A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose

NeurIPS 2021

Shih-Yang Su[1]

Frank Yu[1]

Michael Zollhöfer[2]

Helge Rhodin[1]

[1] University of British Columbia

[2] Meta Reality Labs

View on google slides for better viewing experience: https://docs.google.com/presentation/d/1ps8N8ACJ6NqJ0_T_XSa7g1vZndhe1v200-gIO7JDReM/edit?usp=sharing

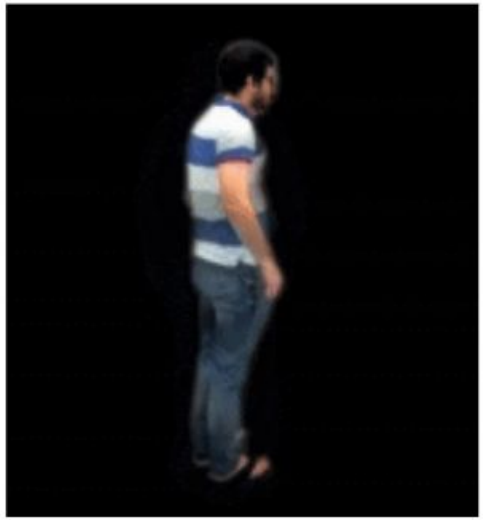
Goal: Learn an animatable human avatar

A-NeRF body models

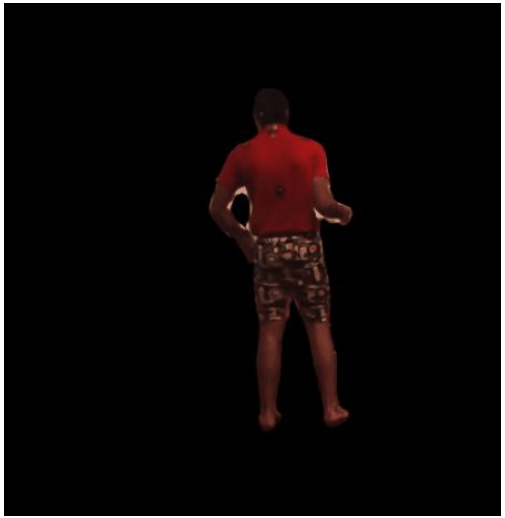
Learned from monocular images and estimated 3D poses

Setting: poses unseen during training

Related works



Textured Neural Avatars [1]
Inconsistent cross views



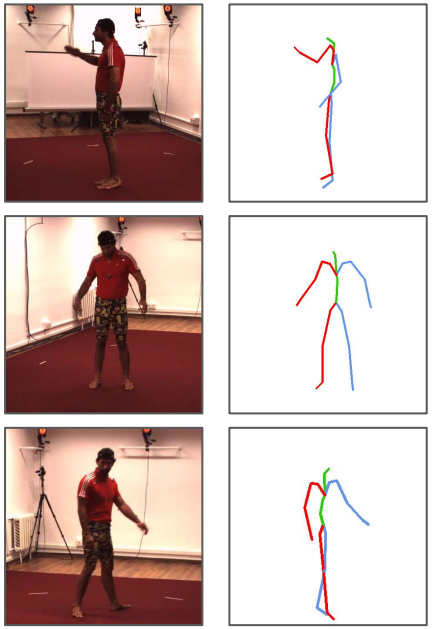
NeuralBody [2]
Needs template mesh

D-NeRF [3]
Cannot control body pose

- **Other works require two or more cameras**

[1] “**Textured Neural Avatars**”, Shysheya et al., CVPR 2019.
[2] “**Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans**”, Peng et al., CVPR 2021.
[3] “**D-NeRF: Neural Radiance Fields for Dynamic Scene**”, Pumarola et al., CVPR 2021

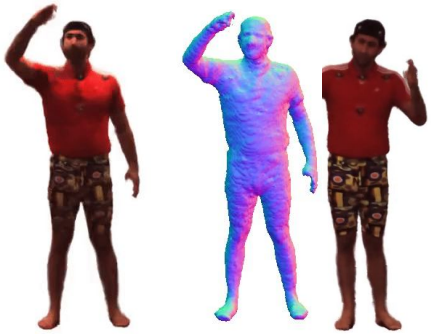
A-NeRF: Articulated Neural Radiance Fields for Animatable Human Model



Images Estimated poses

Articulated
Neural Radiance Fields

Novel views Animation



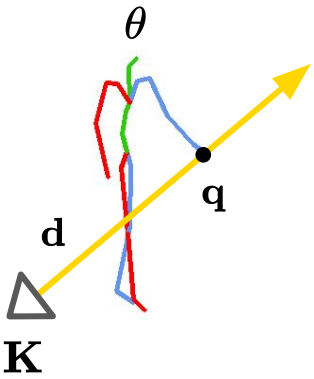
Animatable human model

No template meshes or pre-defined surface [1] needed

[1] Image credit: "**SMPL-X: A new joint 3D model of the human body, face and hands together**", Pavlakos et al., CVPR2019

A-NeRF volumetric body model

Training image



Articulated
Neural Radiance Fields
ϕ

Via volume rendering

A-NeRF joint optimizes the radiance fields and the estimated pose

Training image

$\mathbf{I}_k$

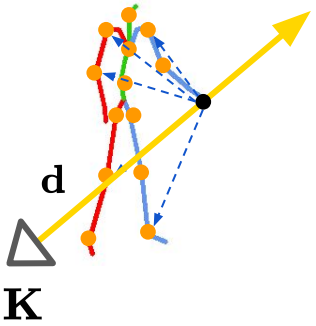
Via volume rendering

$C_\phi(\theta)$

A-NeRF volumetric body model with skeleton-relative encoding

Training image

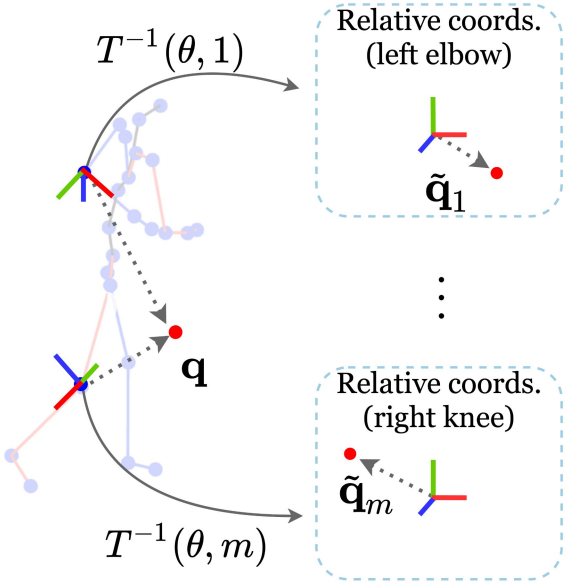
Skeleton-relative encoding



Articulated
Neural Radiance Fields
ϕ

Via volume rendering

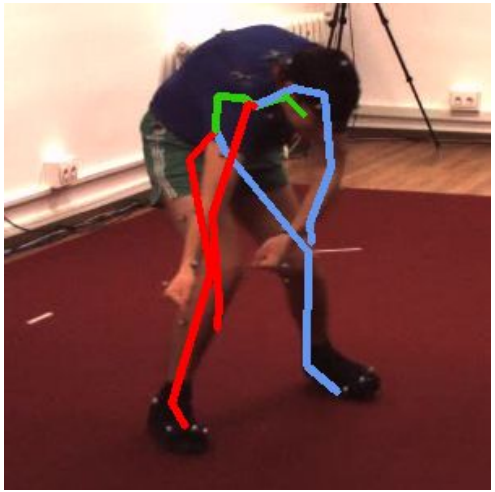
Skeleton-relative encoding



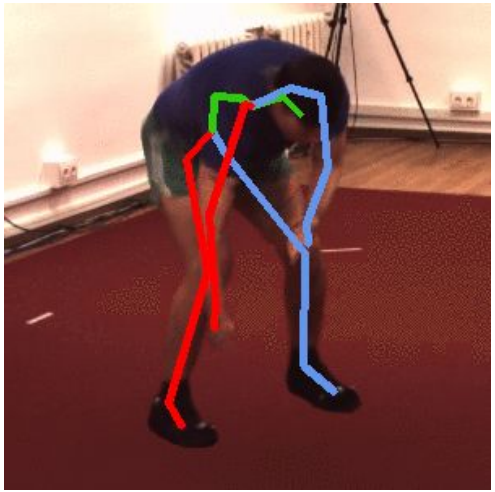
$T^{-1}(\theta, m)$: **World-to-local bone transformation**

$$\tilde{\mathbf{v}}_m = ||\tilde{\mathbf{q}}_m||_2 \quad \text{(Rel. Dist.)}$$
$$\tilde{\mathbf{q}}_m = T^{-1}(\theta, m)\mathbf{q} \quad \text{(Rel. Pos.)}$$
$$\tilde{\mathbf{r}}_m = \frac{\tilde{\mathbf{q}}_m}{||\tilde{\mathbf{q}}_m||_2} \quad \text{(Rel. Dir.)}$$
$$\tilde{\mathbf{d}}_m = \left[T^{-1}(\theta, m)\right]_{3\times 3}\mathbf{d} \quad \text{(Rel. Ray.)}$$

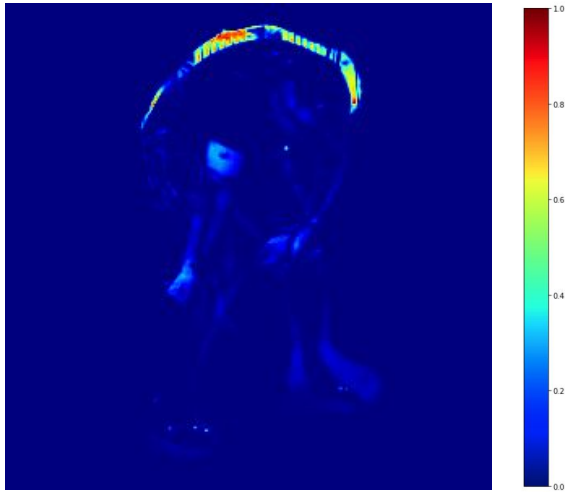
Skeleton-relative encoding enables pose refinement



Reference image
with estimated pose

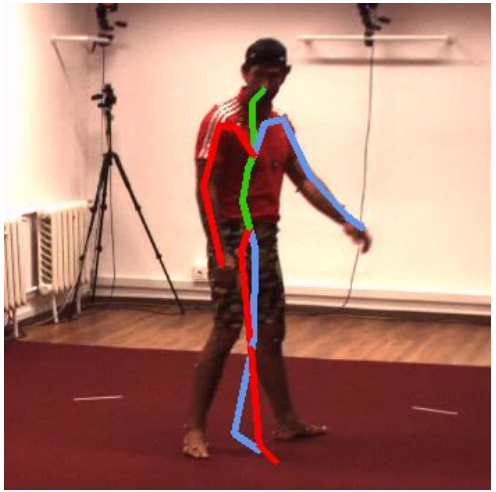


A-NeRF rendering
from initial pose to refined pose

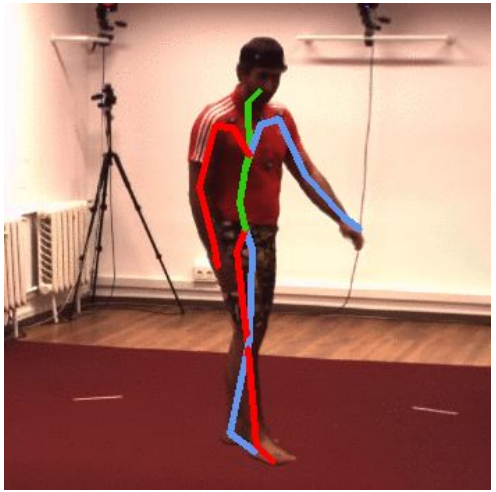


Photometric error
(L2 distance, normalized to [0, 1])

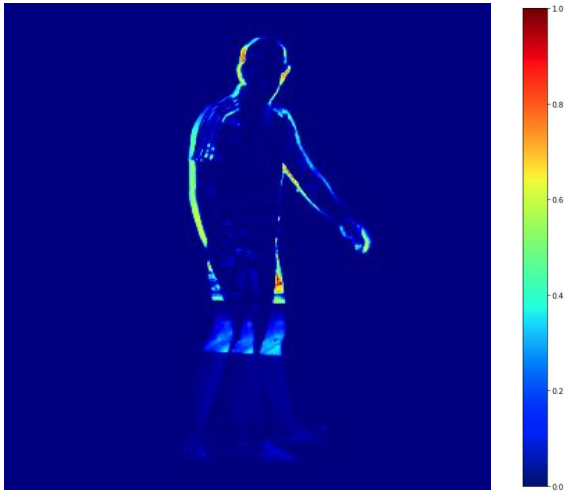
Skeleton-relative encoding enables pose refinement



Reference image
with estimated pose



A-NeRF rendering
from initial pose to refined pose

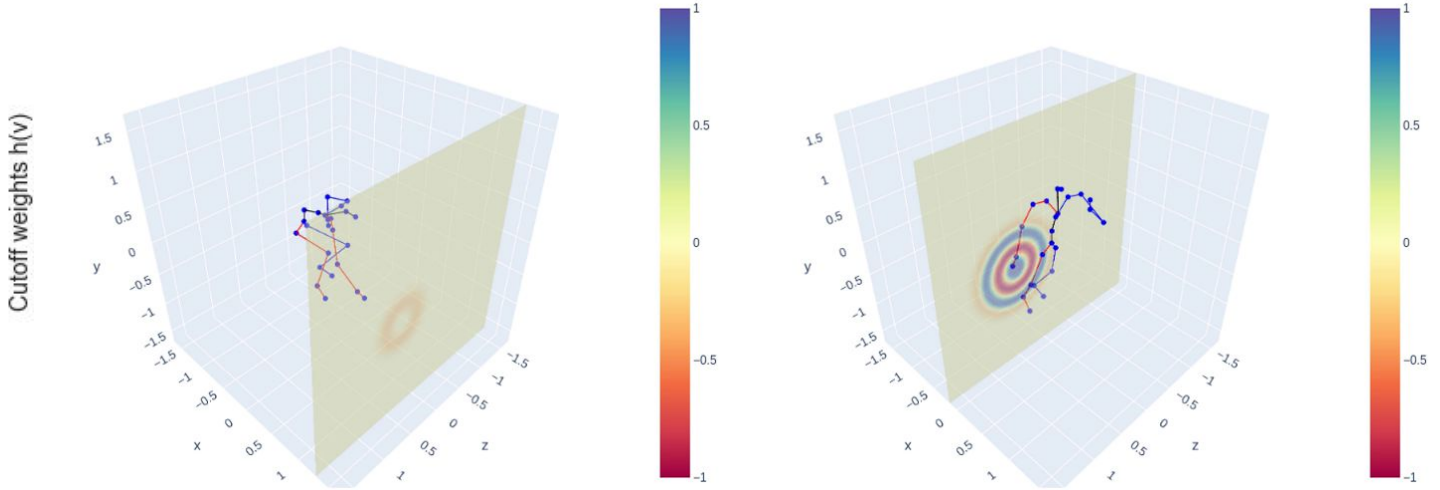


Photometric error
(L2 distance, normalized to [0, 1])

Skeleton-relative encoding with cutoff positional encoding

$$h(\tilde{\mathbf{v}}) = 1 - \text{sigmoid}(\tau \cdot (\tilde{\mathbf{v}} - t))$$

Cutoff positional encoding of **right hand** in different pose



Full skeleton-relative encoding: $\mathbf{e}(\mathbf{q}, \mathbf{d}, \theta) = \left[h(\tilde{\mathbf{v}})\Gamma(\tilde{\mathbf{v}}), \tilde{\mathbf{r}}, h(\tilde{\mathbf{v}})\Gamma(\tilde{\mathbf{d}})\right]$

Animating A-NeRF Body Models

Mixamo[1]

MonoPerfCap [2]

Human 3.6M [3]

SURREAL [4]

Setting: poses unseen during training

[1] "**Mixamo**", Adobe, https://www.mixamo.com/.
[2] "**MonoPerfCap: Human Performance Capture from Monocular Video**", Xu, TOG 2018.
[3] "**Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments**", Ionescu et al., TPAMI 2014.
[4] "**Learning from Synthetic Humans**", Varol et al., CVPR 2017

Novel view synthesis comparisons

Reference image

NeuralBody [1]

A-NeRF (Ours)

Setting: training poses, **unseen camera trajectory**

[1] “**Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans**”, Peng et al., CVPR 2021.

Novel view synthesis comparisons

Reference image

NeuralBody [1]

A-NeRF (Ours)

Setting: training poses, **unseen camera trajectory**

[1] "**Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans**", Peng et al., CVPR 2021.

A-NeRF captures plausible geometry

Reference image [1]

A-NeRF geometry

Reference image [2]

A-NeRF geometry

Setting: training poses, **unseen camera trajectory**

[1] "**MonoPerfCap: Human Performance Capture from Monocular Video**", Xu, TOG 2018.
[2] "**Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments**", Ionescu et al., TPAMI 2014.

A-NeRF captures plausible geometry

Reference image [1]

A-NeRF geometry

Reference image [1]

A-NeRF geometry

Setting: training poses, **unseen camera trajectory**

[1] “**Mixamo**”, Adobe, https://www.mixamo.com/.

A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose

NeurIPS 2021

Shih-Yang Su[1]

Frank Yu[1]

Michael Zollhöfer[2]

Helge Rhodin[1]

[1] University of British Columbia

[2] Facebook Reality Labs