RAC: Reconstructing Animatable Categories from Videos

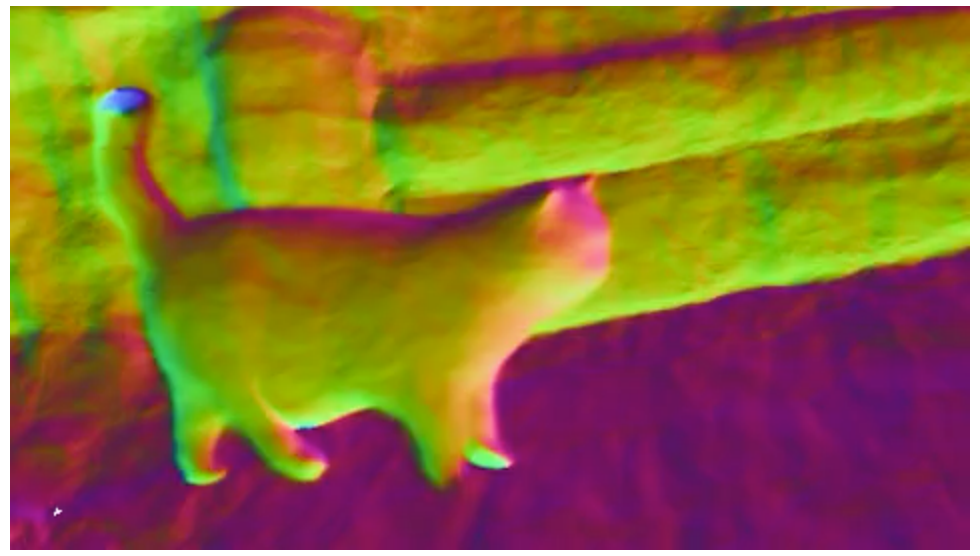
Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, Deva Ramanan

CVPR 2023



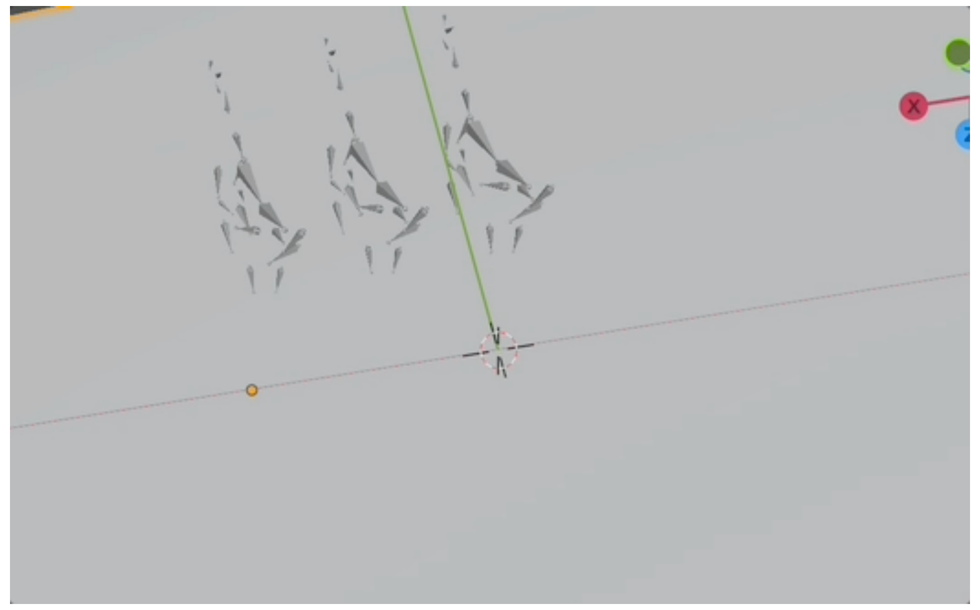

Motivation: 4D Capture from Monocular Videos

Input: videos of the ***same*** instance



Geometry (Normals)

Color, Motion



Editing

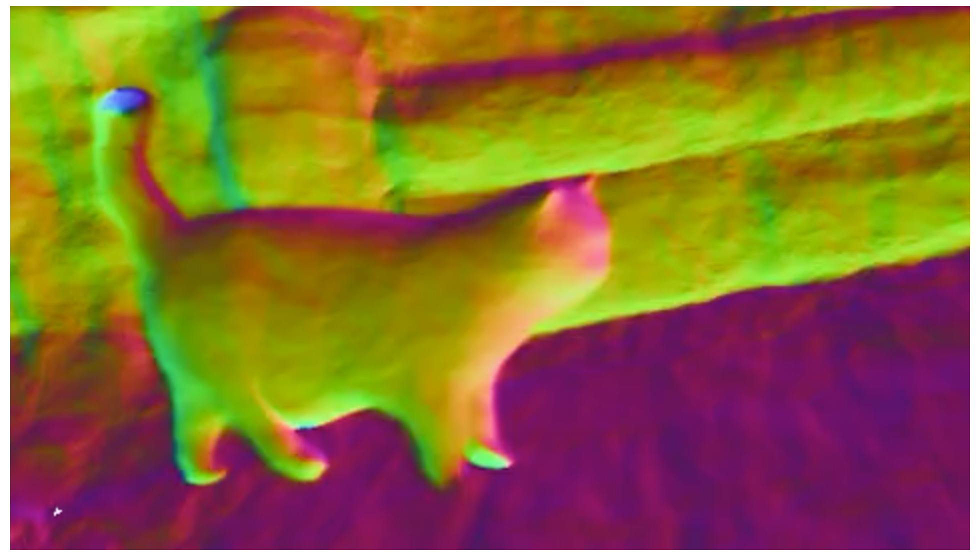
3D printing

BANMo: Building animatable 3d neural models from many casual videos. CVPR 2022.

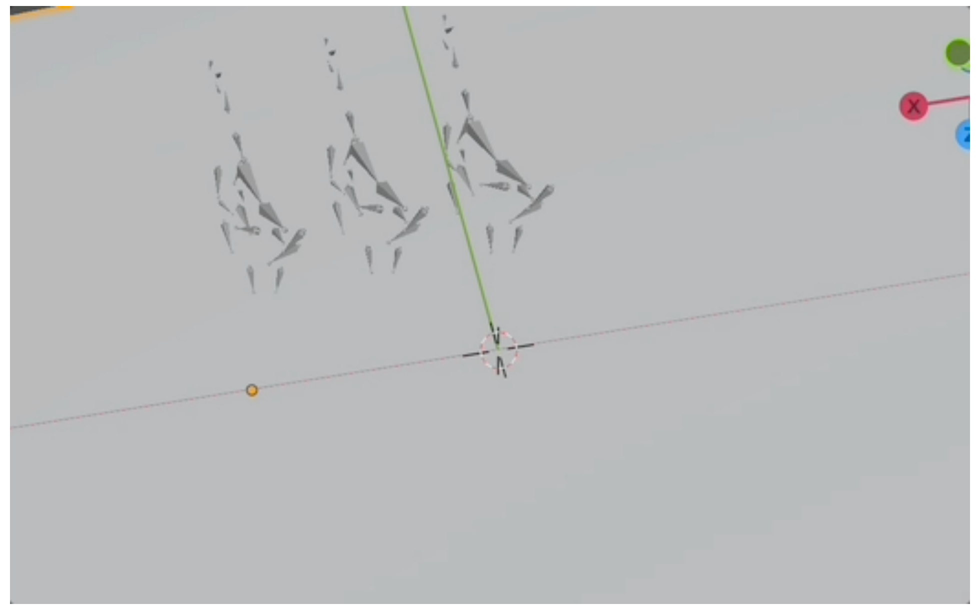
We are interested in reconstructing the dynamic 3D world using casually captured monocular videos, which enables applications such as novel view synthesis, scene editing and asset creation.

Motivation: 4D Capture from Monocular Videos

Input: videos of the ***same*** instance



Geometry (Normals)



Editing

Color, Motion

3D printing

BANMo: Building animatable 3d neural models from many casual videos. CVPR 2022.

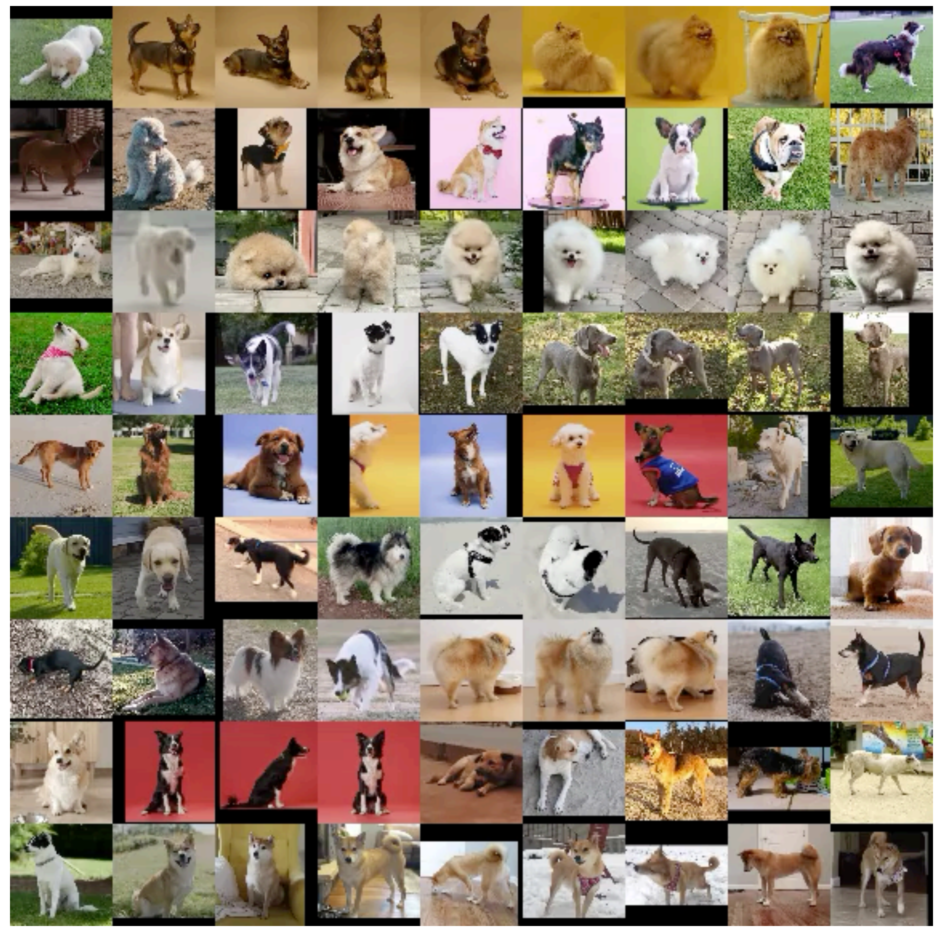
Prior works have shown nice results on building articulated body models, assuming sufficient view coverage given many videos of single instances.

Challenge: Limited Views

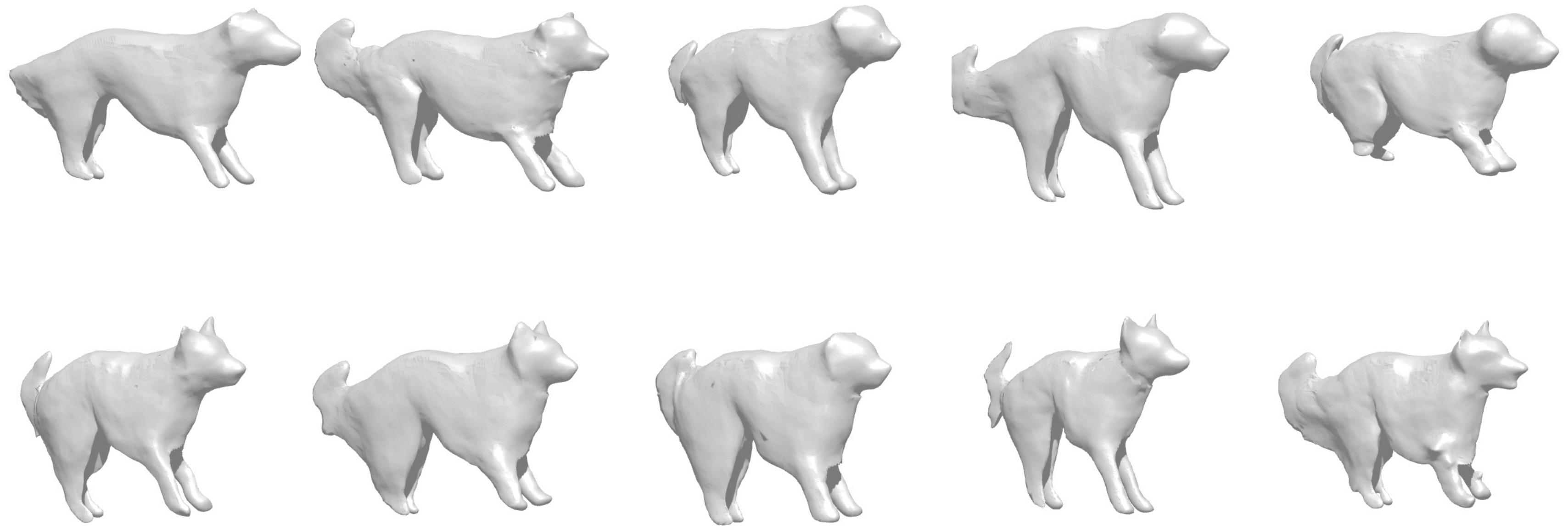
Invisible surface appears distorted

Reconstructing instances with limited view coverage is a challenge. In this example, the reconstruction looks convincing from the reference viewpoint, but the shape and deformation appear squashy from a novel viewpoint, due to lack of constraints.

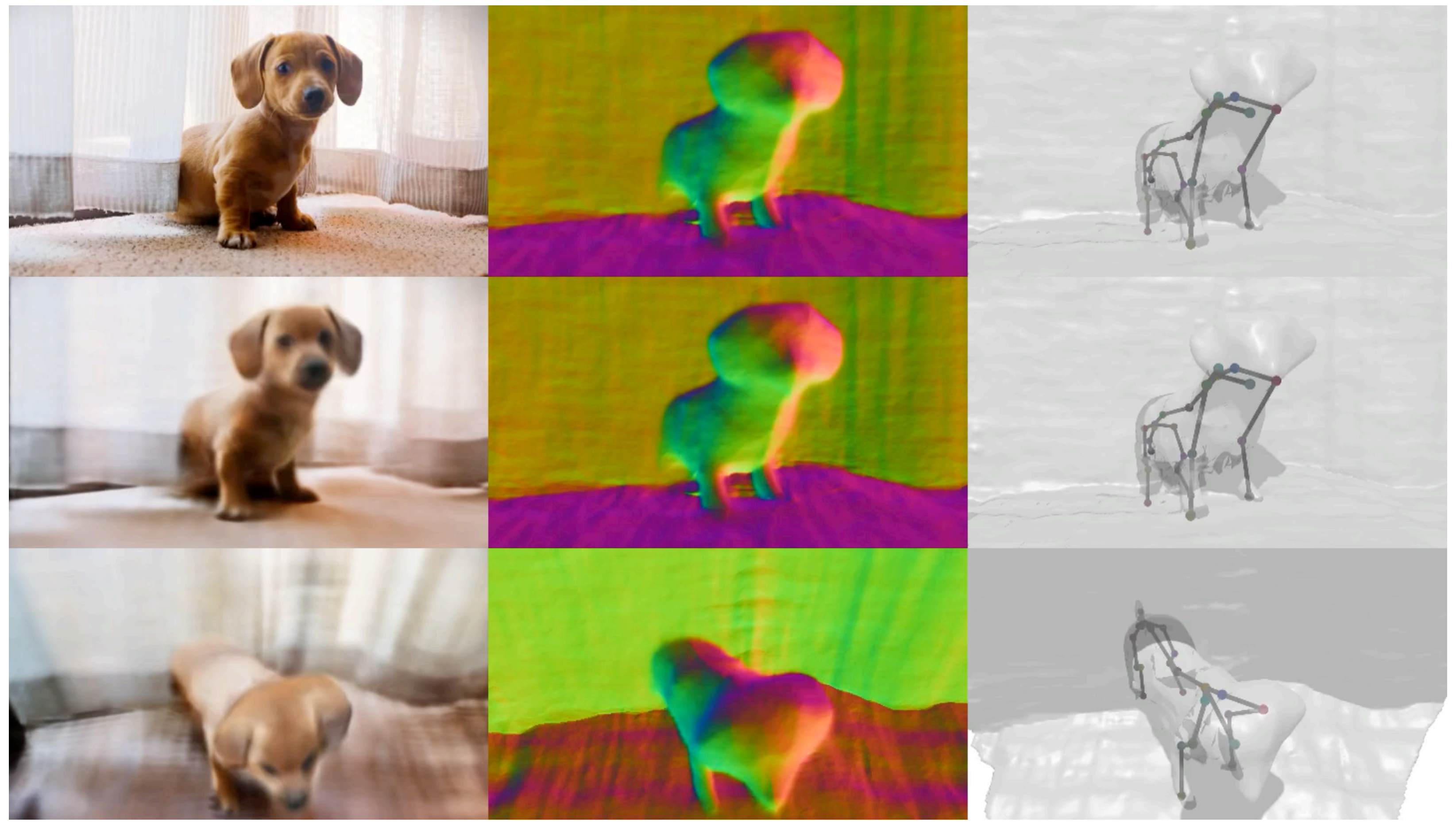
Our Solution: Build Body Prior from Videos



Internet videos



Self-supervised 3D prior

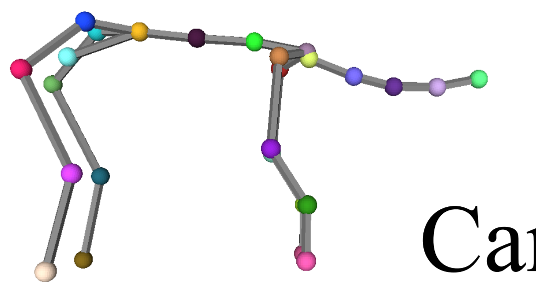


Reconstruction and novel views

In this work, we learn shape and motion model over a category from RGB videos. Such prior are useful for reconstructing instances with limited viewpoints.

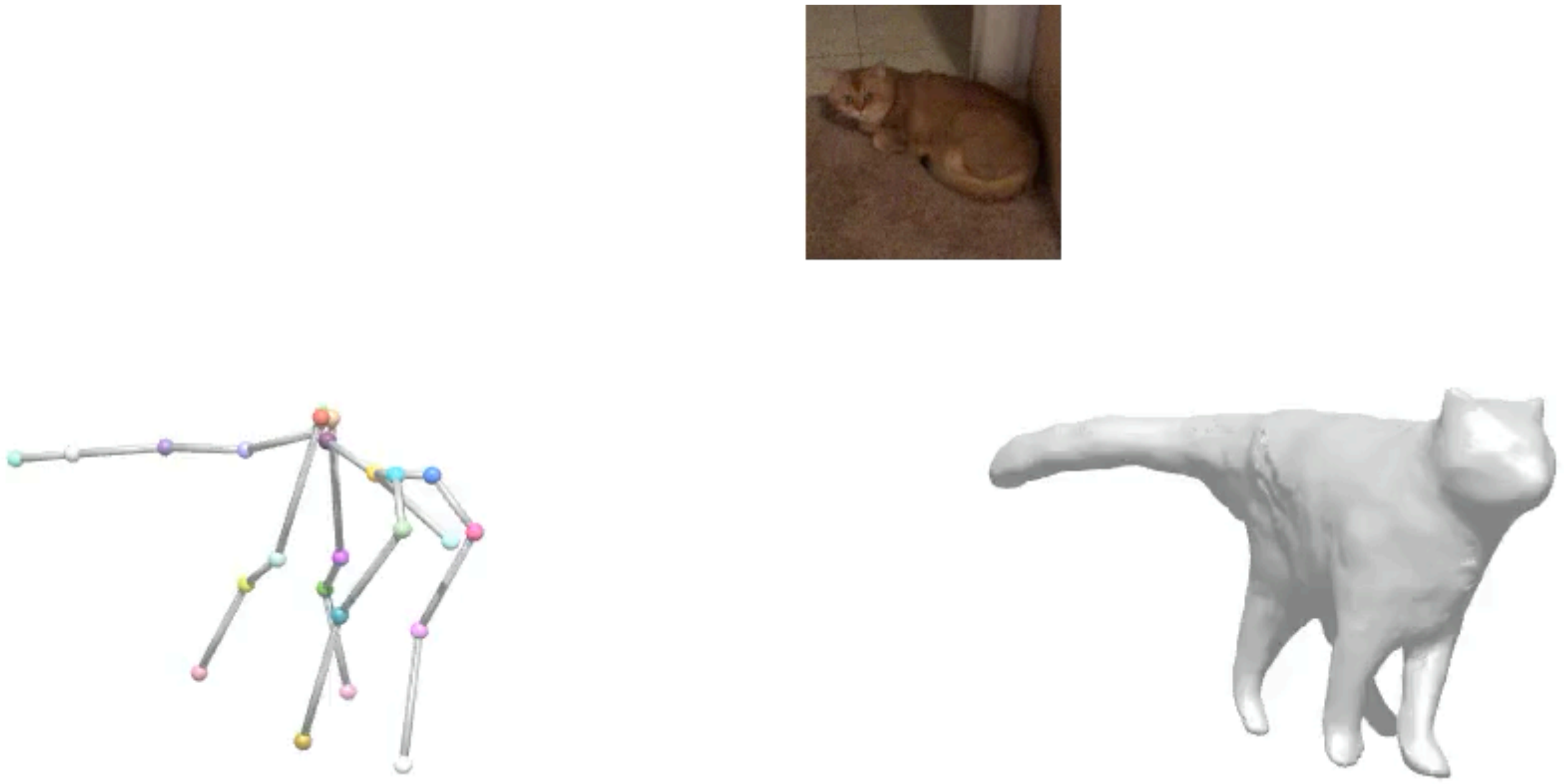
Category Reconstruction from Videos Collections

Problem setup: Given many (~100) videos of a deformable object category, can we build an animatable 3D model?



Canonical Skeleton

Differentiable Rendering



Animatable Cat Model

Given many videos of a deformable object category, for instance, 100 cat videos scraped from the internet, our goal is build an animatable 3D model that faithfully represents their shape and motion.

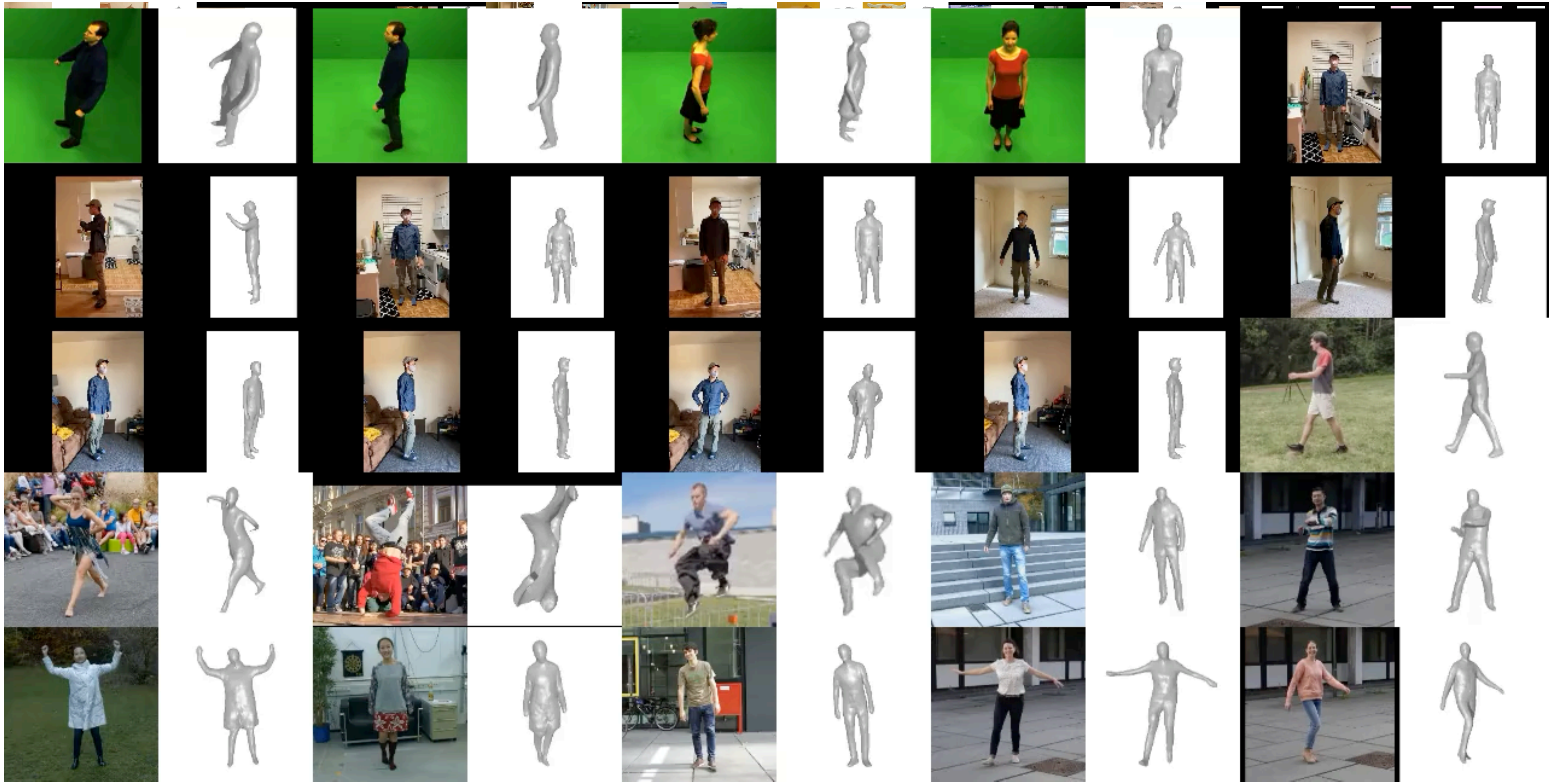
Preview of Results: Reconstruction

We visualize the videos and reconstructed shapes side by side

Preview of Results: Reconstruction

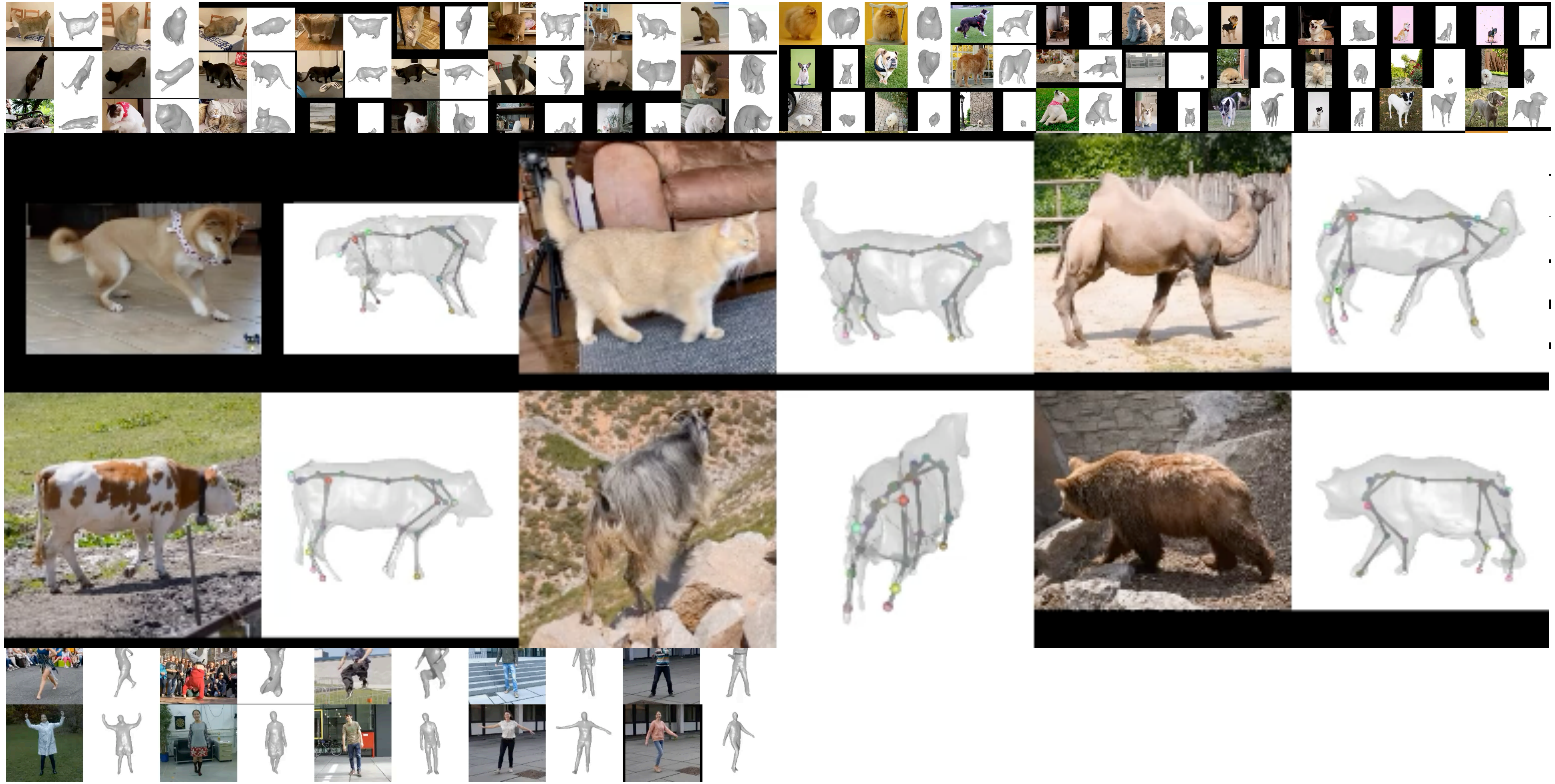
Besides cats, our method also applies to dogs, human, and a generic quadruped category

Preview of Results: Reconstruction



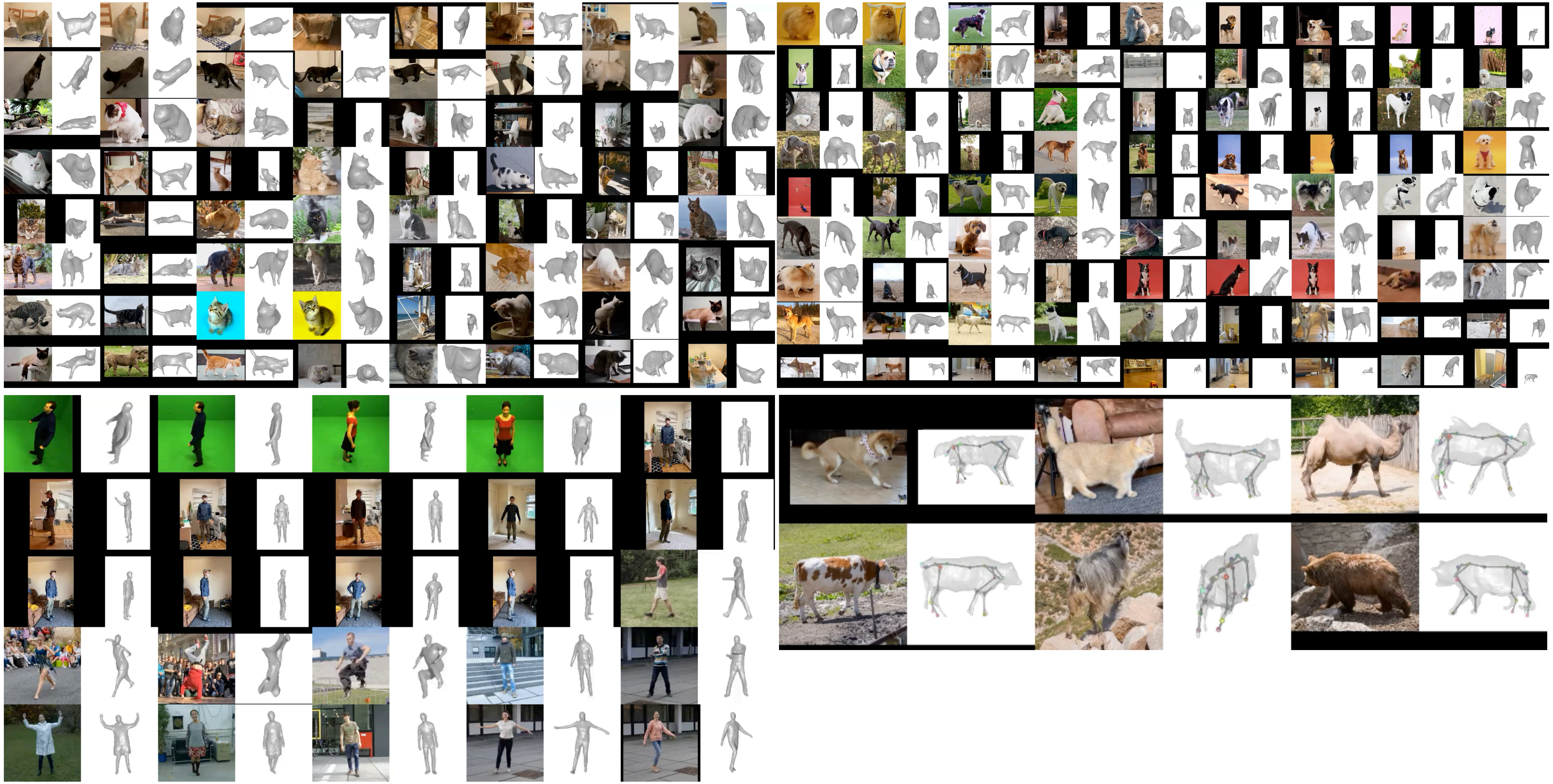
Besides cats, our method also applies to dogs, human, and a generic quadruped category

Preview of Results: Reconstruction



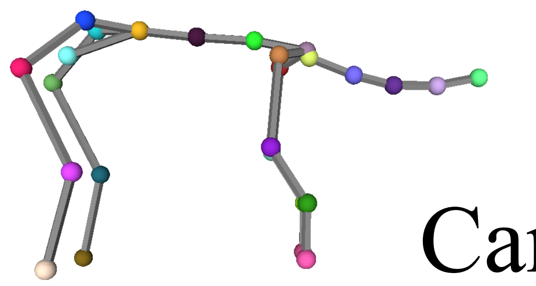
Besides cats, our method also applies to dogs, human, and a generic quadruped category

Preview of Results: Reconstruction



Besides cats, our method also applies to dogs, human, and a generic quadruped category

Preview of Results: Interpolation and Retargeting

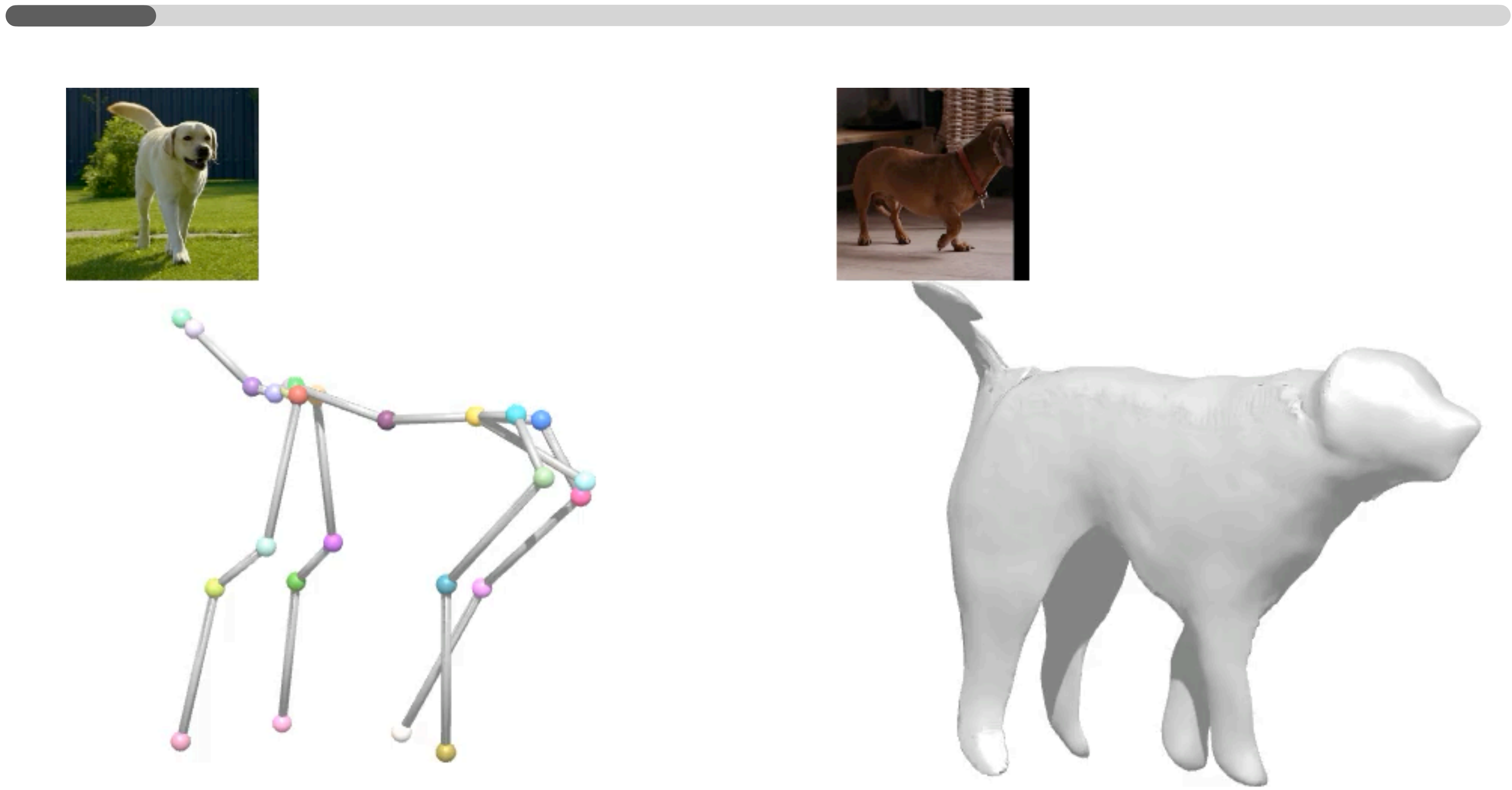


Canonical Skeleton

Monocular Videos

Differentiable Rendering

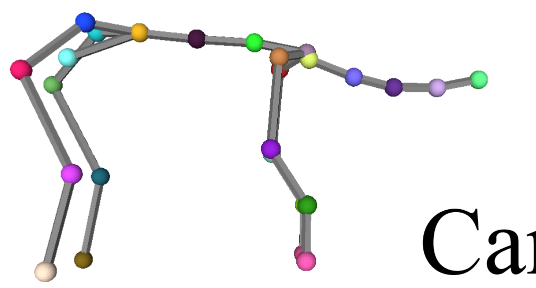
Morphology Code



Animatable Category Model

The reconstructed model allows for shape and skeleton interpolation between two instances.

Preview of Results: Interpolation and Retargeting

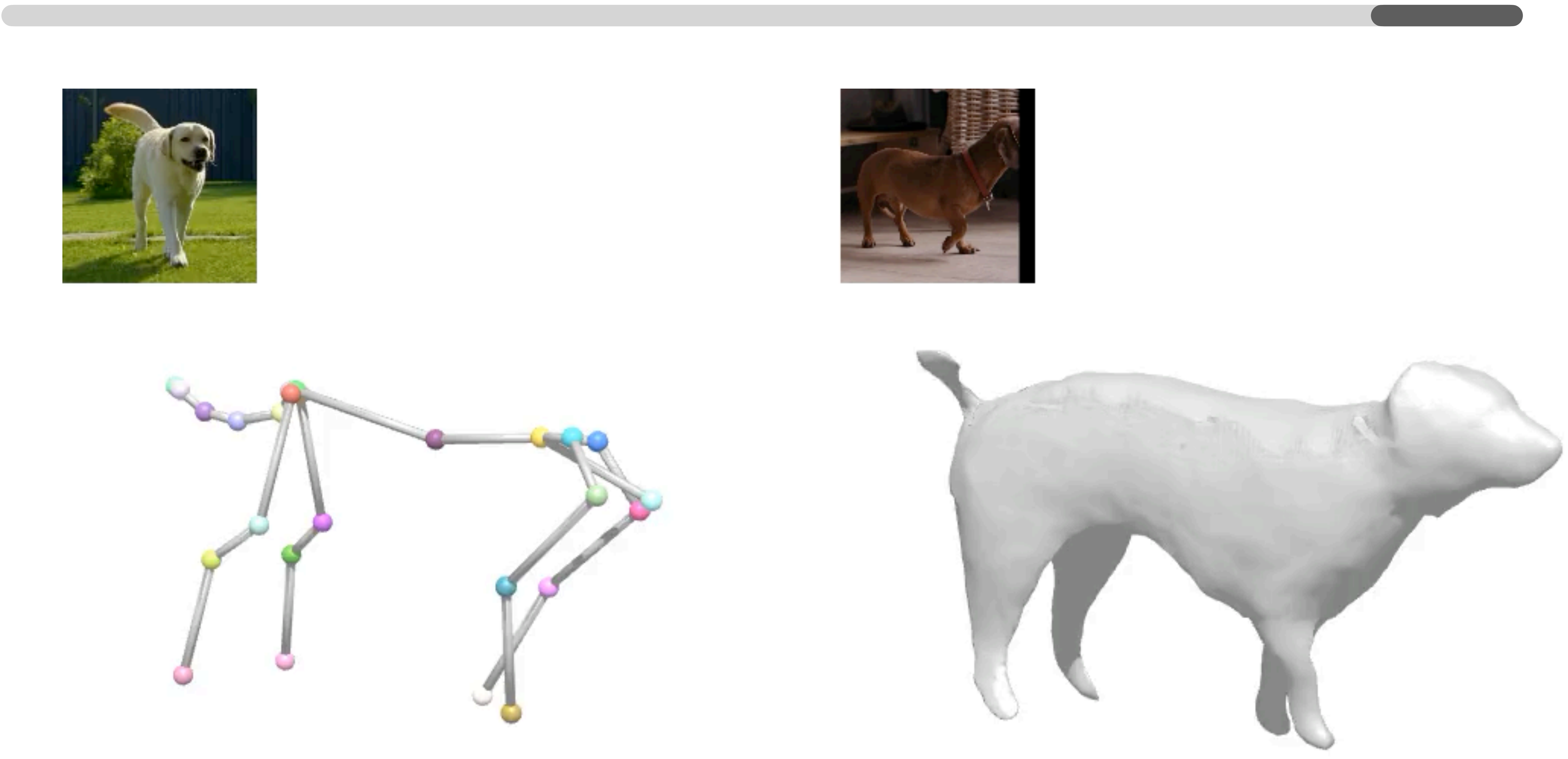


Canonical Skeleton

Monocular Videos

Differentiable Rendering

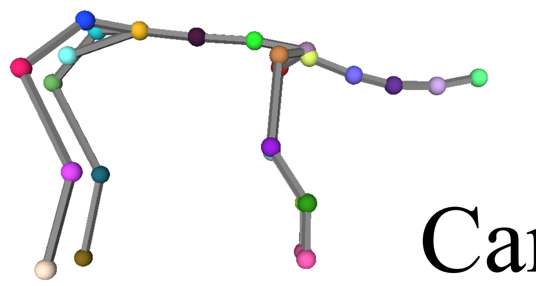
Morphology Code



Animatable Category Model

We can also transfer the motion from one instance to another.

Preview of Results: Interpolation and Retargeting

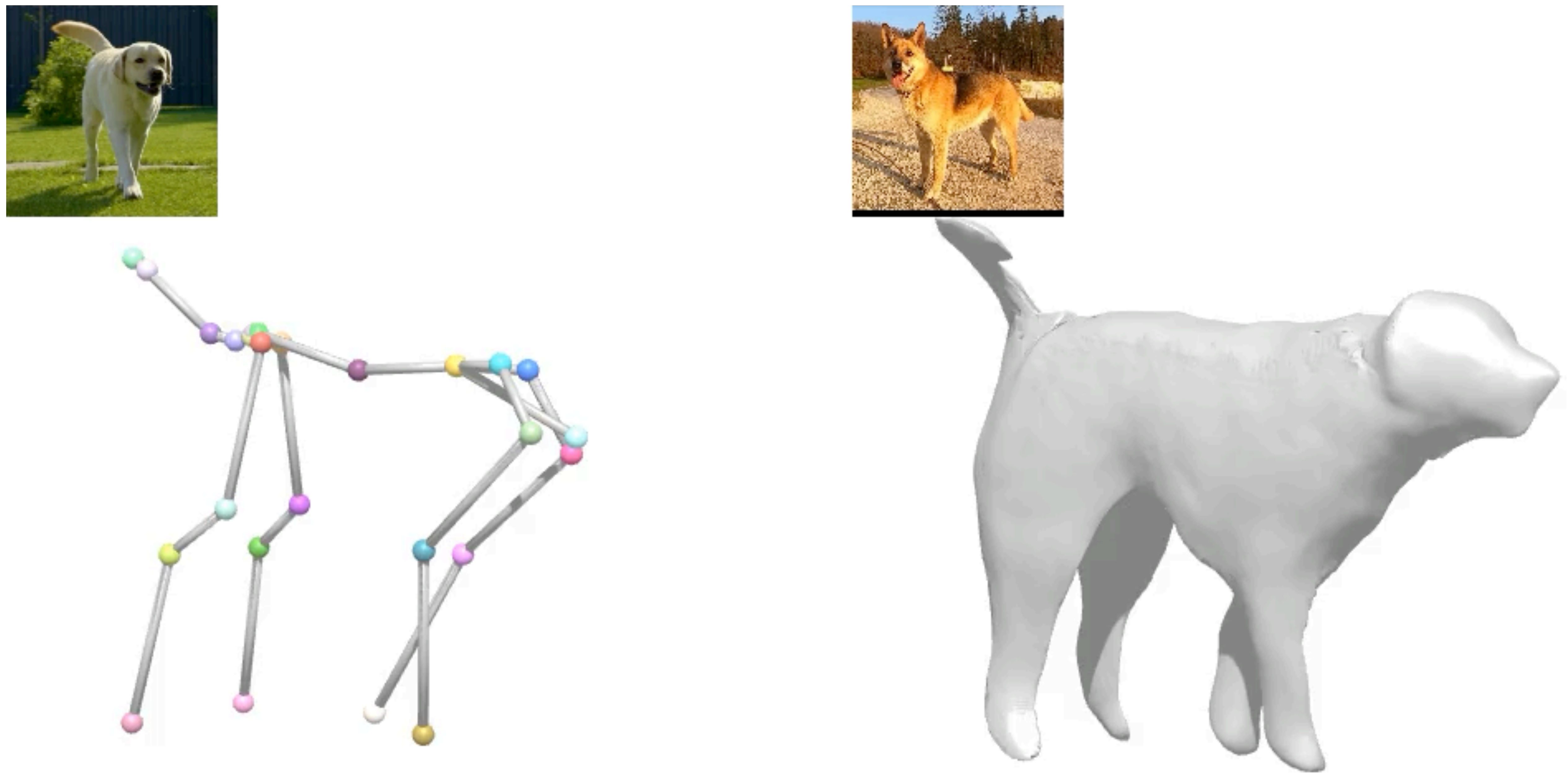


Canonical Skeleton

Monocular Videos

Differentiable Rendering

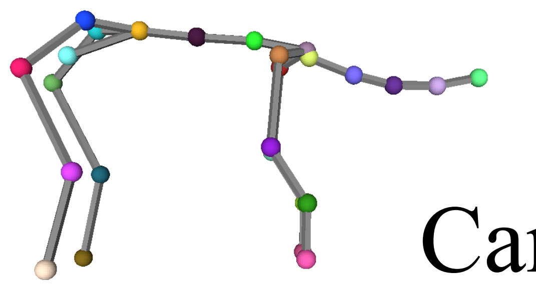
Morphology Code



Animatable Category Model

Here’s another example of interpolation and retargeting.

Preview of Results: Interpolation and Retargeting

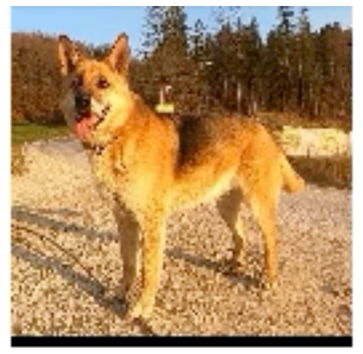


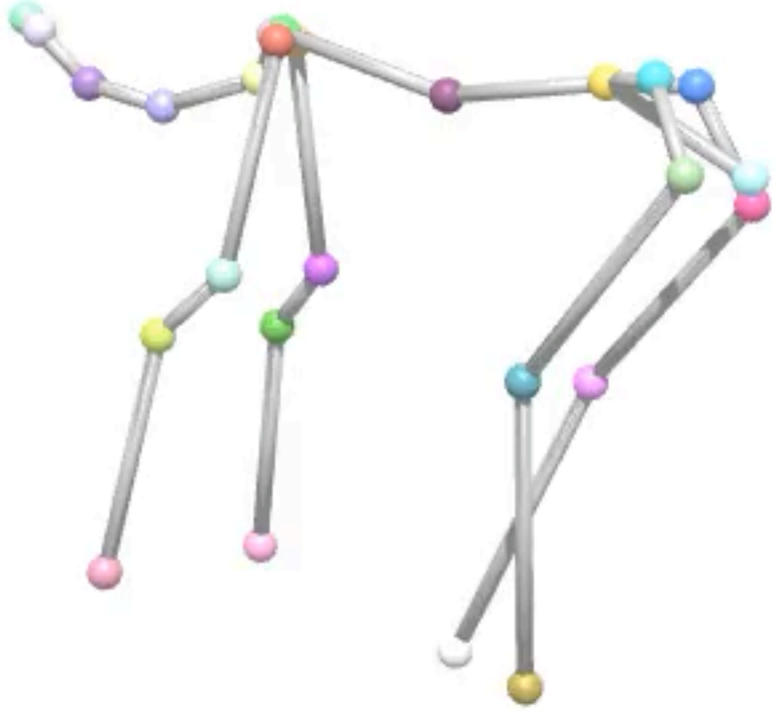
Canonical Skeleton

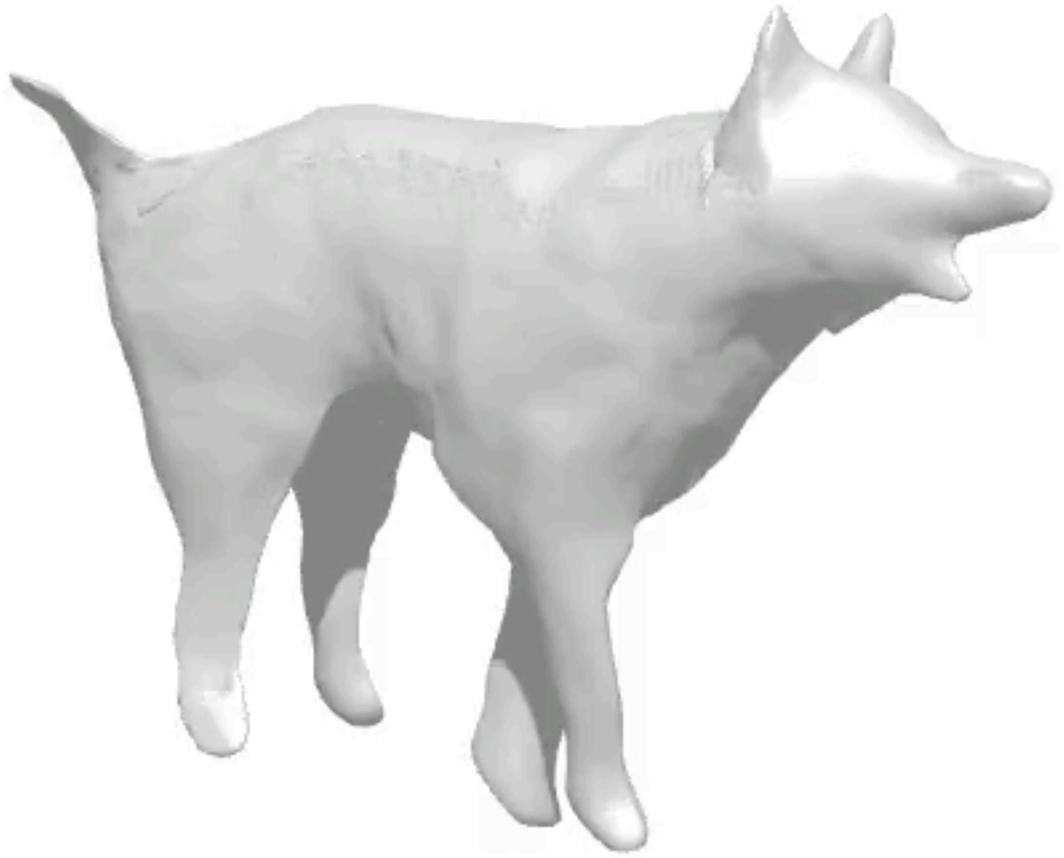
Monocular Videos

Differentiable Rendering

Morphology Code



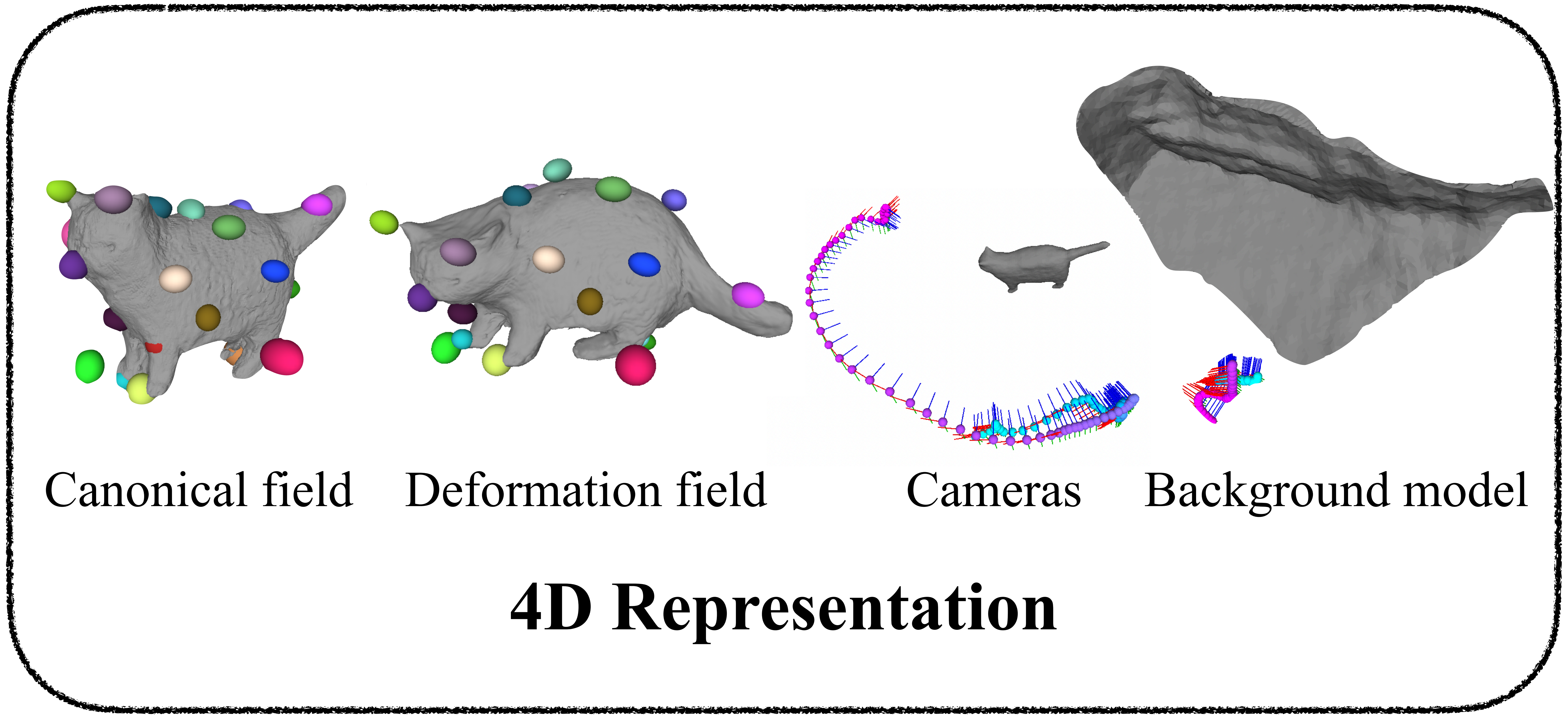




Animatable Category Model

Here’s another example of interpolation and retargeting.

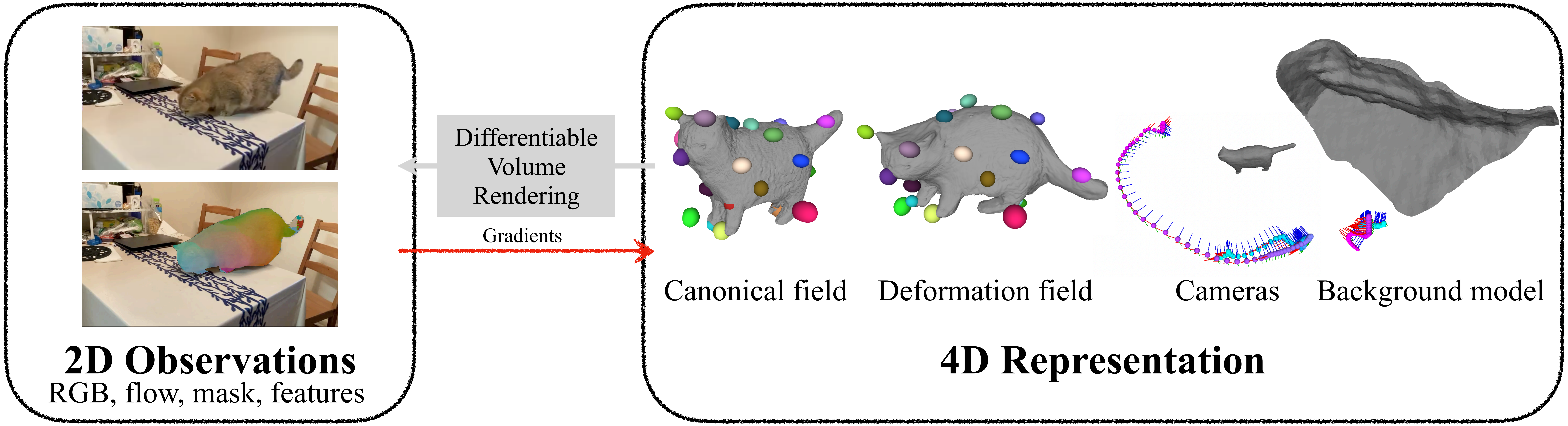
Method: Analysis-by-Synthesis



BANMo: Building animatable 3d neural models from many casual videos. CVPR 2022.

Similar to BANMo, we represent the deformable 3D object as a canonical shape, deformation fields, and camera poses. We additionally use a background neural field to explain the non-object component.

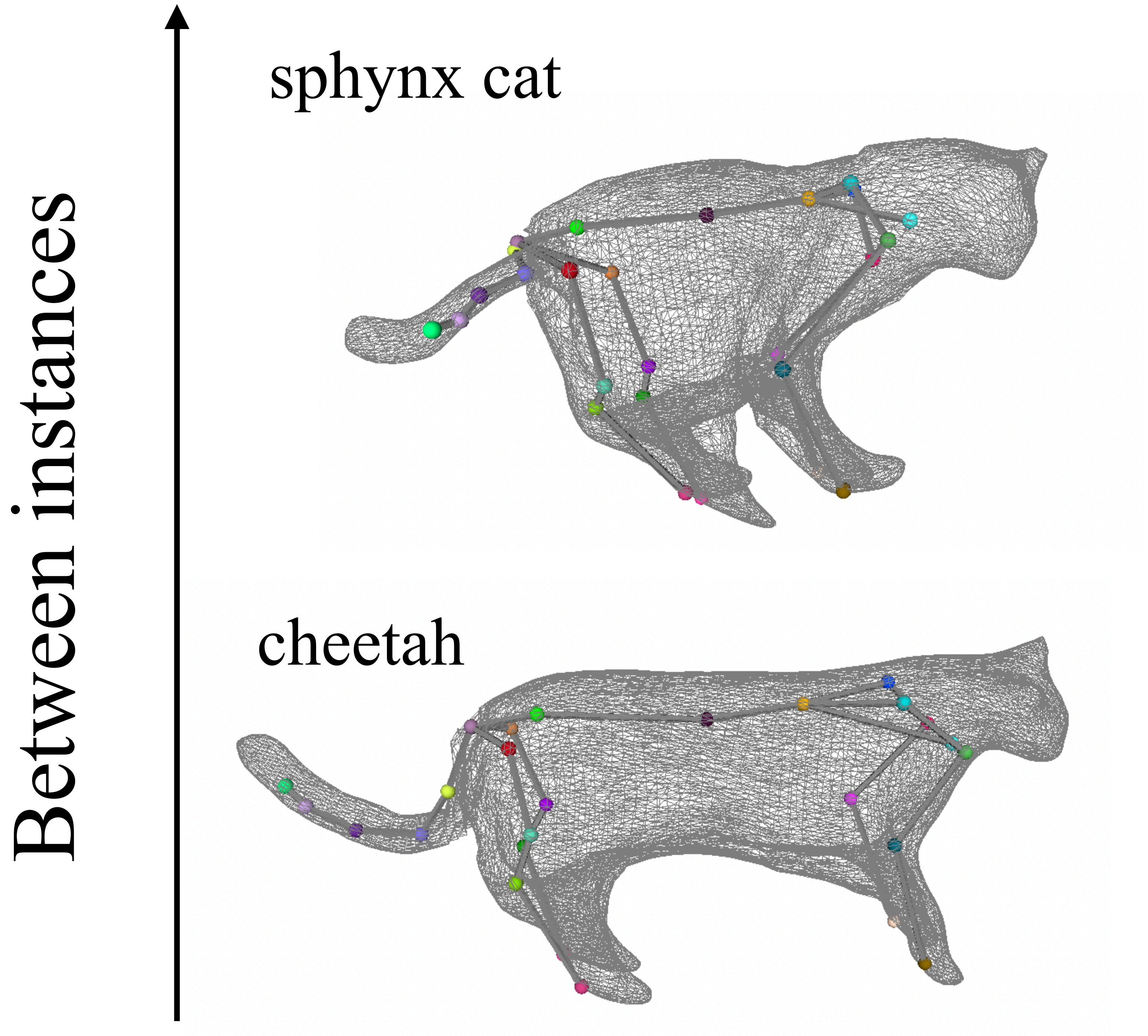
Method: Analysis-by-Synthesis



BANMo: Building animatable 3d neural models from many casual videos. CVPR 2022.

During optimization, we render the model parameters w/ a differentiable engine and minimizes the difference between the rendered images and the observed ones. Besides color, we found it helps convergence to reconstruct the other image properties, such as optical flow, segmentation, and pixel features obtained from pre-traeind models.

Disentangling *Morphology* versus *Motion*

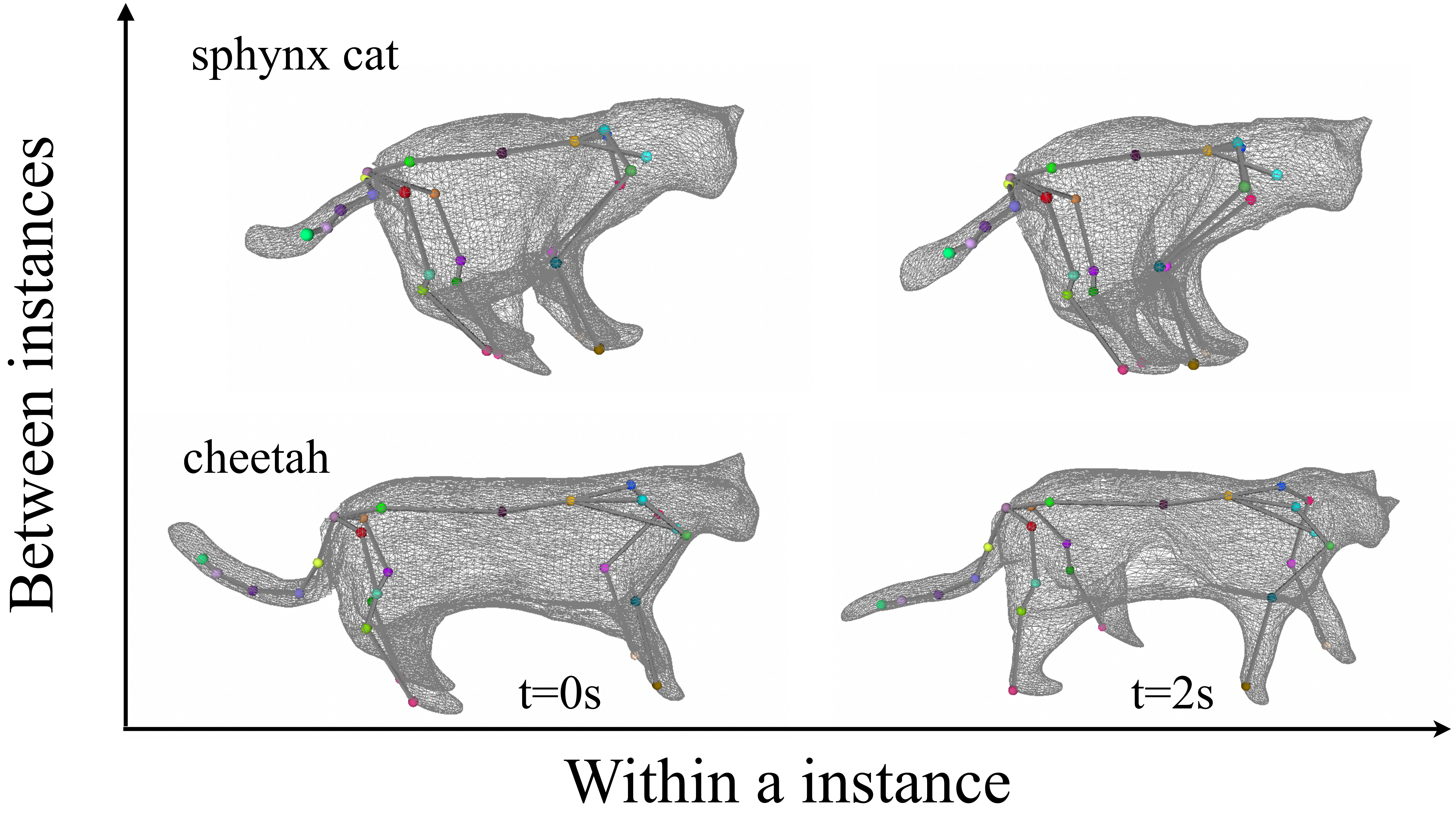


Between-Instance Variations: Morphology

including: bone dimension, body shape and appearance

Here, one key challenge is how to disentangle the variations in the input videos. We notice there are two types of variations in the input videos. One is variation between instances. For example, cheetah has long limbs and round ears, but sphinx cat has shorter limbs and pointed ears.

Disentangling *Morphology* versus *Motion*



Between-Instance Variations: Morphology

including: bone dimension, body shape and appearance

Within-Instance Variations: Motion

Including: skeleton articulations, soft deformations

Another type of variation is motion, which includes changes over time, such as bone articulation and soft tissue deformations.

Modeling Morphology

Morphology code *[HyperNeRF, 2022]*

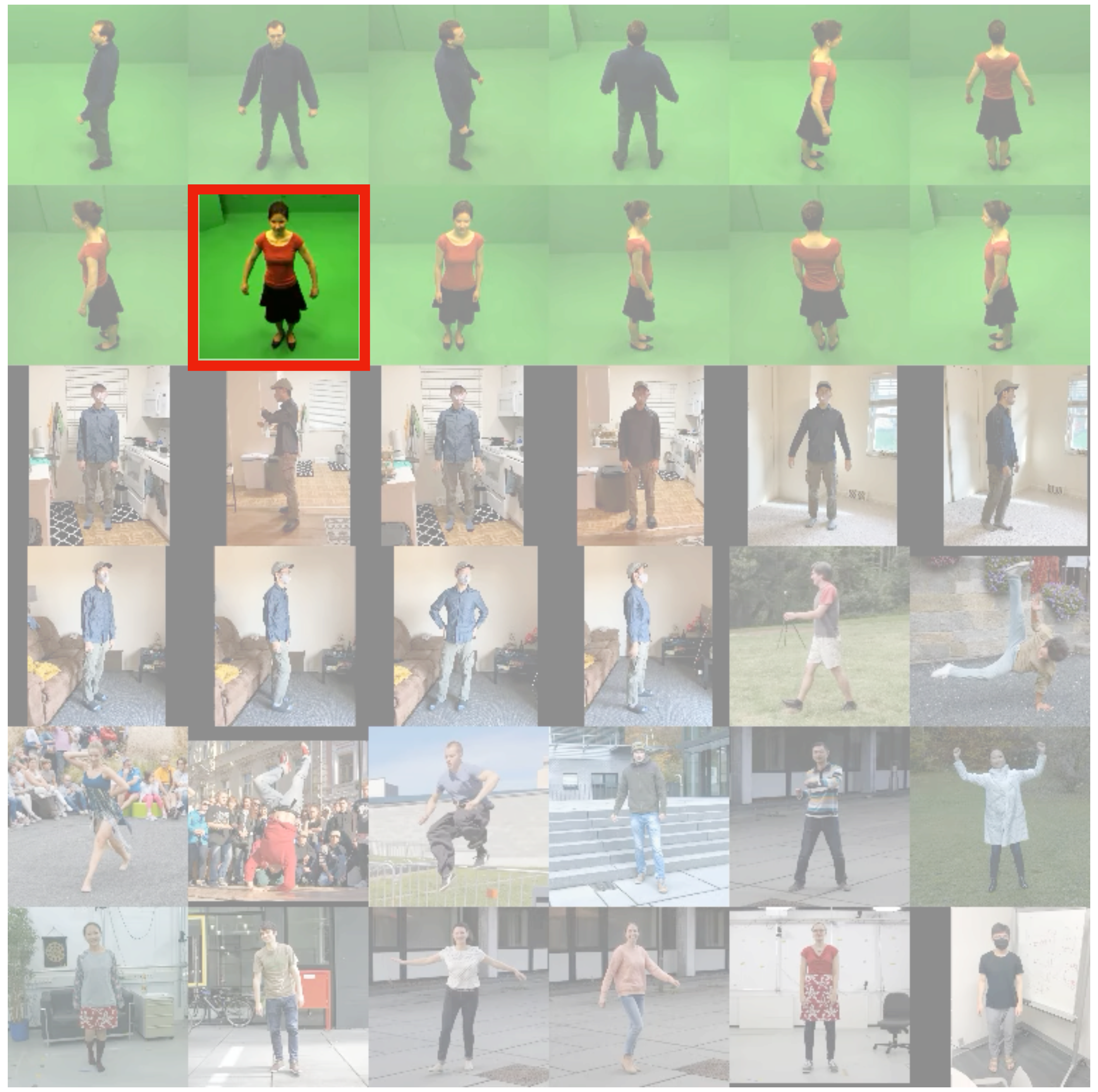
$$\sigma = \phi(\mathbf{x}, \boldsymbol{\beta}_i \in \mathbb{R}^{32})$$

Morphology: Between-Instance Variations

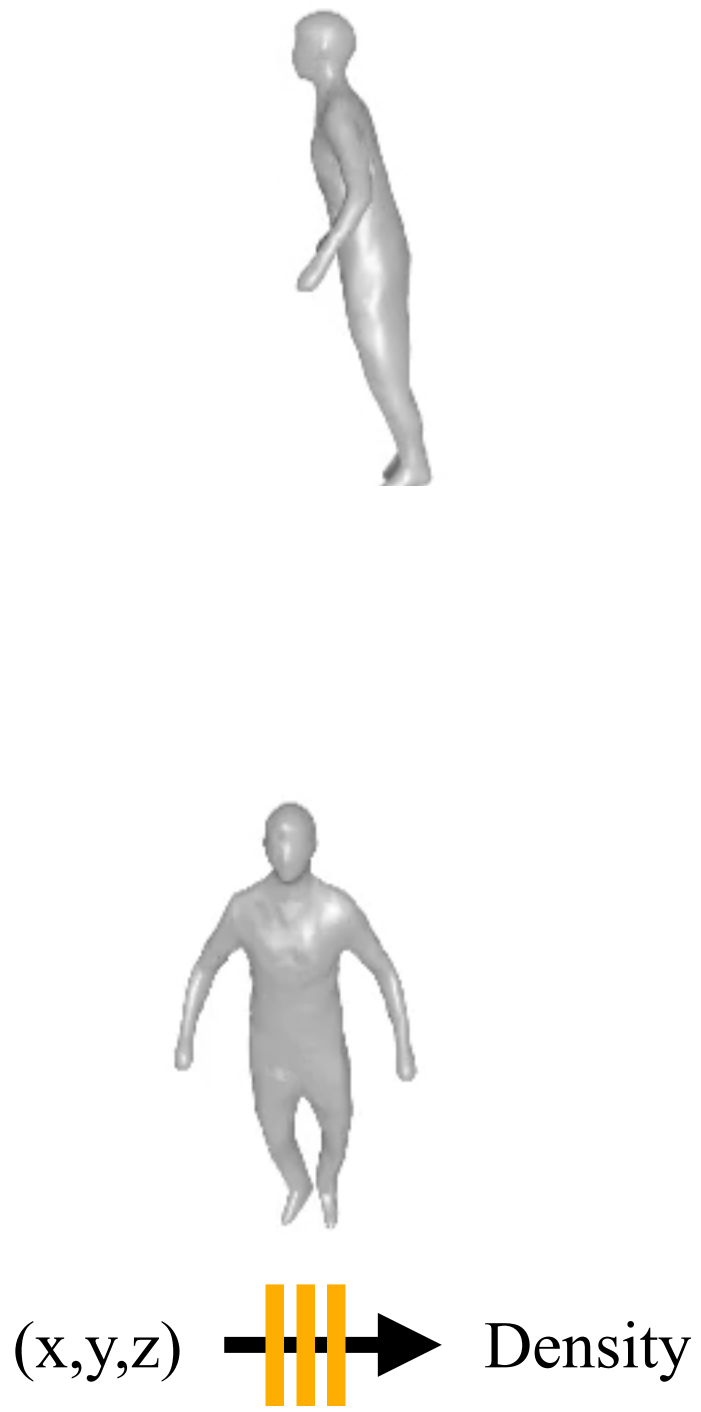
including: bone dimension, body shape and appearance

To model morphological variations over a category, we adopt the conditional NeRF formulation, and use a morphology code to control the shape and color of the reconstruction.

Modeling Morphology



Reference video

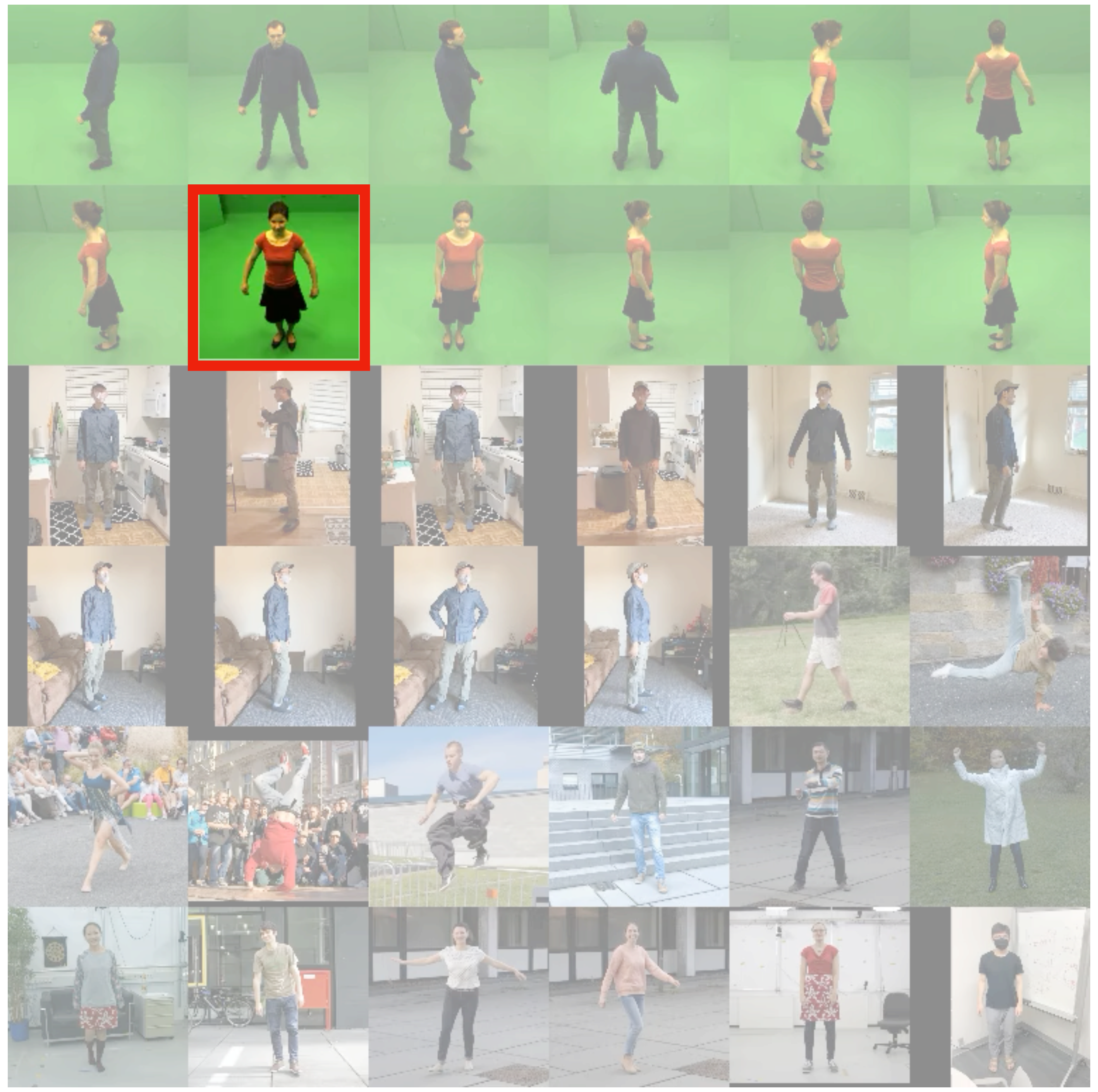


Shared morphology (BANMo)

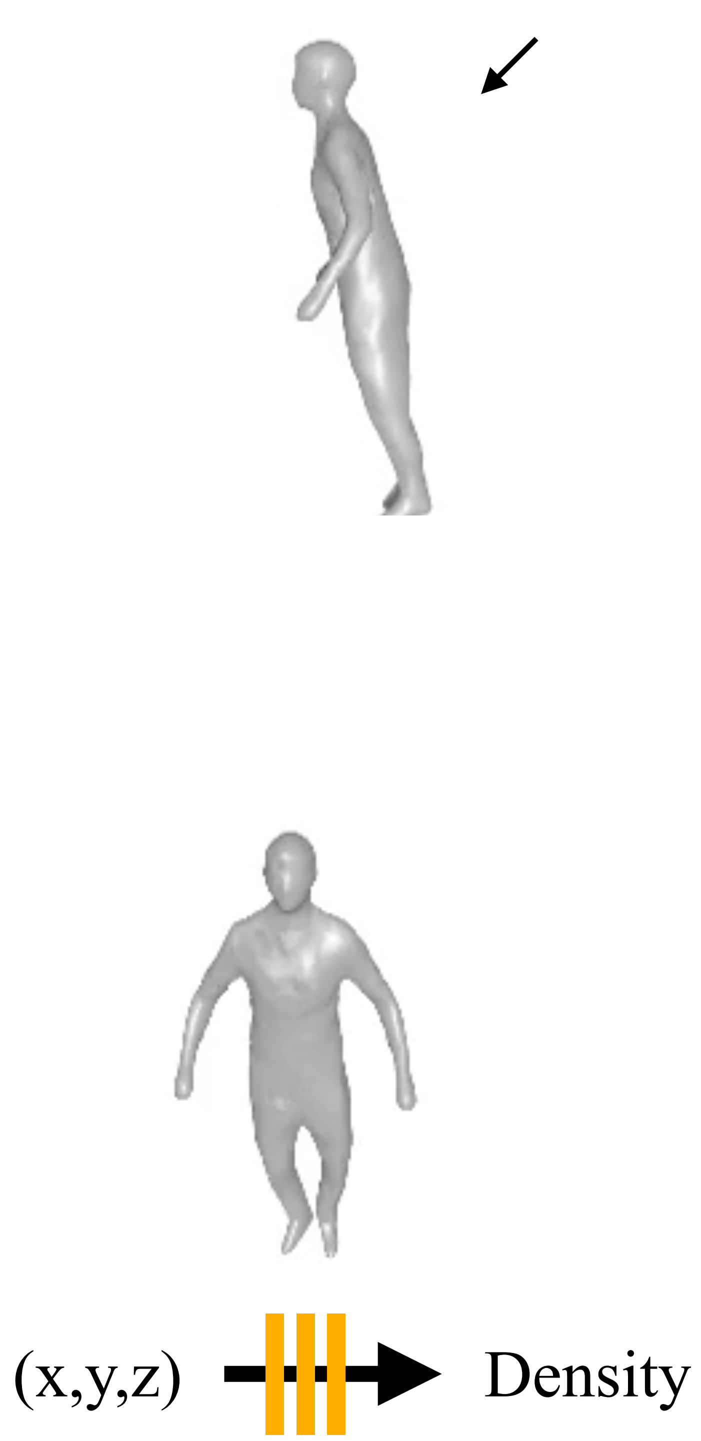
Observation 1: A shared morphology washed out *instance details* (e.g., hair, cloth).

We observe that methods assuming a shared morphology fail to reconstruct instance details.

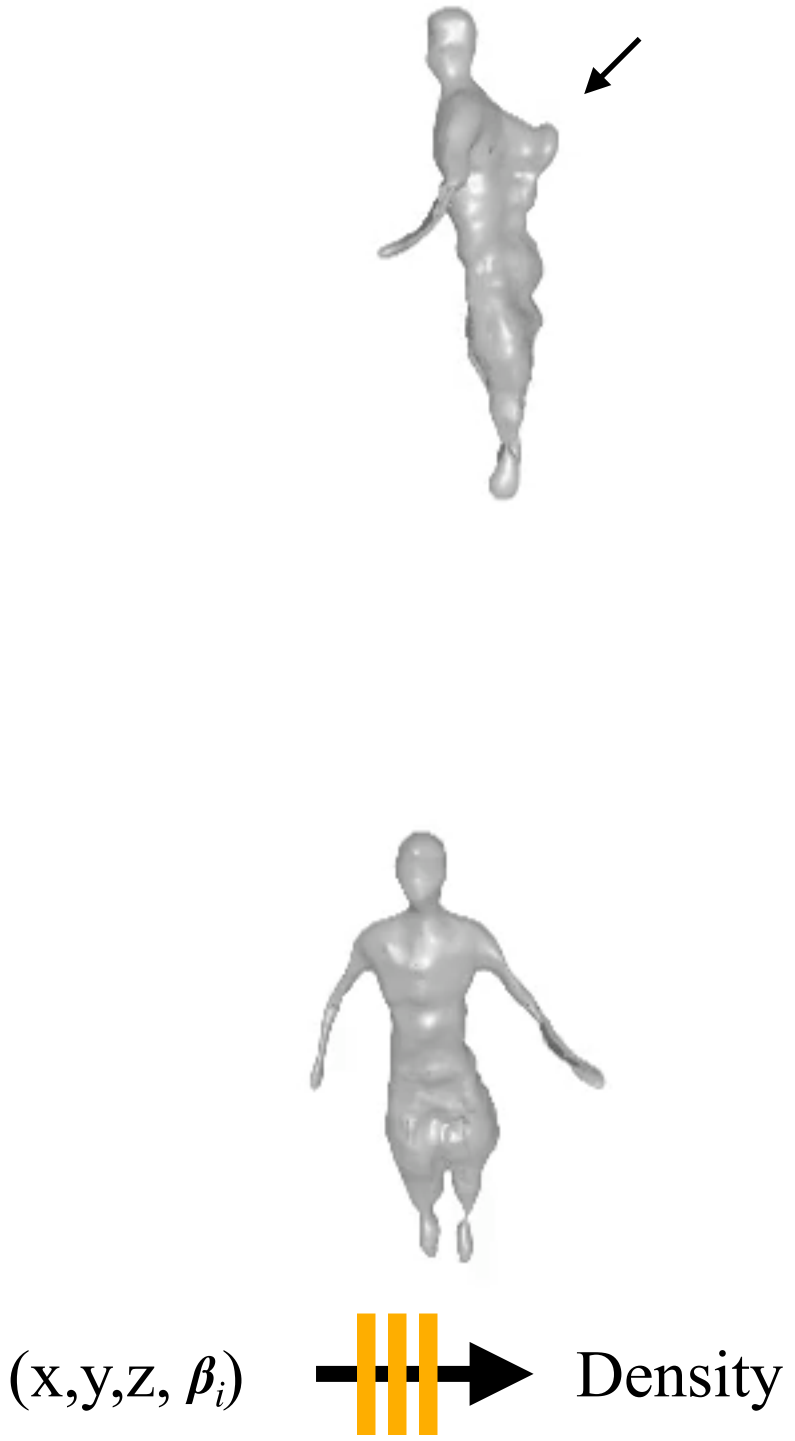
Modeling Morphology



Reference video



Shared morphology (BANMo)



Instance morphology

Observation 1: A shared morphology washed out *instance details* (e.g., hair, cloth).

Observation 2: Instance-specific morphology fails to hallucinate the invisible surface.

On the contrary, adding the morphology code improves the instance details from the reference viewpoint. However, the reconstruction appears abnormal from a novel viewpoint.

Morphology Code Regularization

Morphology code

$$\sigma = \phi(\mathbf{x}, \boldsymbol{\beta}_i \in \mathbb{R}^{32})$$

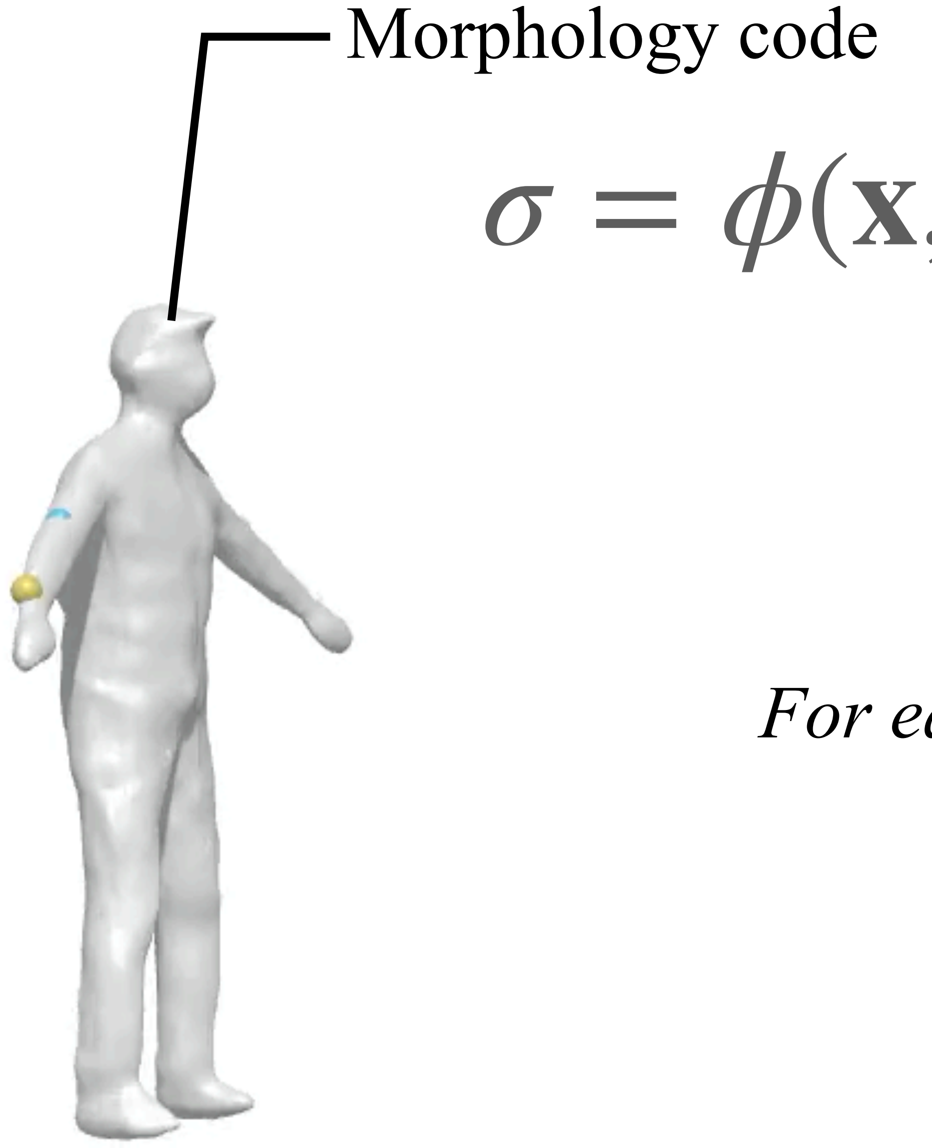
Morphology: Between-Instance Variations

including: bone dimension, body shape and appearance

Code Annealing asks a latent code to represent *any* data point with an annealing schedule.

To solve this problem, we use a code annealing strategy. At the beginning of optimization, we ask the latent code to represent any instance in the data, and then gradually specialize the latent codes to their corresponding videos.

Morphology Code Regularization



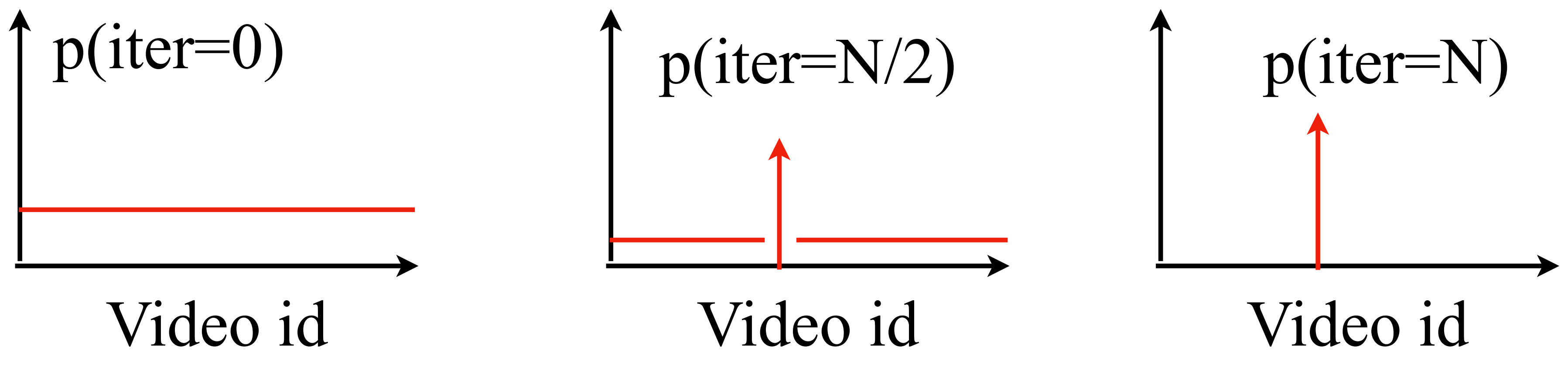
$$\sigma = \phi(\mathbf{x}, \boldsymbol{\beta}_i \in \mathbb{R}^{32})$$

Morphology: Between-Instance Variations

including: bone dimension, body shape and appearance

For each video i, $\cancel{\min_{\beta} ||\mathbf{c}_i - \mathscr{R}ender(\beta_i)||}$

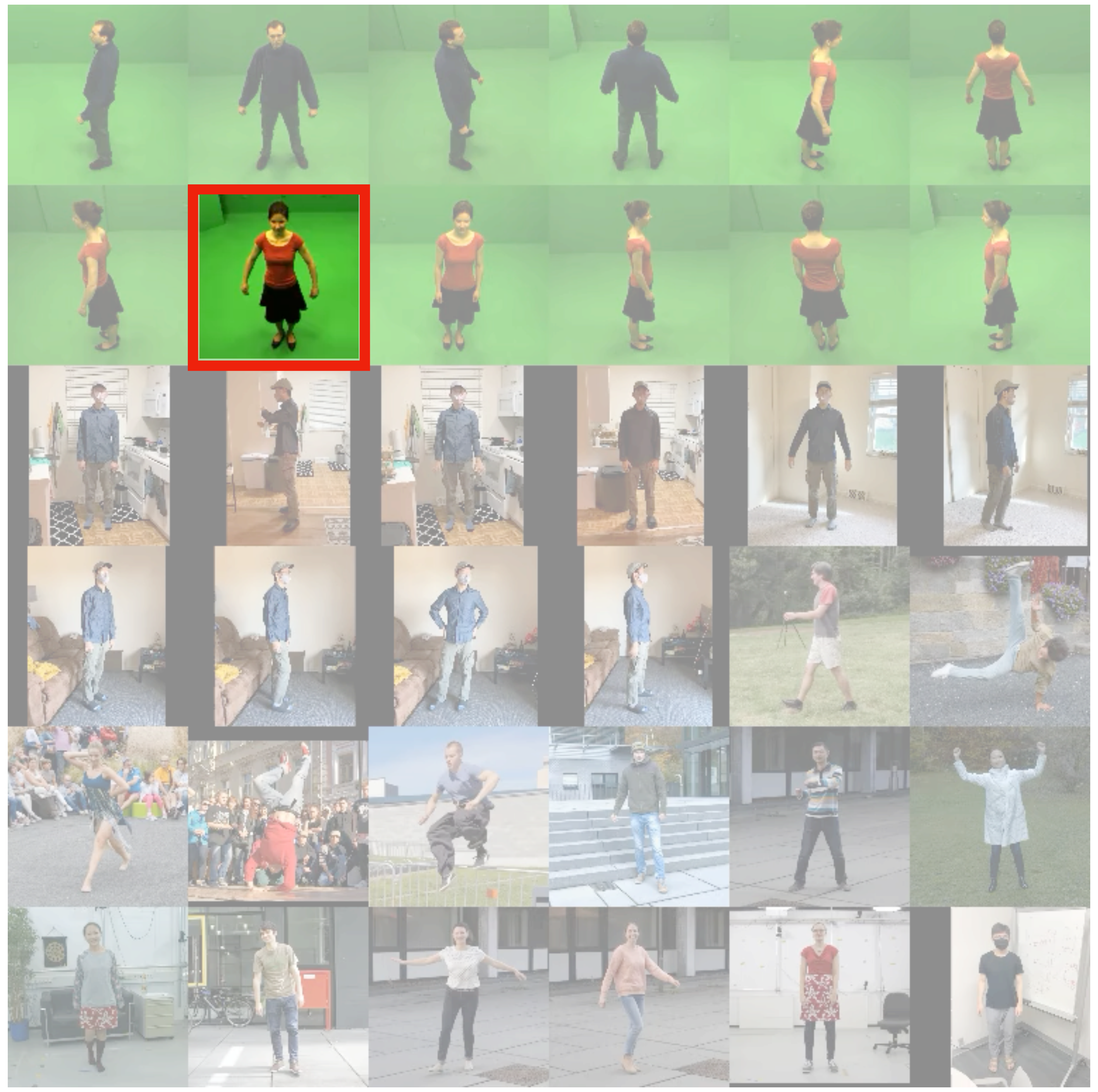
$$\min_{\beta} ||\mathbf{c}_i - \mathscr{R}ender(\beta_j)||, \quad j \sim D_t(i)$$



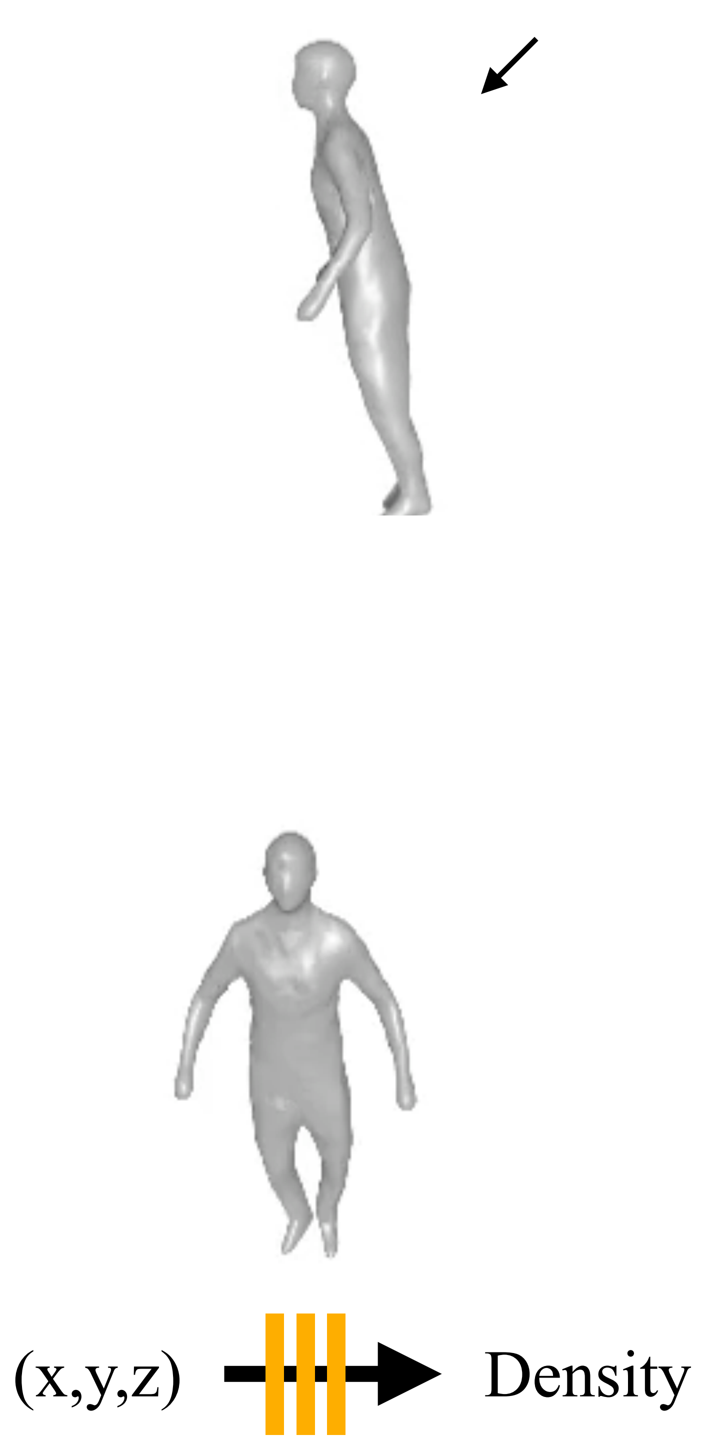
Code Annealing asks a latent code to represent *any* data point with an annealing schedule.

This can be operationalized by manipulating the probability of latent code sampling. Instead of using the latent code corresponding to video i, we randomly sample the code with a probability that goes from 1 to a small value.

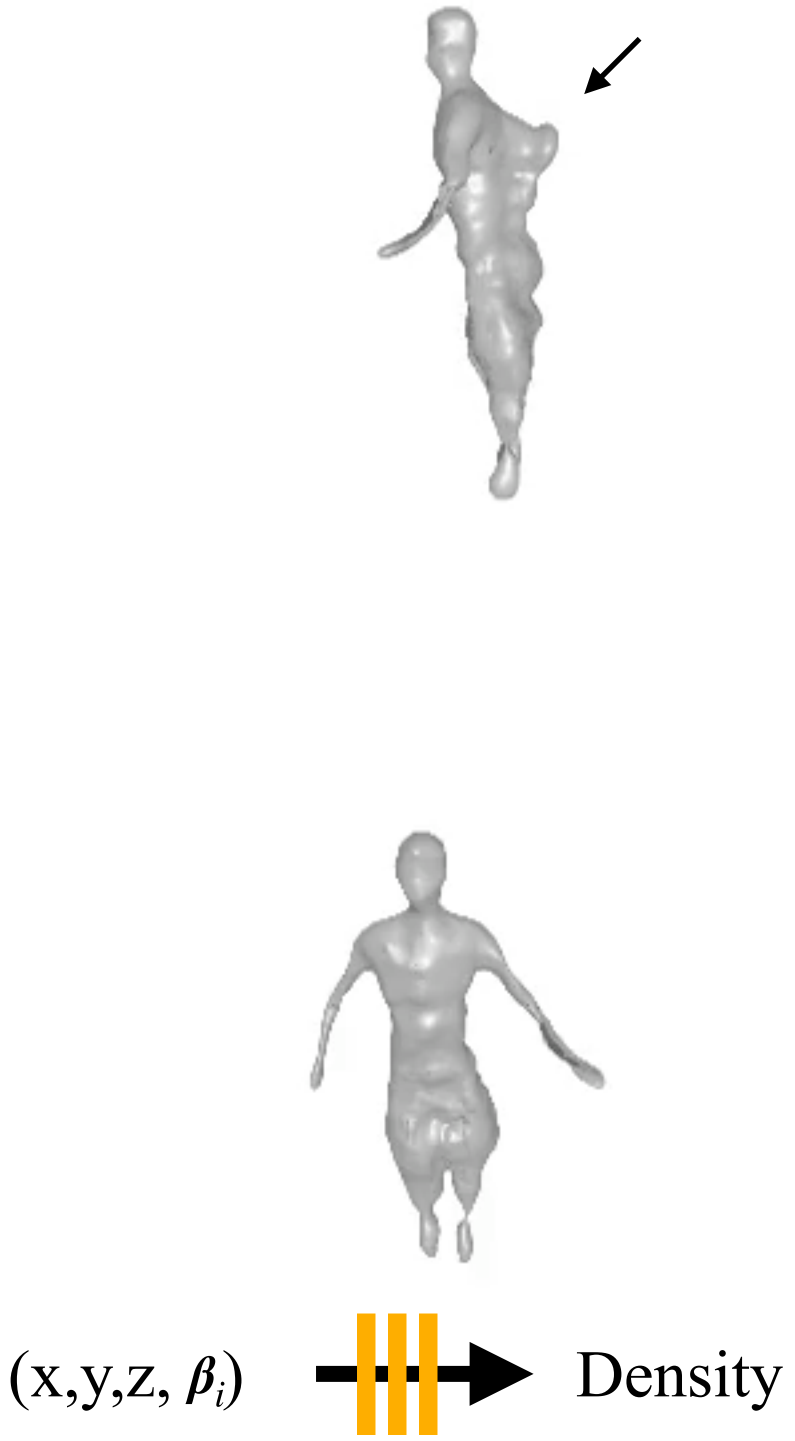
Morphology Code Regularization



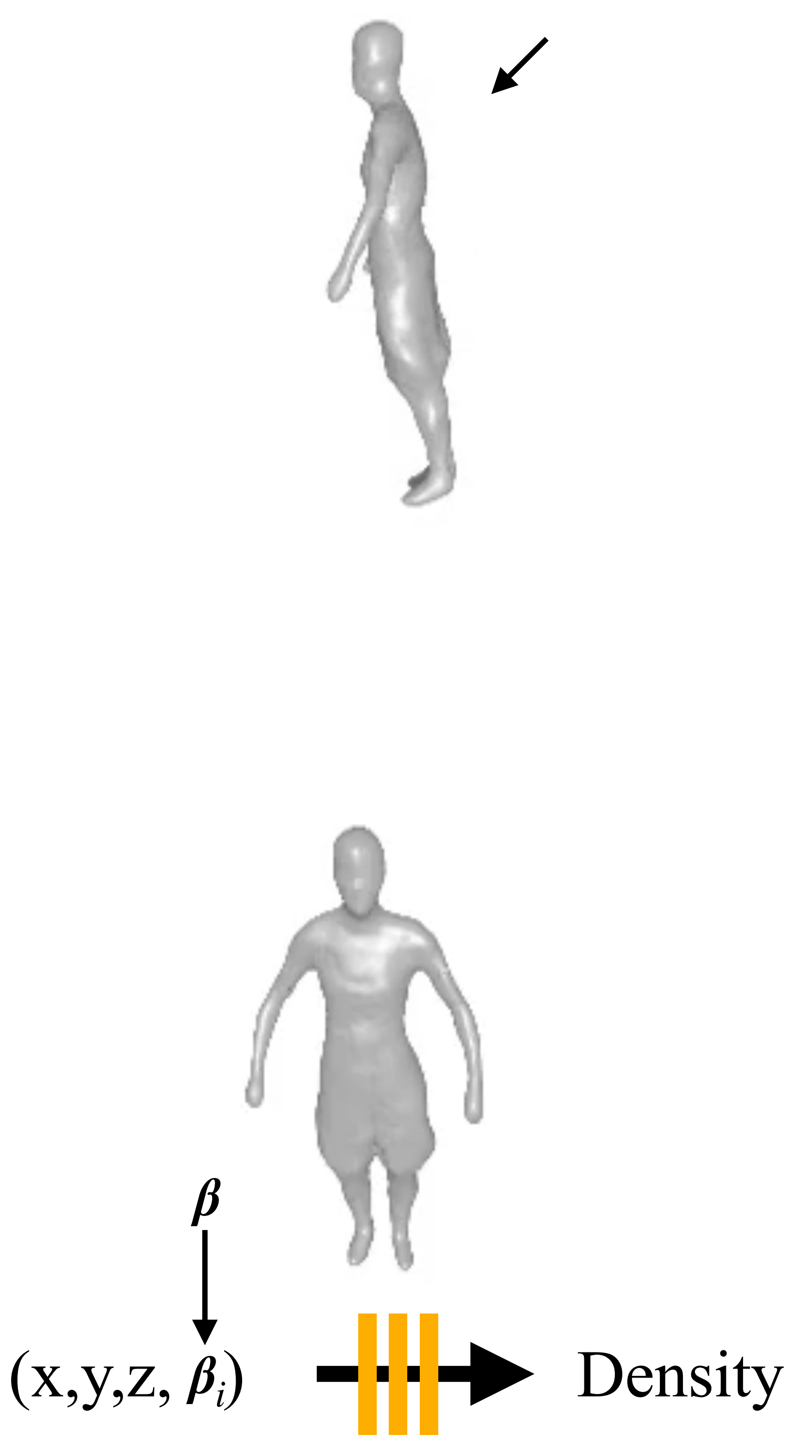
Reference video



Shared morphology (BANMo)



Instance morphology



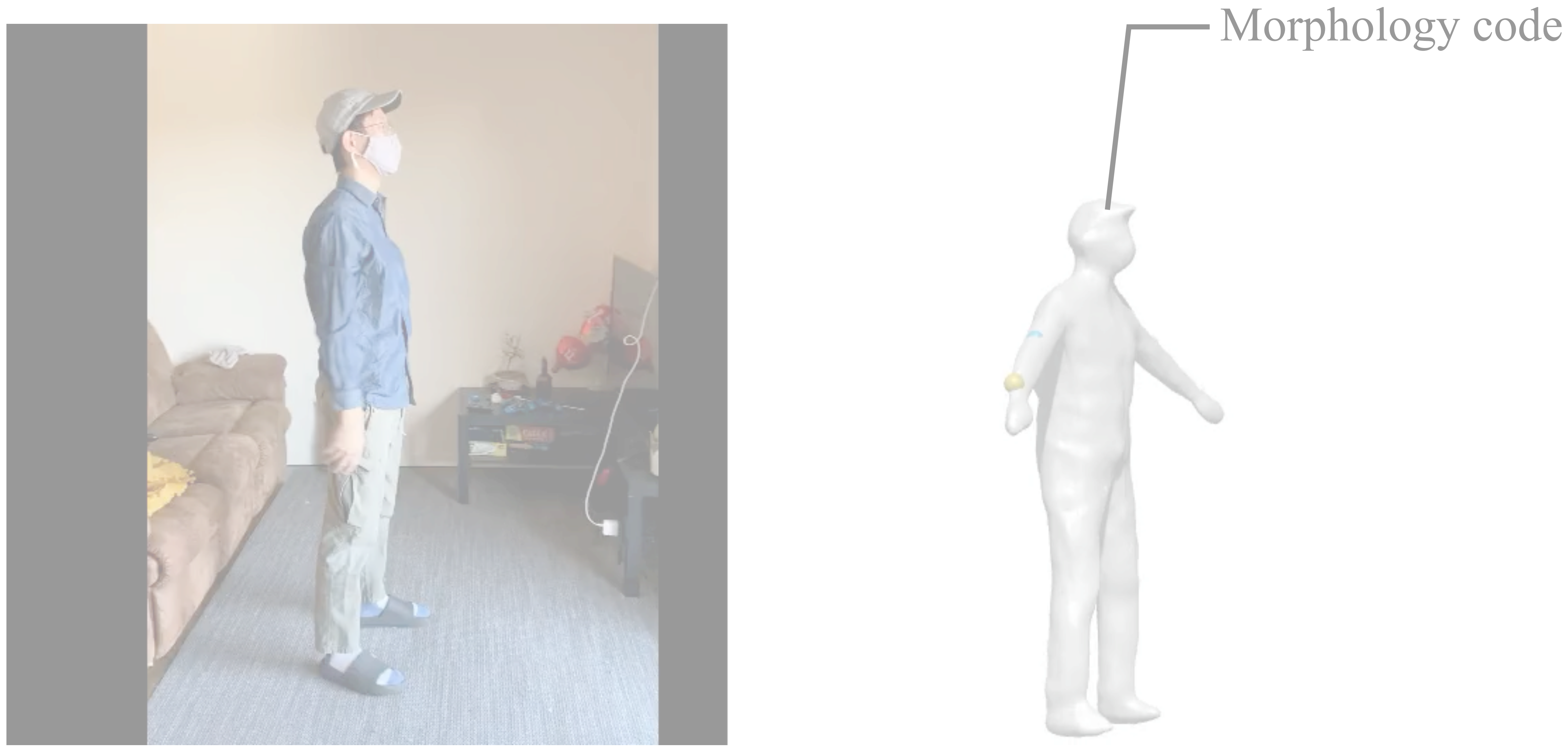
Code Annealing

Observation 3: Code annealing finds a trade-off between sharing and specialization.

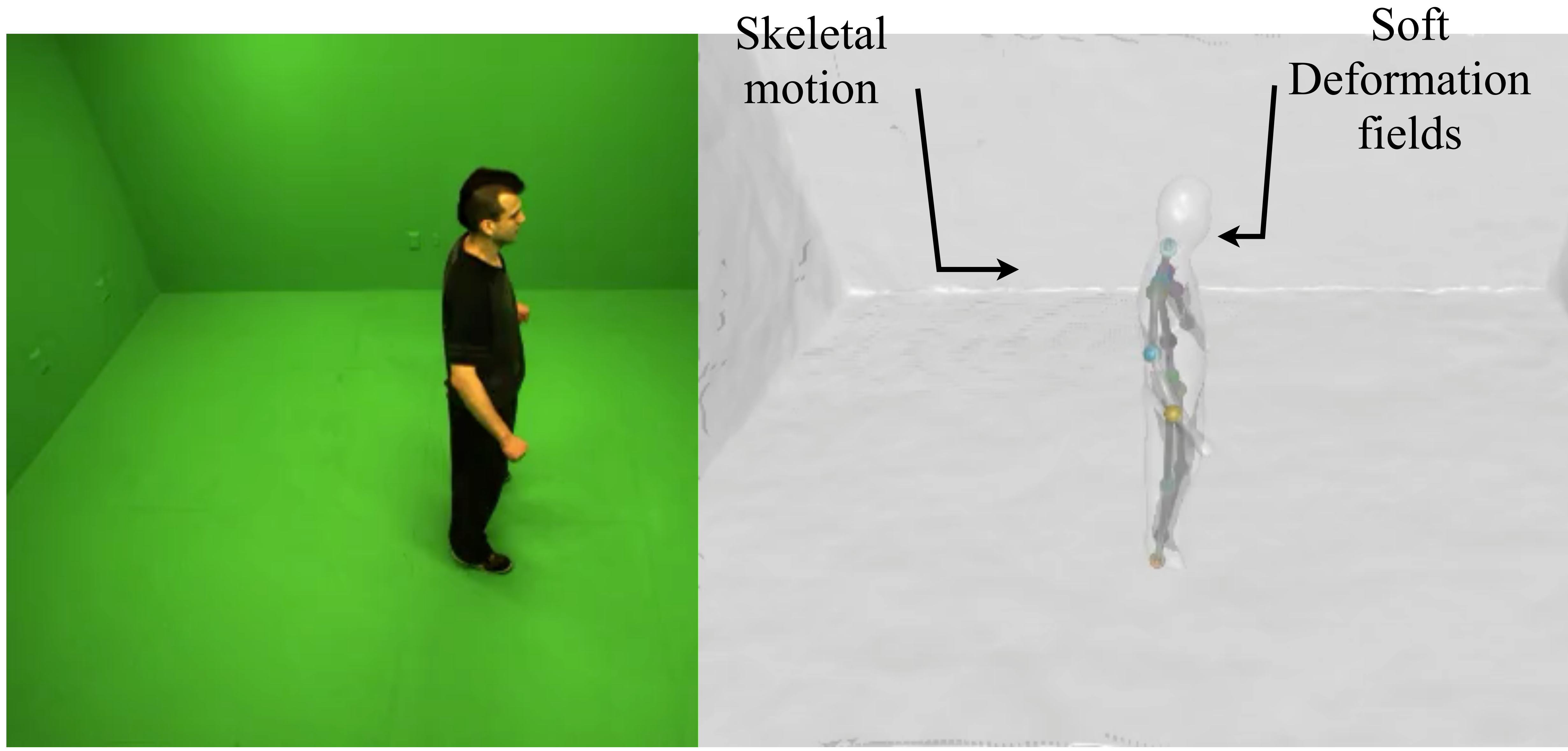
We found that code annealing regularizes the latent representation, and helps recovering the invisible surface.

Modeling Motion

[HumanNeRF 2022]



Morphology: Between-Instance Variations
including: bone dimension, body shape and appearance



Motion: Within-Instance Variations
Including: skeleton articulations, soft deformations

To model motion, we use a hybrid representation including a skeleton-based blend skinning field that represents the majority of the motion, as well sa a soft-deformation field to represent those not explained by the skeleton, such as cloth deformation.

Modeling Motion: Control Points versus Skeleton

Motion source Retargeting

Control-point-based deformation
(BANMo)

- Control points are too flexible (e.g., does not preserve length), making deformations squashy.

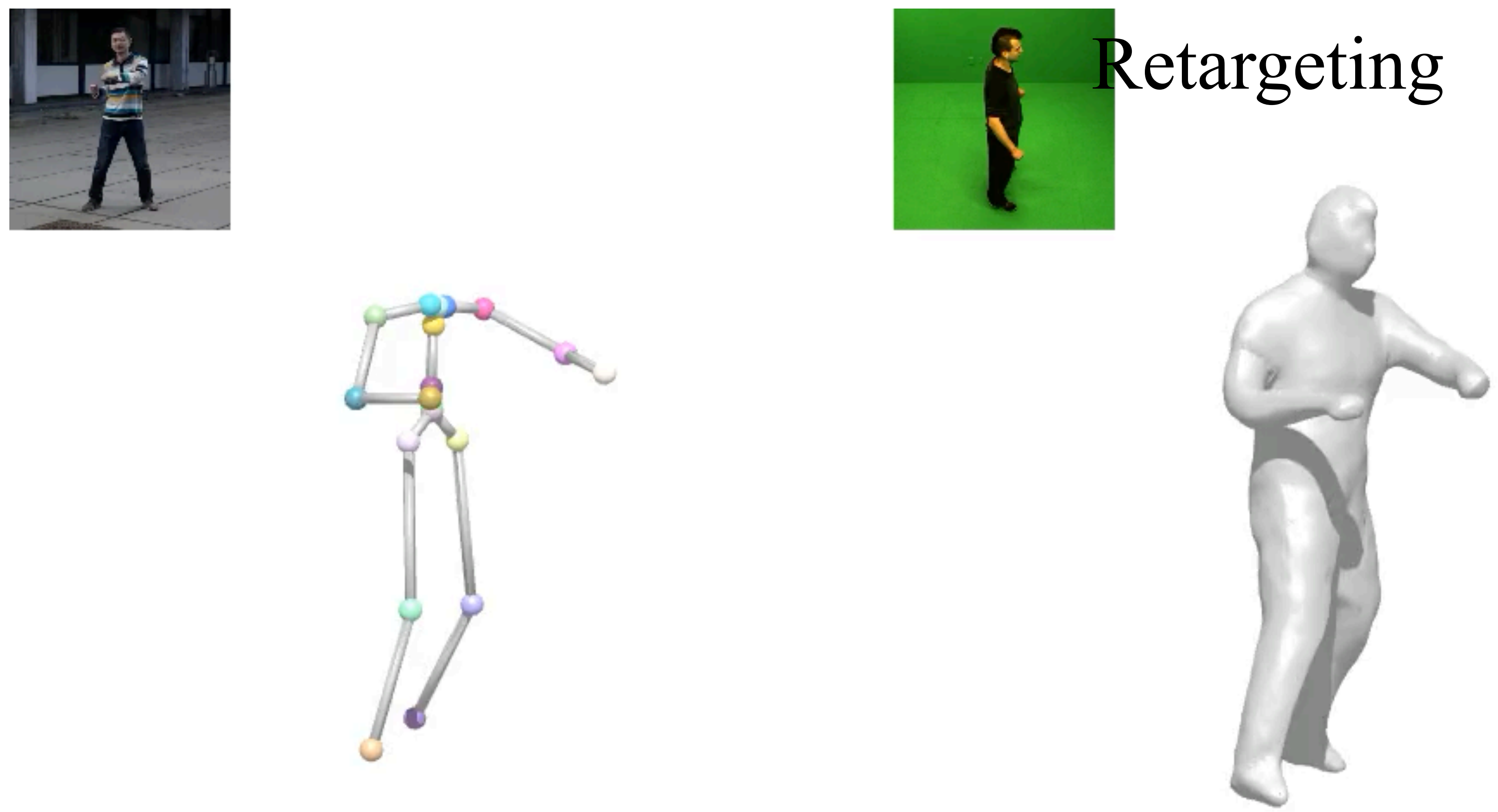
We compare the deformation field representation with control-point-based deformation in BANMo. The motion is transfered from the left to the human subject on the right. Note that control-point-based deformation looks squashy since it does not preserve length over time.

Modeling Motion: Control Points versus Skeleton

Motion source　　Retargeting

Control-point-based deformation
(BANMo)

Motion source



Skeleton deformation

- Control points are too flexible (e.g., does not preserve length), making deformations squashy.
- Solution: Using a skeleton to force fixed bone length over a video.

On the contrary, the skeleton-based deformation forced a fixed bone-length within a video, and produces better reconstruction and transfer quality.

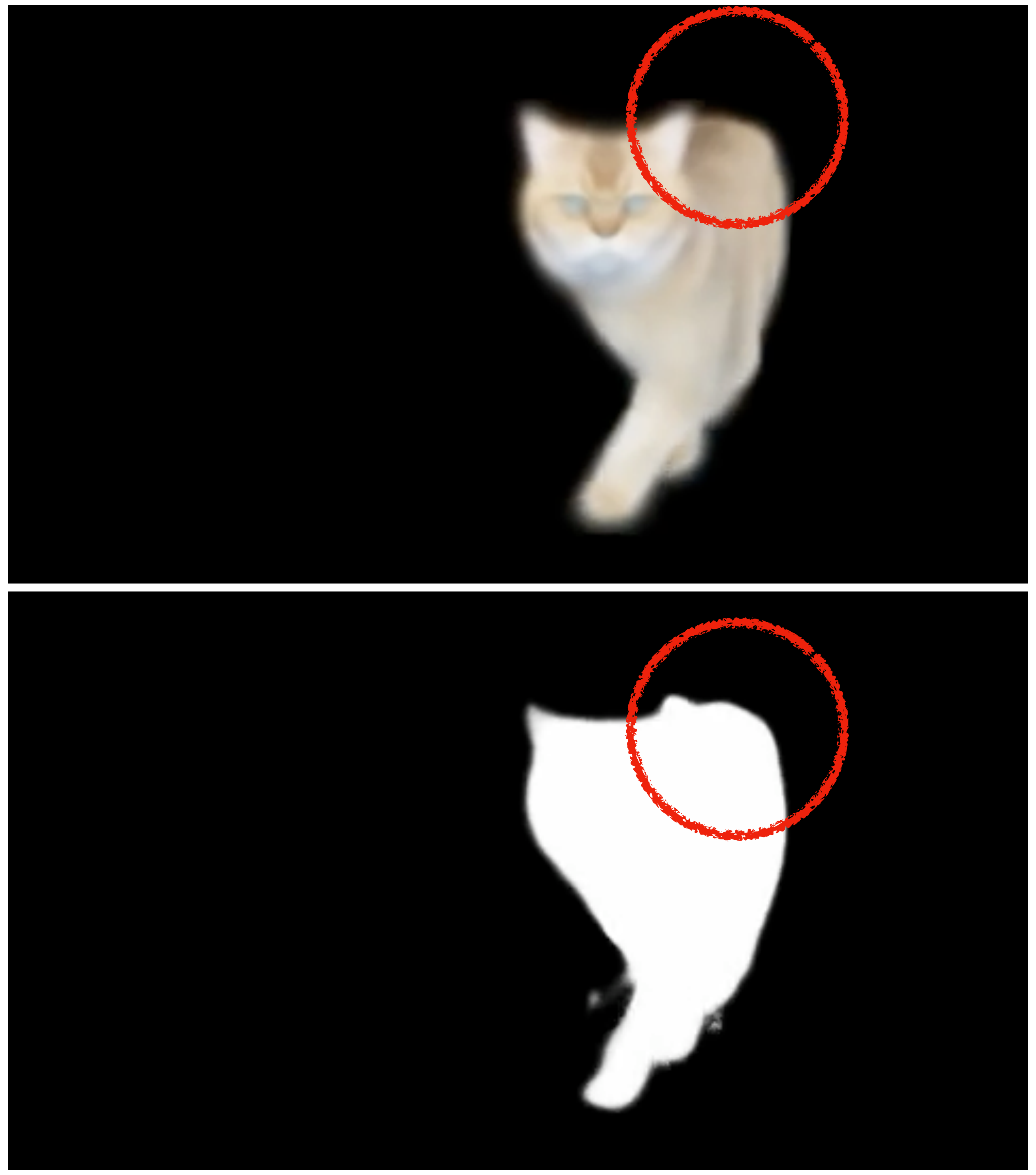
Modeling 3D Background

Reference video

To deal with the noisy input segmentation, we jointly optimize a background model with compositional rendering. In this example, the tail of cat is segmented as part of the background.

Modeling 3D Background

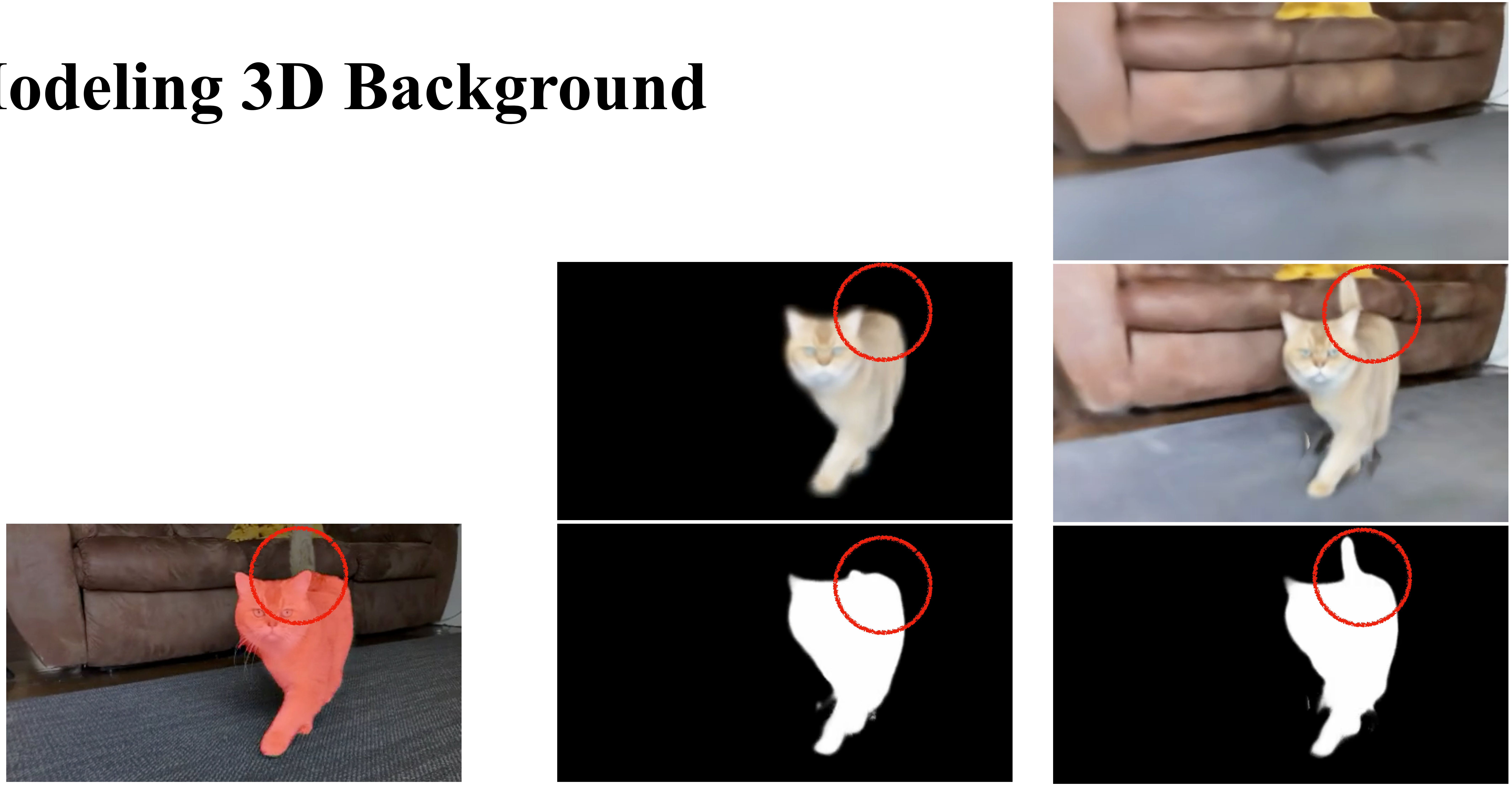
Reference video



W/o background

Without background modeling, the reconstruction of the tail appears hidden behind the body.

Modeling 3D Background



Reference video   W/o background   With background

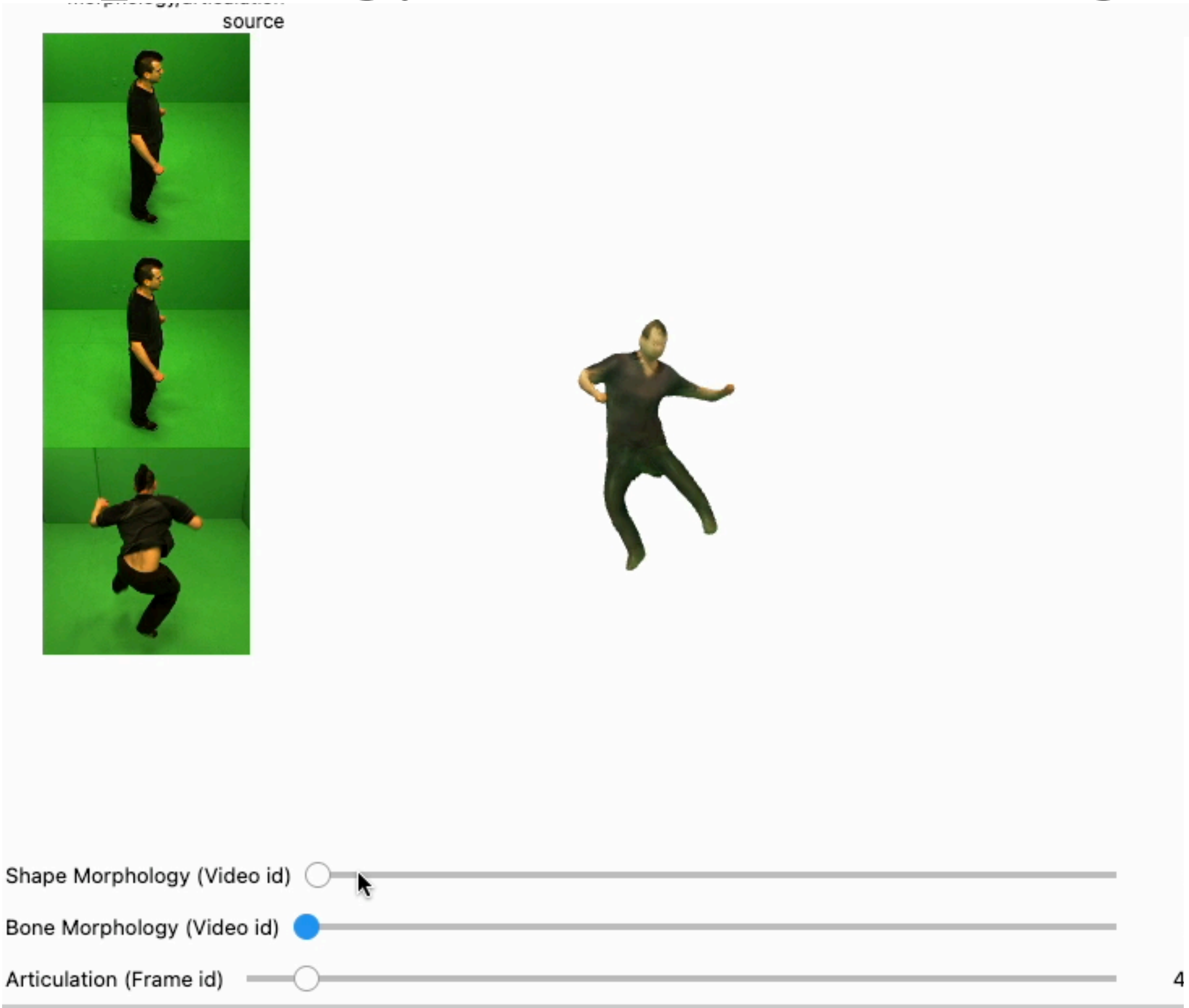
With background modeling, we are able to remove the inaccurate silhouette constraints and more faithfully segment the object and the background.

Application: Morphology-Motion Mixing

Shape source

Bone source

Articulation source



We show a demo of morphology-motion mixing. We first modify the morphology given the same pose, and then modify the pose while keeping the morphology unchanged.

Takeaways

- We leverage category prior for few view reconstruction.
- A code annealing method for disentangling morphology and motion.
- Skeleton and background modeling helps reconstruction and motion retargeting.