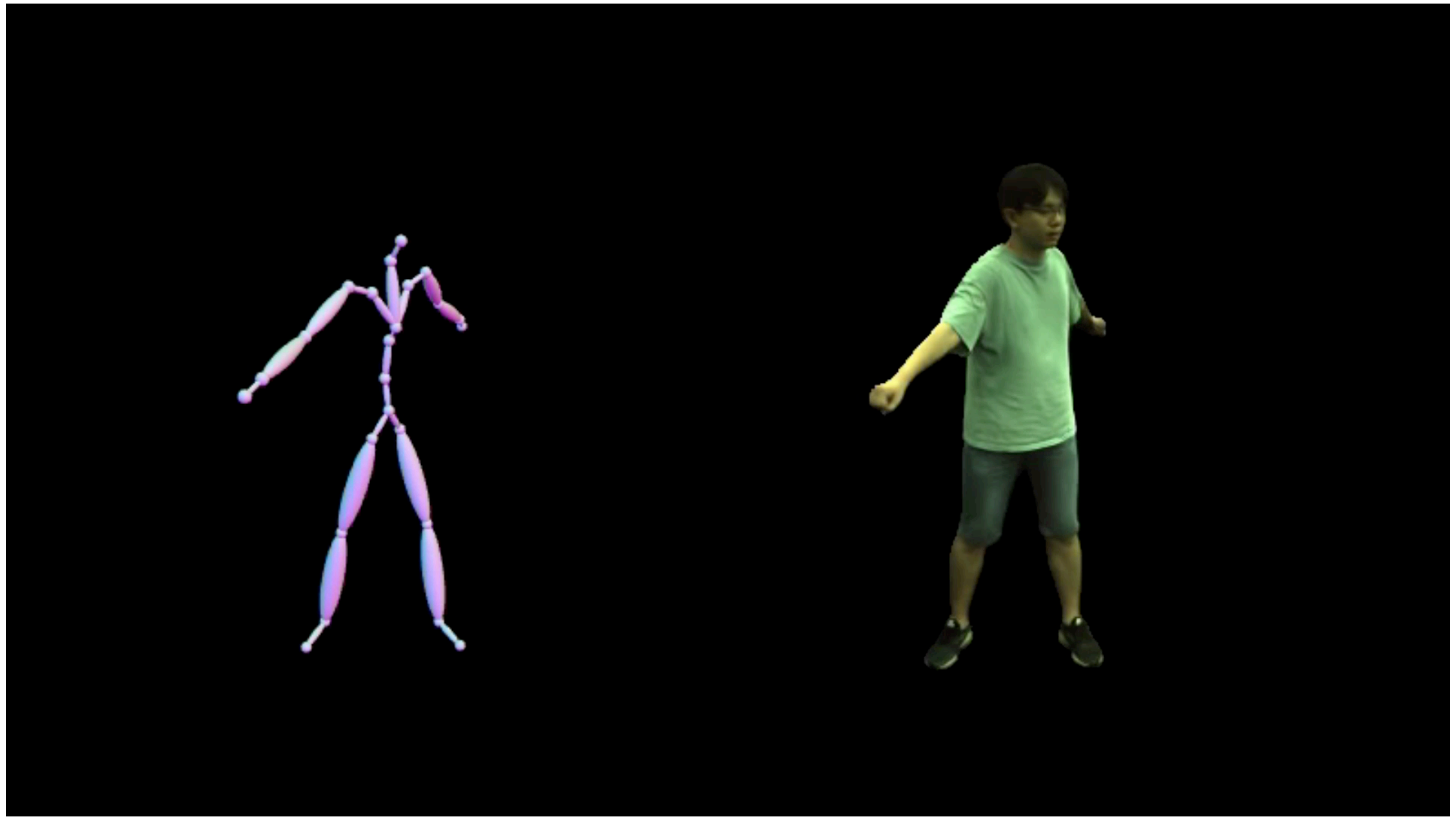
Two advantages of using LBS models

1. Human skeletons can be observed from images, and thus we only need to optimize the blend weight fields.



Two advantages of using LBS models

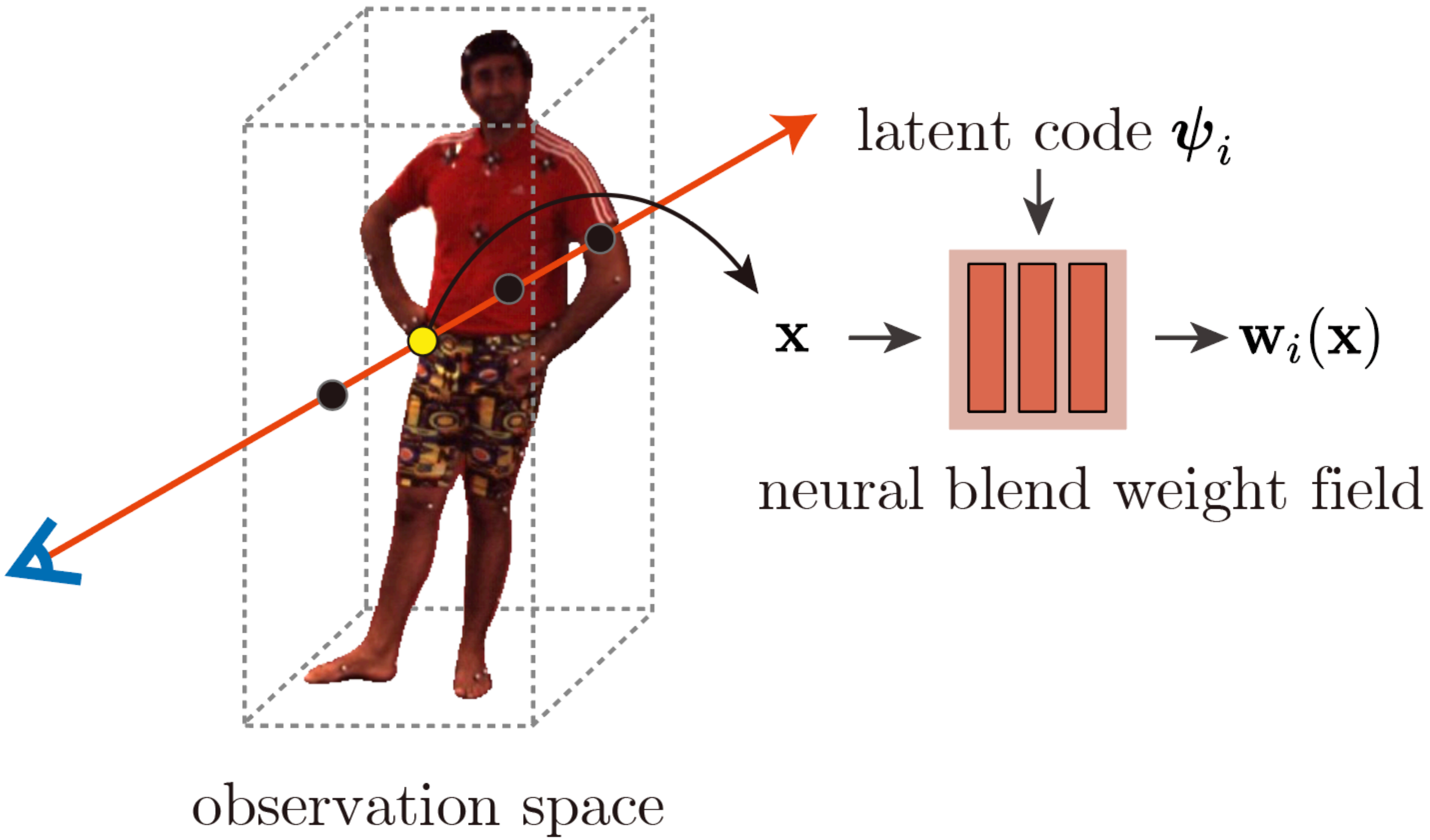
1. Human skeletons can be observed from images, and thus we only need to optimize the blend weight fields.
2. The learned blend weight fields can be combined with new human skeletons to animate human models.



Overview of the proposed pipeline

How to learn the blend weight fields

It is ill-posed to learn the blend weight fields from scratch



How to learn the blend weight fields

Given an initial blend weight, we learn a residual vector, resulting in the neural blend weight.

SMPL blend weight + Residual vector = Neural blend weight

$$\mathbf{w}_i(\mathbf{x}) = \mathrm{norm}(F_{\Delta \mathbf{w}}(\mathbf{x}, \psi_i) + \mathbf{w}^{\mathrm{s}}(\mathbf{x}, S_i))$$

Image credit: Bhatnagar, Bharat Lal, et al. NeurIPS 2020.

Learn canonical blend weights with consistency loss

We need to learn the blend weights at the canonical space for animation.

Training

$$
\begin{cases}
\text{Image loss: } L_{\text{rgb}} = \sum_{r \in \mathcal{R}} \|\tilde{\mathbf{C}}_i(\mathbf{r}) - \mathbf{C}_i(\mathbf{r})\|_2 \\
\text{Consistency loss: } L_{\text{nsf}} = \sum_{\mathbf{x} \in \mathcal{X}_i} \|\mathbf{w}_i(\mathbf{x}) - \mathbf{w}^{\text{can}}(T_i(\mathbf{x}))\|_1
\end{cases}
$$

Animation with the trained model

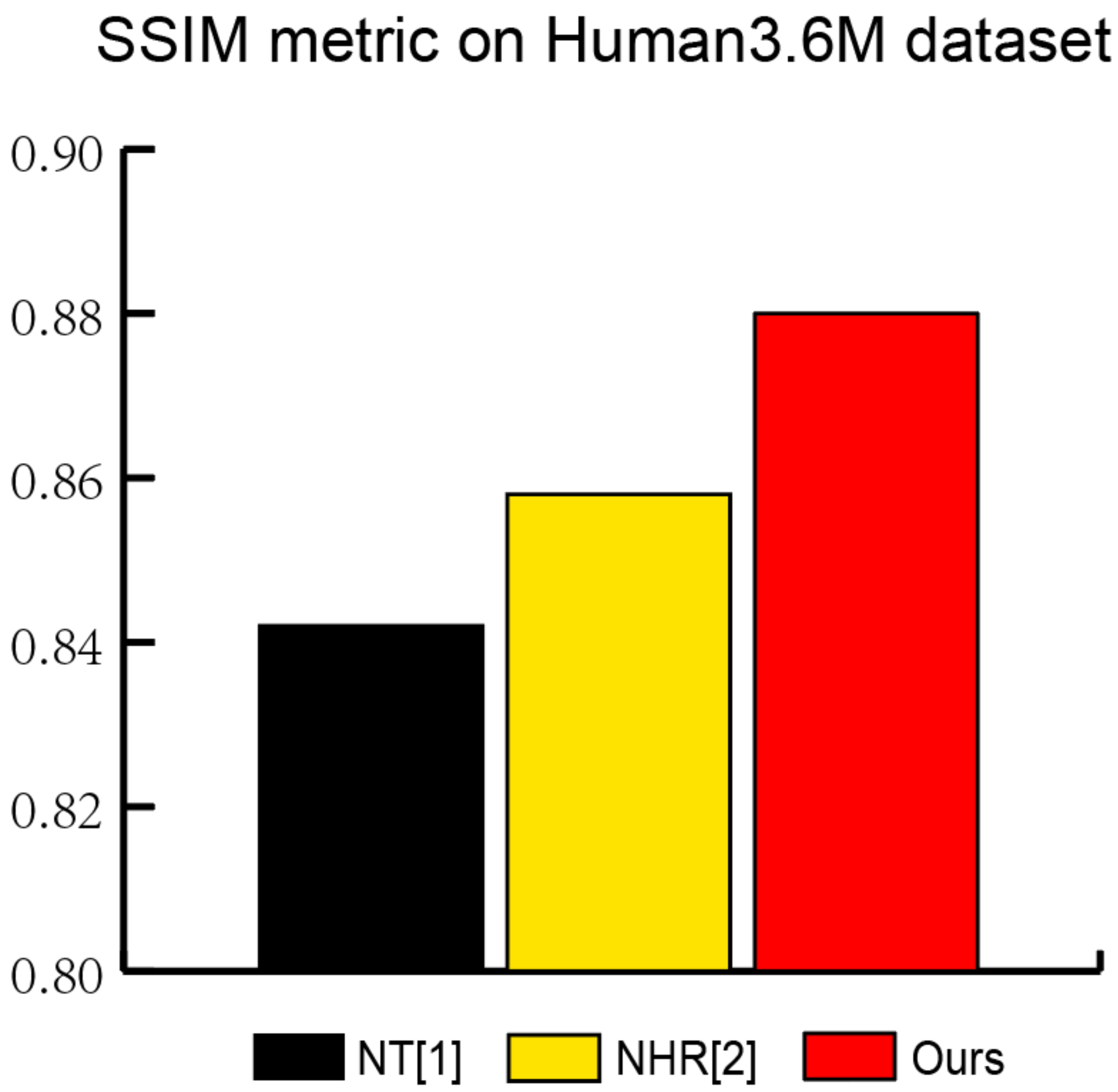
Given an unseen human pose, we need to generate the blend weights at this pose for animation.

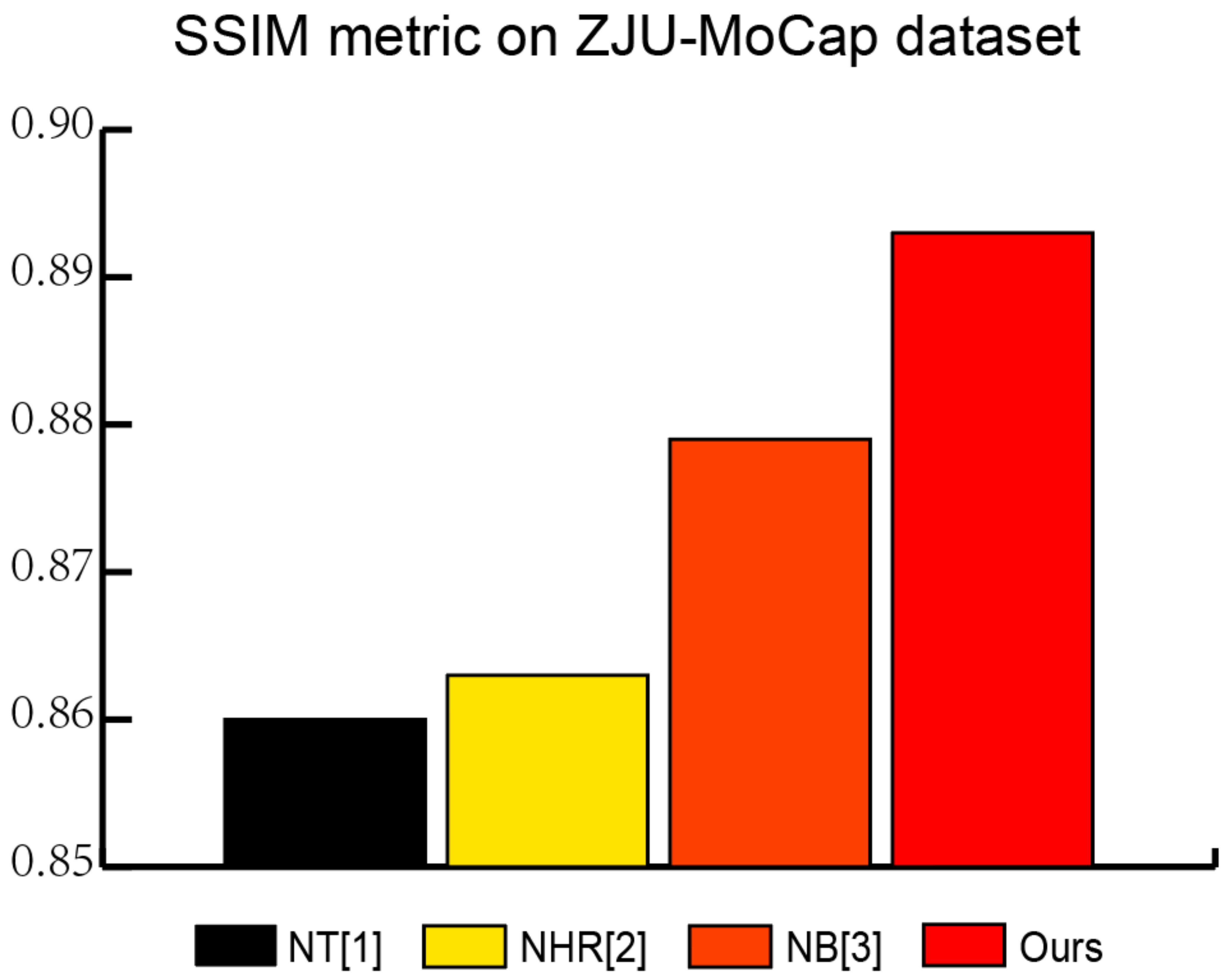
Learn blend weights under unseen human poses

The blend weights at the canonical space are used to train the blend weights under unseen human poses.

$$L_{\text{nsf}} = \sum_{\mathbf{x} \in \mathcal{X}_i} \| \mathbf{w}_i(\mathbf{x}) - \mathbf{w}^{\text{can}}(T_i(\mathbf{x})) \|_1$$

Quantitative comparison on novel pose synthesis





[1] Thies, Justus, et al. Deferred neural rendering: Image synthesis using neural textures. In ACM TOG, 2019.
[2] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.
[3] Peng, Sida, et al. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In CVPR, 2021.

Qualitative comparison on novel pose synthesis

Neural Textures [1] NHR [2] Neural Body [3] Ours

[1] Thies, Justus, et al. Deferred neural rendering: Image synthesis using neural textures. In ACM TOG, 2019.
[2] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.
[3] Peng, Sida, et al. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In CVPR, 2021.

Qualitative comparison on novel pose synthesis

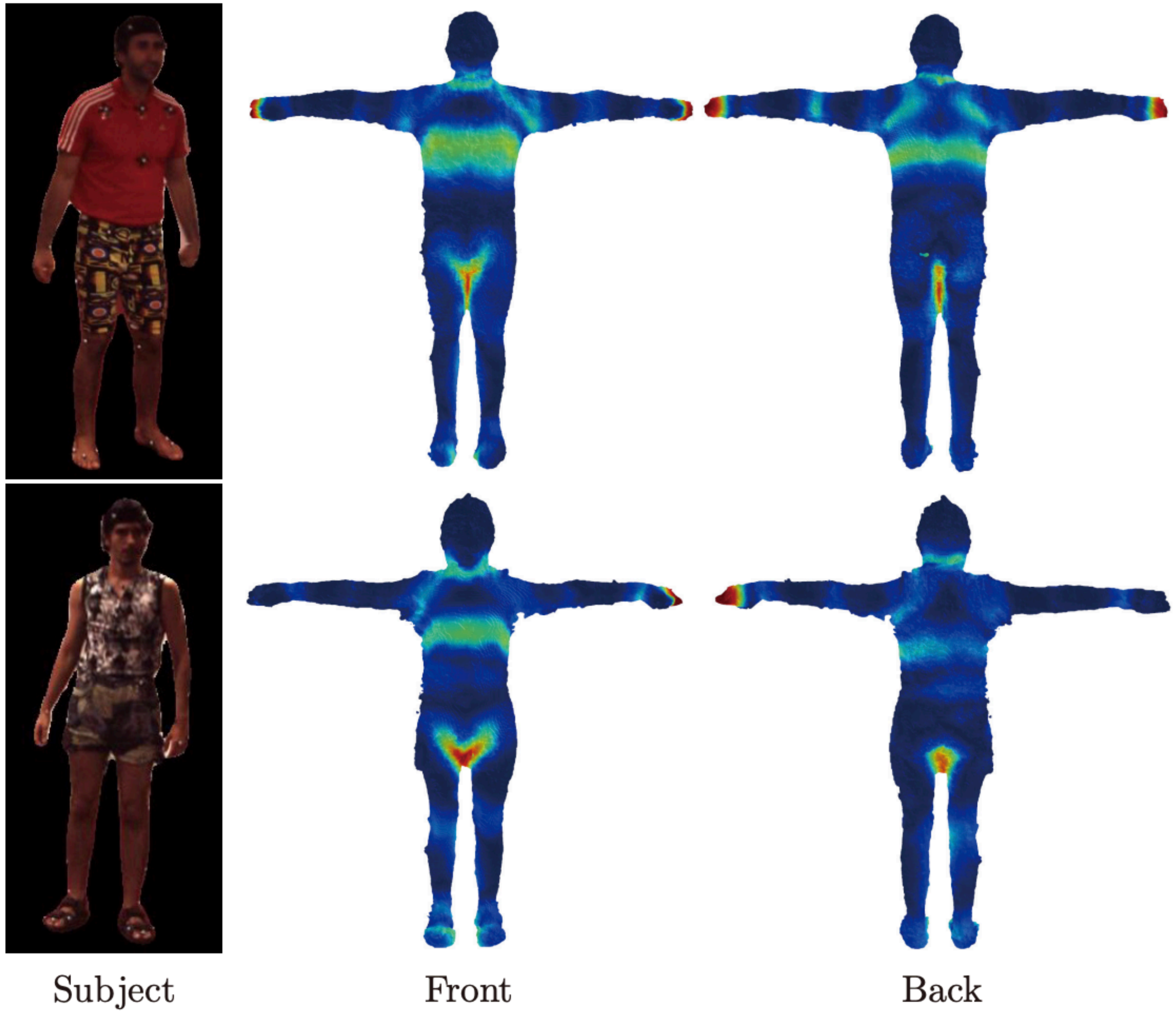
Neural Textures [1] NHR [2] Neural Body [3] Ours

[1] Thies, Justus, et al. Deferred neural rendering: Image synthesis using neural textures. In ACM TOG, 2019.
[2] Wu, Minye, et al. Multi-View Neural Human Rendering. In CVPR, 2020.
[3] Peng, Sida, et al. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In CVPR, 2021.

Ablation studies: neural blend weight field

	PSNR	SSIM
Neural blend weight field	**23.72**	**0.886**
SMPL blend weight field	21.65	0.850

Table 3: **Comparison between neural blend weight field and SMPL blend weight field** on subject "S9".



Visualization of blend weight residuals

Ablation studies: human pose accuracy

	PSNR	SSIM
Marker-based pose estimation	**23.72**	**0.886**
Marker-less pose estimation	22.27	0.858

Table 4: **Comparison between models trained with human poses** from marker-based and marker-less pose estimation methods on subject “S9”.

Ground Truth Marker-less Marker-based

Limitations

- Animatable NeRF adopts the LBS model, which can only represent articulated motions, making us difficult to handle human performers wearing loose clothes.

Limitations

- Animatable NeRF adopts the LBS model, which can only represent articulated motions, making us difficult to handle human performers wearing loose clothes.
- Animatable NeRF cannot generalize across different human subjects.

Limitations

- Animatable NeRF adopts the LBS model, which can only represent articulated motions, making us difficult to handle human performers wearing loose clothes.

- Animatable NeRF cannot generalize across different human subjects.

- The animation stage requires us to optimize neural blend weight fields for novel human poses, which is slow.

Thanks!

1. Project page: https://zju3dv.github.io/neuralbody
2. Project page: https://zju3dv.github.io/animatable_nerf

4-view video

Free-viewpoint video