

Virtual Humans – Winter 23/24

Lecture 11_2 – Human Behavior Reconstruction from Images

Prof. Dr.-Ing. Gerard Pons-Moll

University of Tübingen / MPI-Informatics

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

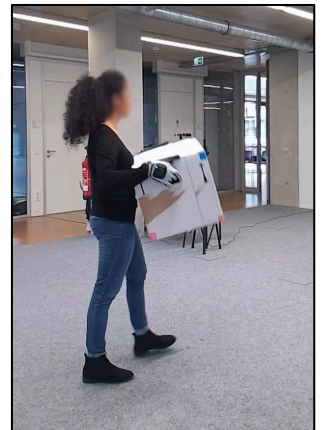
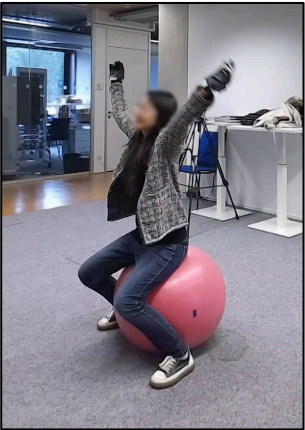


In this lecture...

- Classical work on human object interaction from images.
- Use instance specific optimization to fit SMPL + object mesh to an image.
- Learning to reconstruct human object interaction from data.
- Importance of pixel aligned learning.

Motivation

- Human-object interactions key to understanding the world.
- Images most common source of data.
- Can we identify action from an image?



[Early work from 2009]

Predict human-object interaction from video.


Predict from a video, \mathcal{V} :

- Object, $O = \{\text{class, xy-location}\}$
- Human motion, $m^h = \{\text{motion type, start time, end time, location}\}$
- Object motion, $m^o = \{\text{start time, location}\}$

[Early work from 2009, Pre-Deep Learning]

Predict human-object interaction from video.

Model task as Bayesian inference:

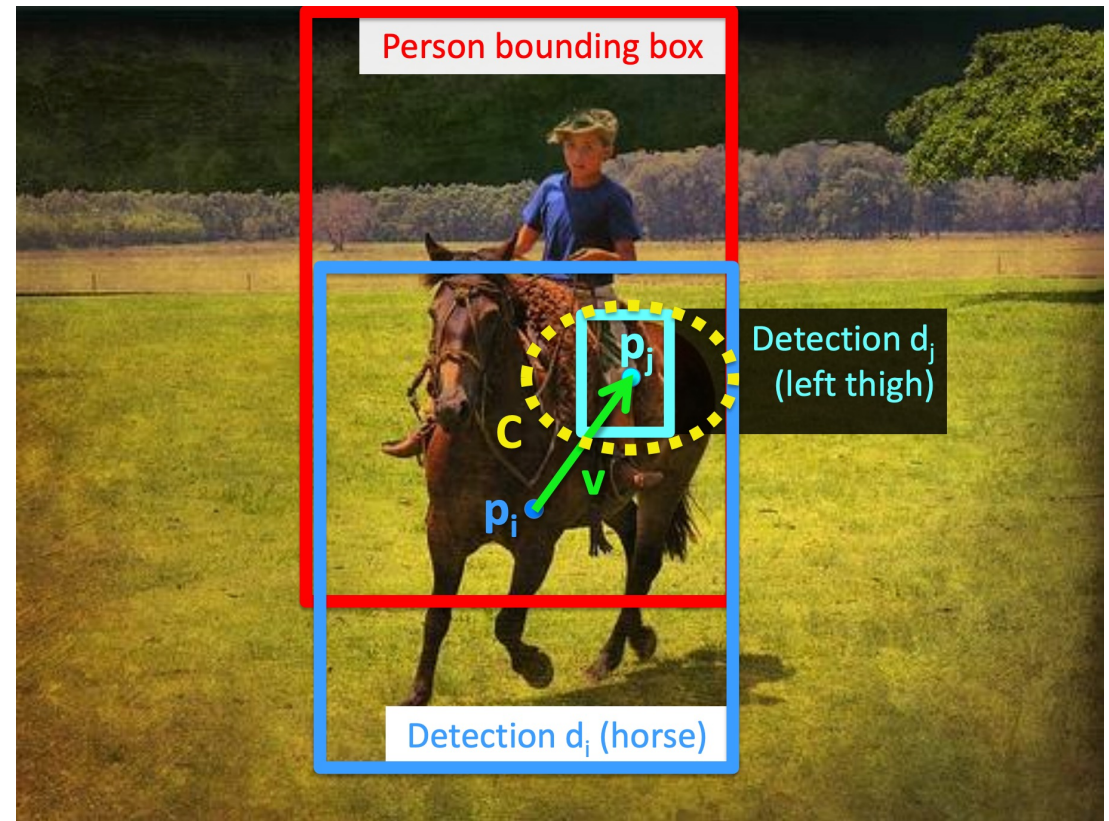
$$P(O, m^h, m^o | \mathcal{V}) \propto P(O | \mathcal{V}) \cdot P(m^h | O) \cdot P(m^h | \mathcal{V}) \cdot P(m^o | O, m^h) \cdot P(m^o | \mathcal{V})$$


All these terms had to be manually defined based on heuristics.
Difficult to scale.

[Early work from 2011, learning based]

Can we identify the interaction in an image?

- Detect in an image, I :
 - Human (b^h) and object (b^o) bounding boxes
 - Object category (o).
- Interaction = $f(I, b^h, b^o, o)$, where $f(\cdot)$ is a classifier.



- With availability of 2D annotated data, tasks like human and object detection/segmentation, tracking, action recognition have been well explored in the past
- What about modelling interactions in 3D?

BEHAVE captured H+O with 4 RGBD cameras.

RGB sequence



Tracking with BEHAVE model



Recap: BEHAVE

- BEHAVE a powerful recording setup, but **required 4 RGBD cameras**.
- Can we **reconstruct** human and object from **a single image**?

Perceiving 3D Human-Object Spatial Arrangements from a Single Image in the Wild (PHOSA)

Jason Y. Zhang*¹, Sam Pepose*², Hanbyul Joo², Deva Ramanan^{1,3}, Jitendra Malik^{2,4}, Angjoo Kanazawa⁴

¹ Carnegie Mellon University ² Facebook AI Research ³ Argo AI

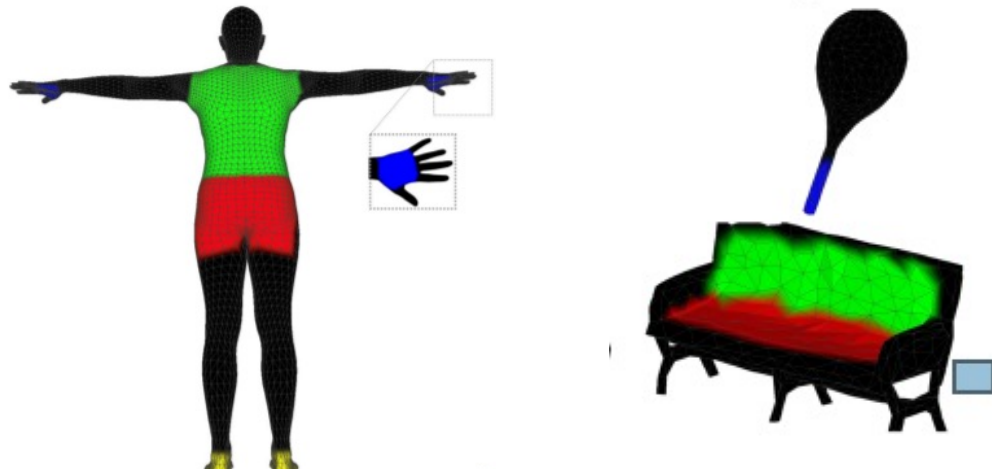
⁴ University of California, Berkeley

Goal

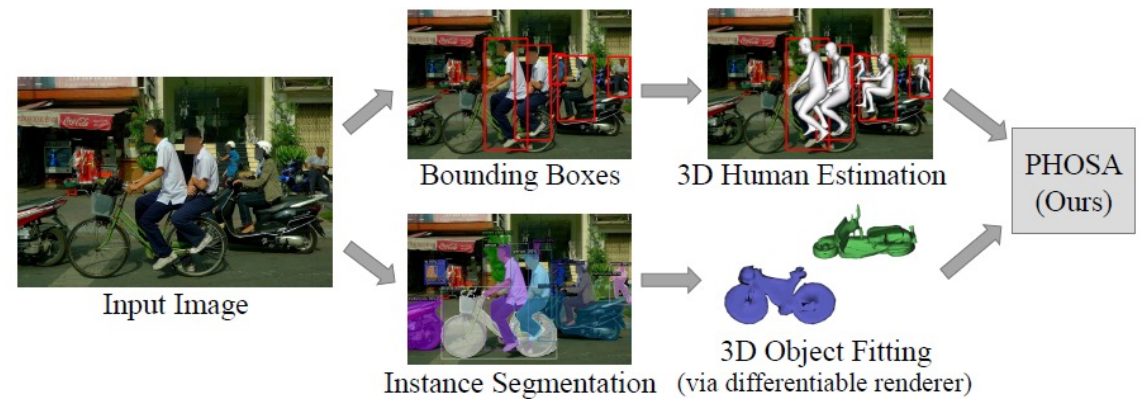
- Reconstruct human (SMPL) and object mesh from a single RGB image.
- Before BEHAVE. **No dataset existed for learning.**
- Purely optimization and heuristic based.

PHOSA Key Ideas

1. Use predefined contact part pairs, eg.
 - Body back – bench/chair back
 - Body butt – bench/chair seat



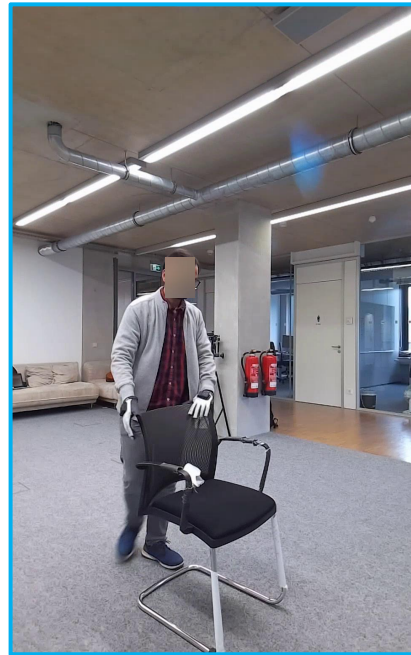
2. Reconstruct human & object separately.
Optimize based on contact pairs.



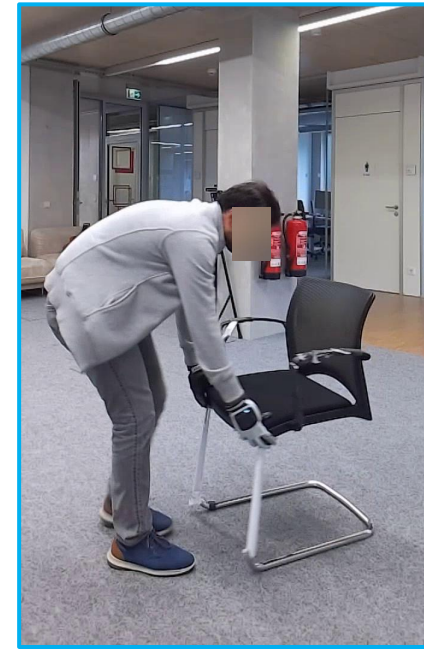
But, one can interact with the object in different ways...



Sit



Lean

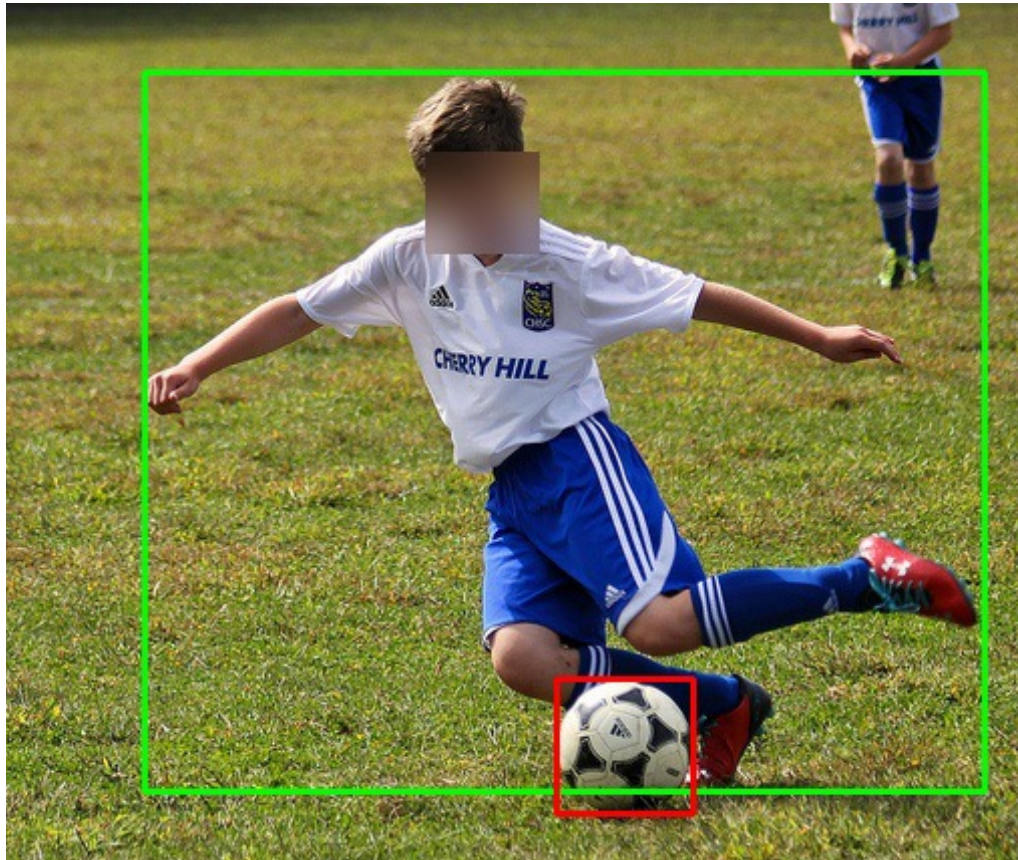


Move

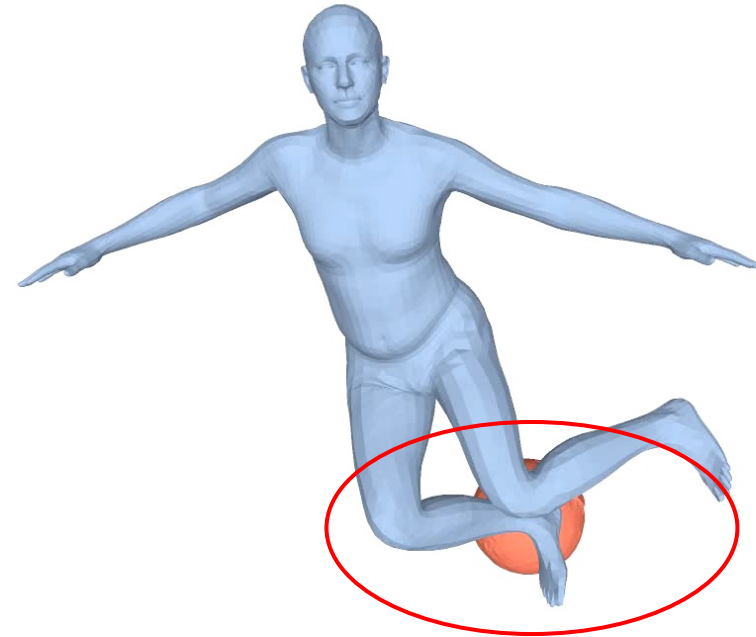
PHOSA: Zhang et al. ECCV'20

PHOSA cannot jointly reason about the human-object arrangement

Input Image



Reconstruction from PHOSA



PHOSA: Zhang et al. ECCV'20

Summary of PHOSA

- PHOSA used innovative heuristics to reconstruct H+O.
- Heuristics difficult to scale.
- **Can we use learning to reconstruct HOI?**



CHORE: Contact, Human and Object REconstruction from a single RGB Image

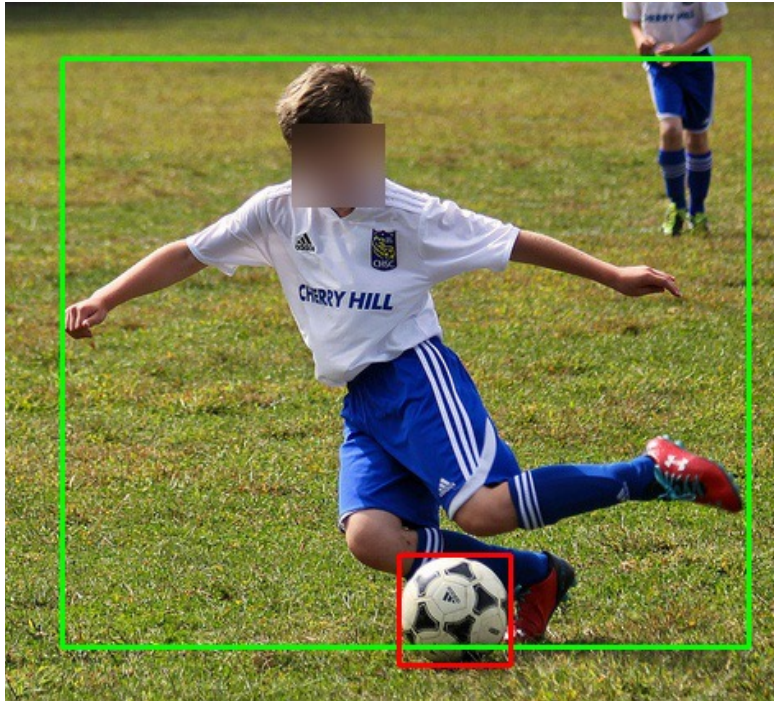
Xianghui Xie², Bharat Lal Bhatnagar^{1,2}, Gerard Pons-Moll^{1,2}

¹University of Tübingen, Germany

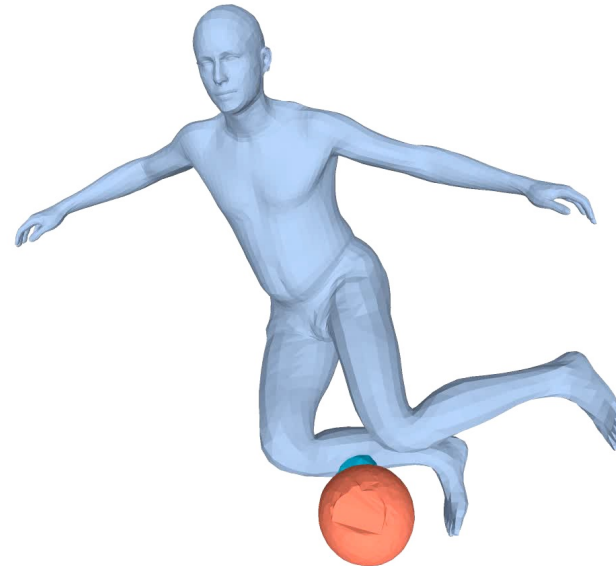
²Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

GOAL: Joint reconstruction of human, object and the contacts between them

Input Image



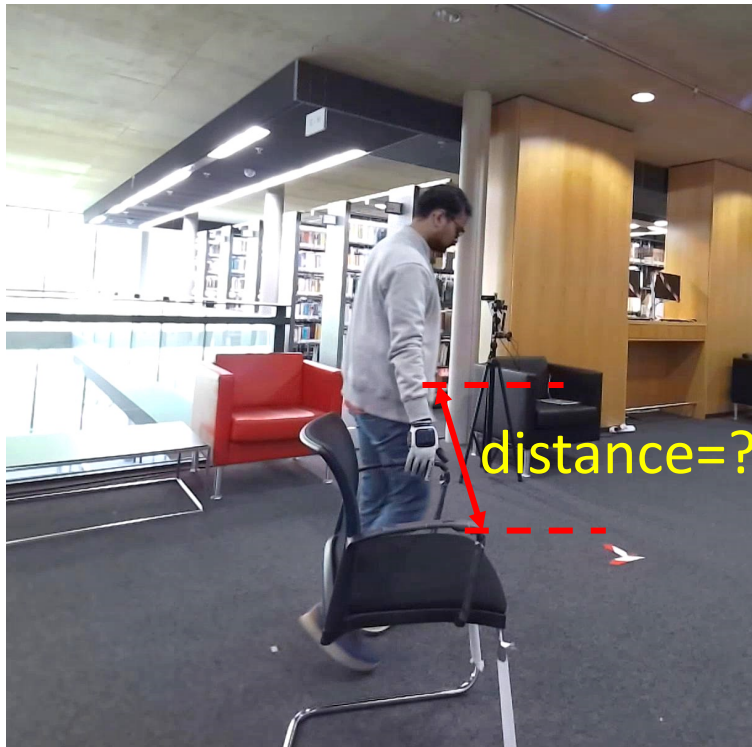
Ours



Why is this a hard problem?

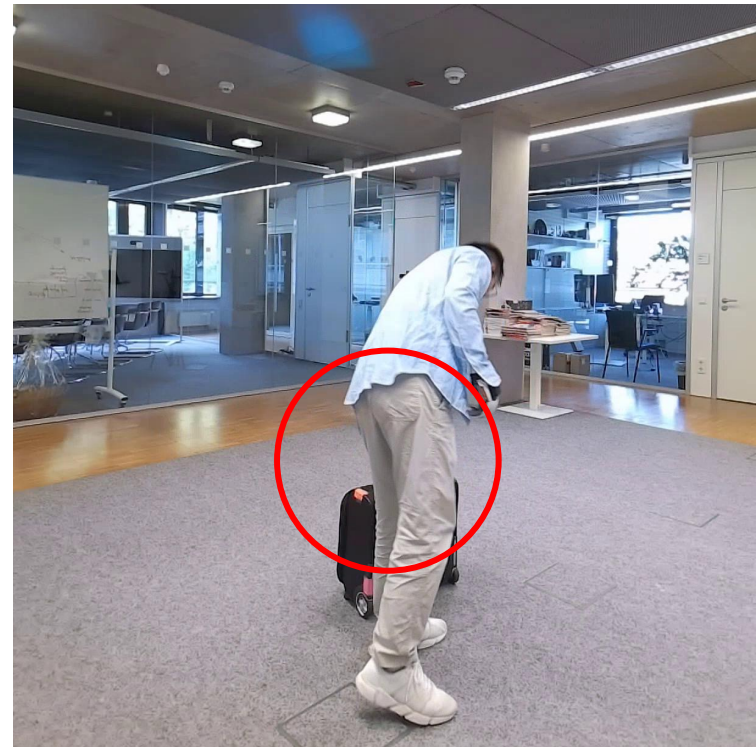
- Depth ambiguity

Distance between human and object unknown from single view.

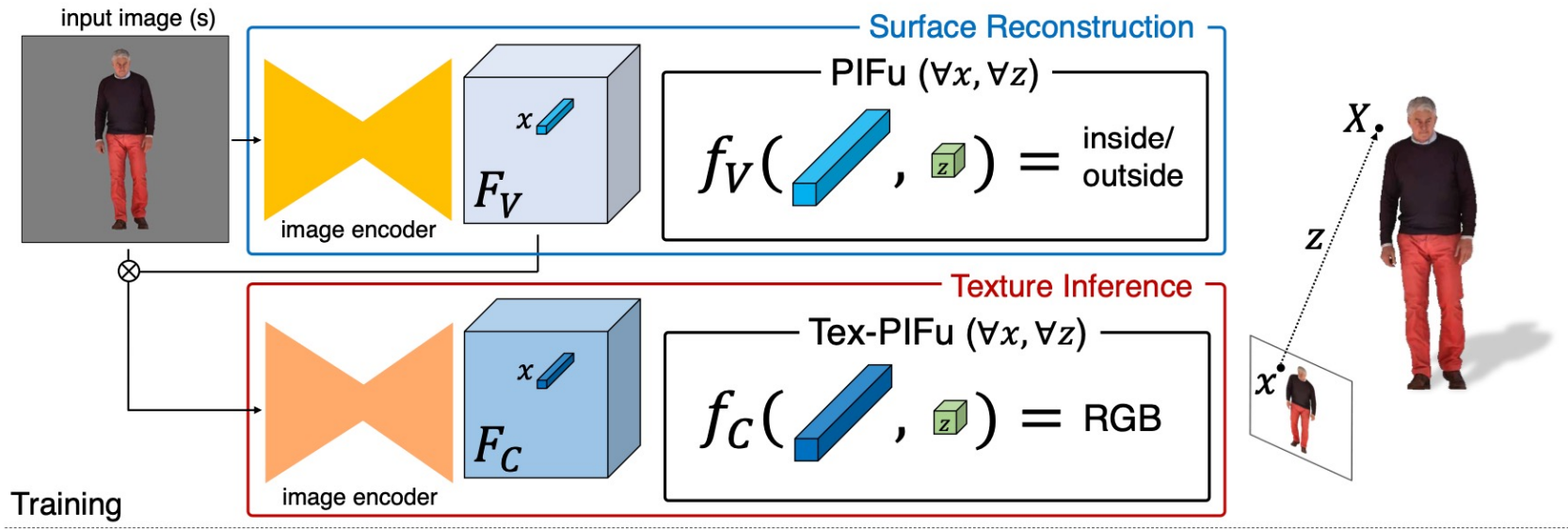


- Heavy occlusion

Human and object occlude each other during interaction.

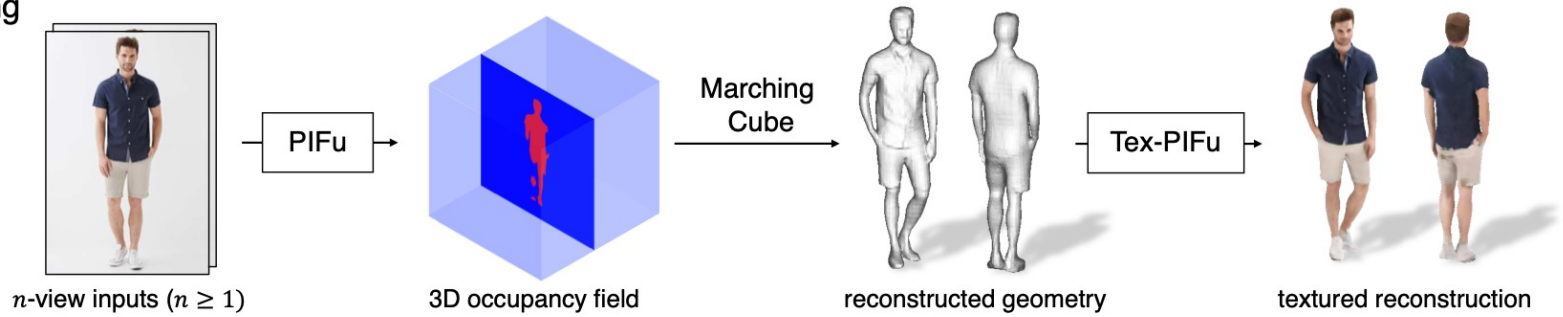


PiFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization



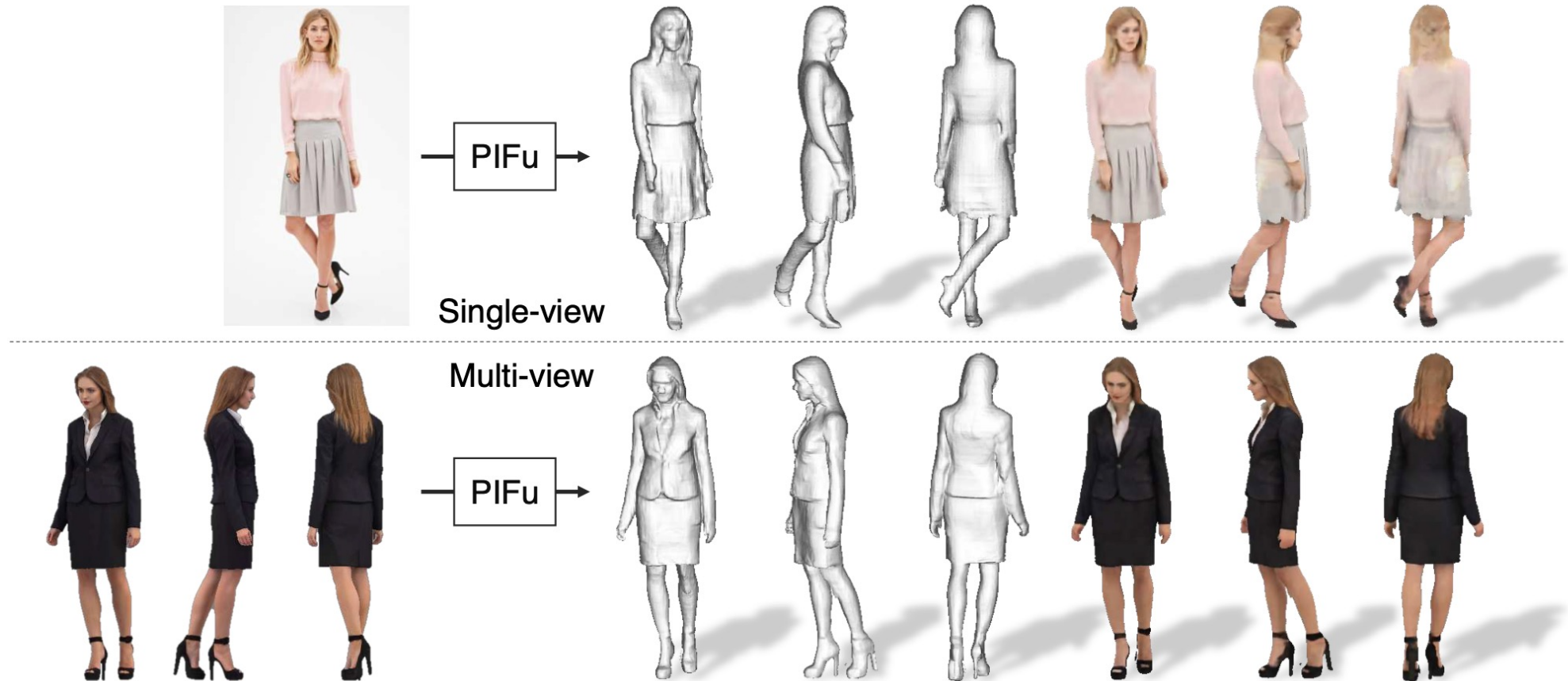
Training

Testing



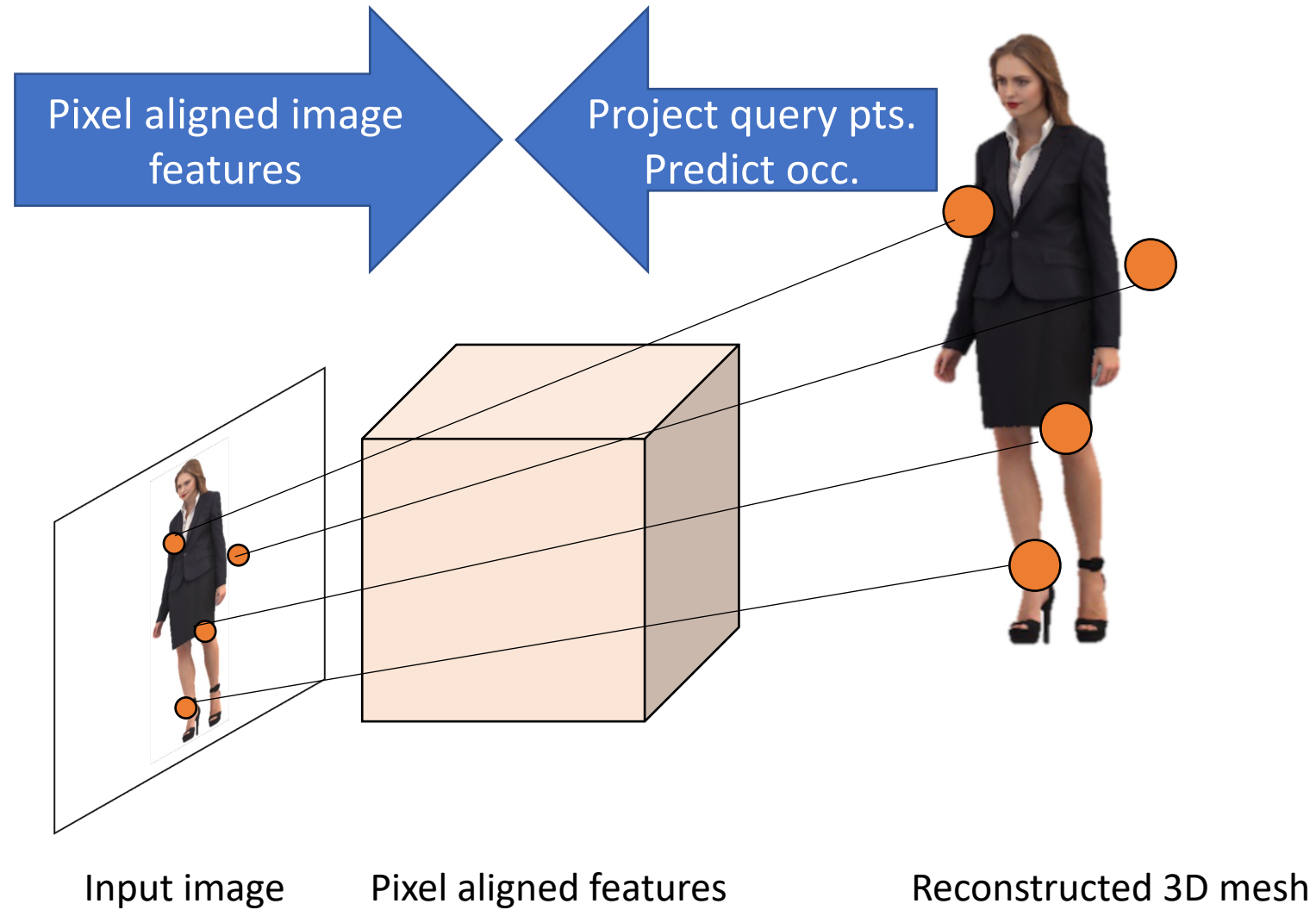
[PiFu. ICCV'19]

PiFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization

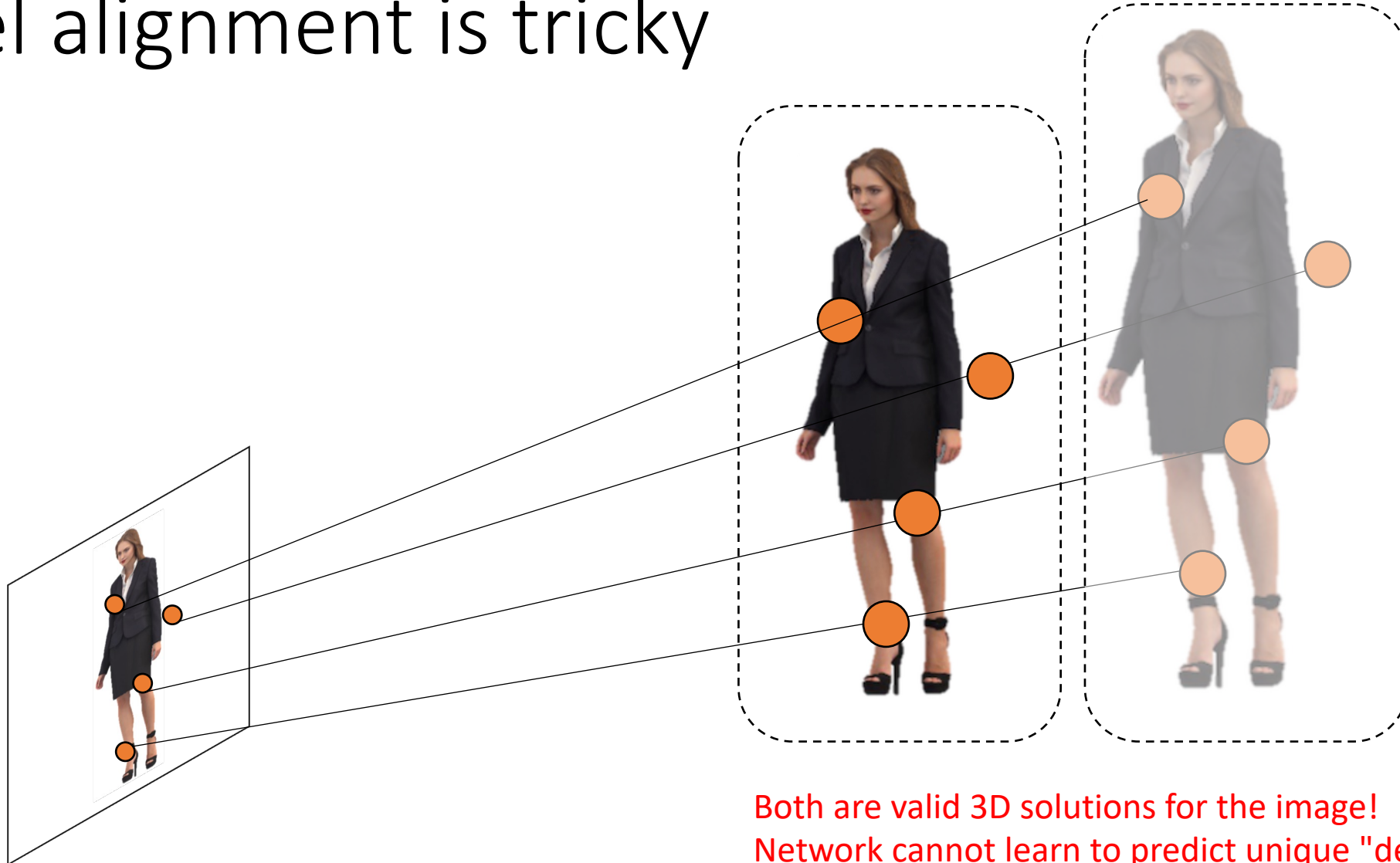


[PiFu. ICCV'19]

Recall: Pixel aligned implicit reconstruction.



Pixel alignment is tricky



Both are valid 3D solutions for the image!
Network cannot learn to predict unique "depth-scale"

PiFU summary

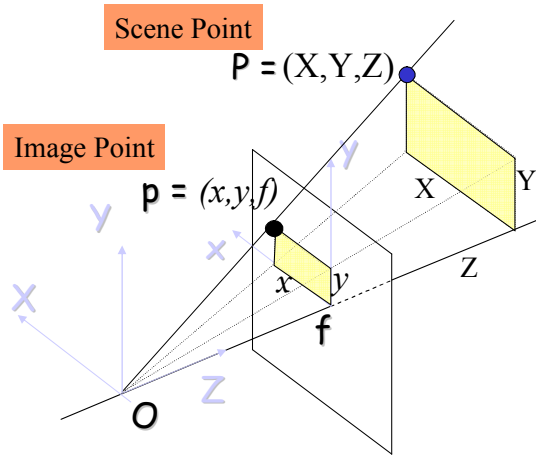
- ✓ Fix depth, reconstruct only scale.
- ✓ Maintain pixel alignment.
- ✗ In PiFu, since the training data comes from 3D scans, this can be done easily by controlling the rendering.

But, what if we have real captured images? How do we fix depth while keeping pixel alignment?

Recall: perspective effects

Robert Collins
CSE486, Penn State

Basic Perspective Projection



Perspective Projection Eqns

$$x = f \frac{X}{Z} + c$$
$$y = f \frac{Y}{Z} + c$$

3 things affect the size of an object/human in the image:

- The focal length f
- The physical size of the object X, Y
- The depth Z

We want to:

1. Change depth from $Z \mapsto Z'$
2. Maintain alignment with image, i.e. (x, y) should remain same.

From the laws of perspective:

- Scaling mesh vertices V by a scale s does not change its projection to the image.
- Proof:

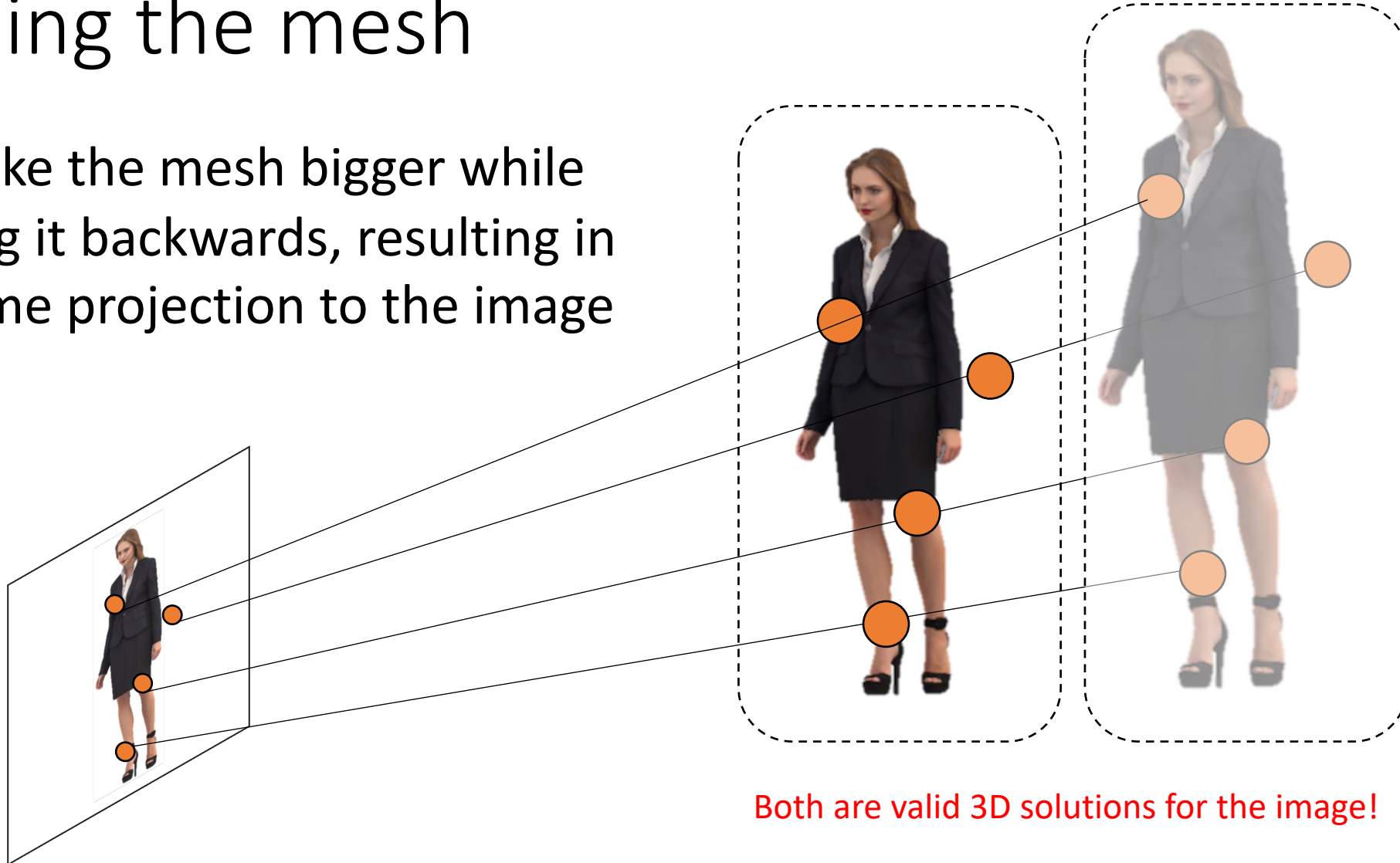
$$\text{Let } \mathbf{v} = (\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z) \in \mathbf{V} \quad \mathbf{v}' = (s\mathbf{v}_x, s\mathbf{v}_y, s\mathbf{v}_z) \in s\mathbf{V}$$

Projection to image

$$\pi(\mathbf{v}) = \left(f_x \frac{\check{\mathbf{v}}_x}{\mathbf{v}_z} + c_x, f_y \frac{\mathbf{v}_y}{\mathbf{v}_z} + c_y \right) = \pi(\mathbf{v}') = \left(f_x \frac{s\mathbf{v}_x}{s\mathbf{v}_z} + c_x, f_y \frac{s\mathbf{v}_y}{s\mathbf{v}_z} + c_y \right) = \pi(\mathbf{v})$$

Scaling the mesh

We make the mesh bigger while pushing it backwards, resulting in the same projection to the image



Both are valid 3D solutions for the image!

How to scale to center the mesh at a fixed depth z_0 ?

$$s = \frac{z_0}{\mu_z}, \text{ where } \mu_z = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_z^i$$

Proof: using the linearity of empirical mean, we obtain:

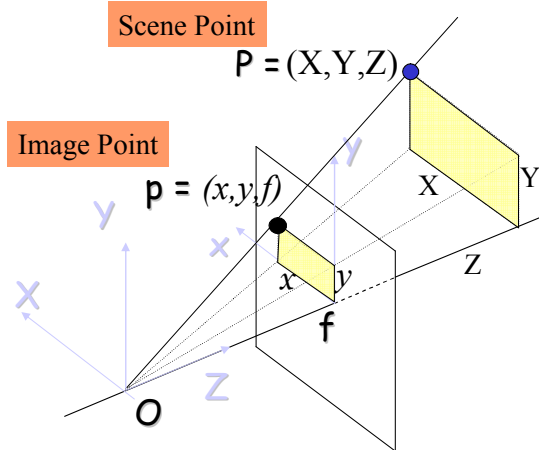
$$\mu'_z = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_z^{i'} = \frac{1}{n} \sum_i \mathbf{v}_z^i \frac{z_0}{\mu_z} = \mu_z \frac{z_0}{\mu_z} = z_0$$

Hence, the mesh is centered at z_0

Recall: perspective effects

Robert Collins
CSE486, Penn State

Basic Perspective Projection



Perspective Projection Eqns

$$x = f \frac{X}{Z} + c$$
$$y = f \frac{Y}{Z} + c$$

3 things affect the size of an object/human in the image:

- The focal length f
- The physical size of the object X, Y
- The depth Z

We want to:

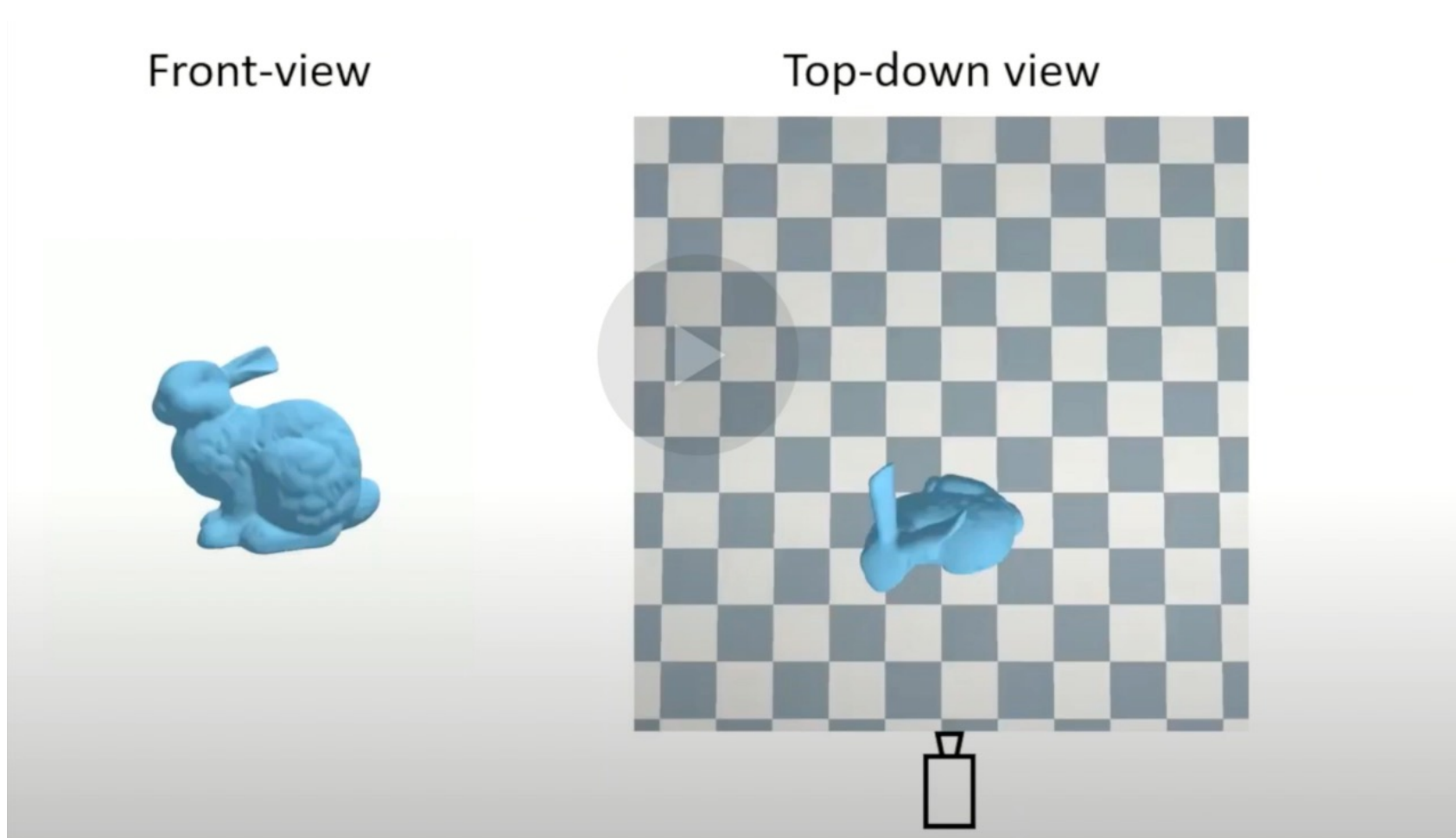
1. Change depth from $Z \mapsto z_0$
2. Maintain alignment with image, i.e. (x, y) should remain same.

We can scale the object proportionally to depth:

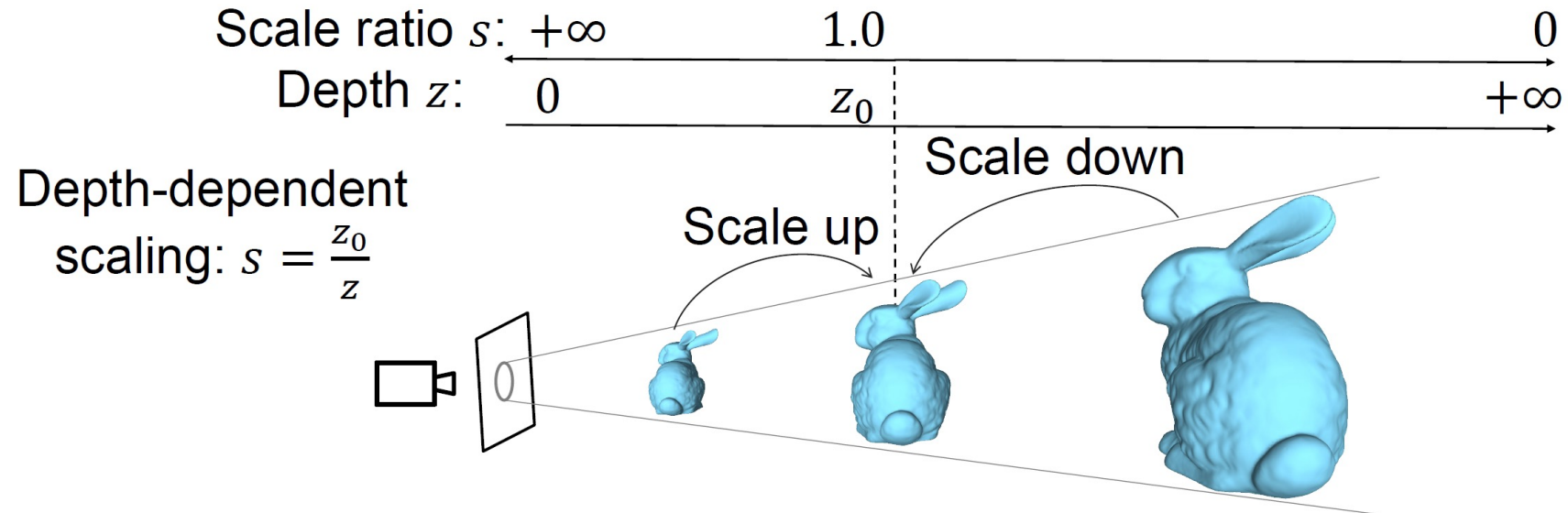
$$x \rightarrow X * \frac{z_0}{Z} \quad y \rightarrow Y * \frac{z_0}{Z} \quad Z \rightarrow z_0$$

$$x = f \frac{X \frac{z_0}{Z}}{z_0} + c = f \frac{X}{Z} + c \quad y = f \frac{Y \frac{z_0}{Z}}{z_0} + c = f \frac{Y}{Z} + c$$

Pixel aligned scaling

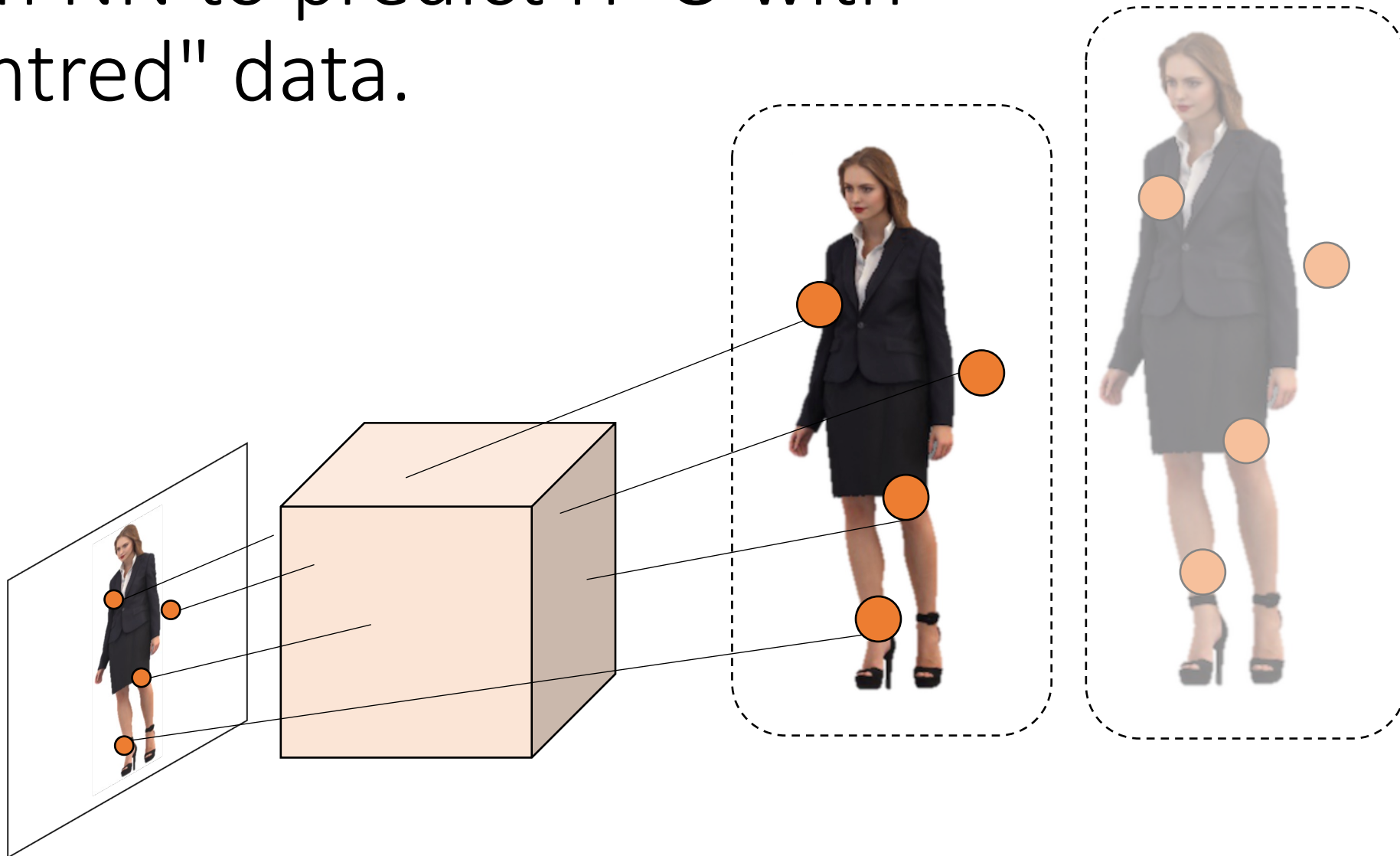


Scale aware alignment



Our scaling strategy brings uncentered object to the same depth, allowing to train neural distance fields more efficiently from *real data*.

Train NN to predict H+O with "centred" data.



CHORE Method: joint reconstruction

1. Jointly reconstruct human and object simultaneously.
2. Use network to learn spatial arrangement priors conditioned on input image, instead of hand-crafted rules.



Input Image

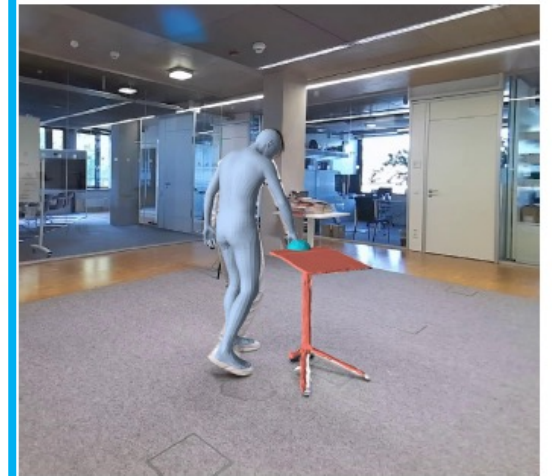
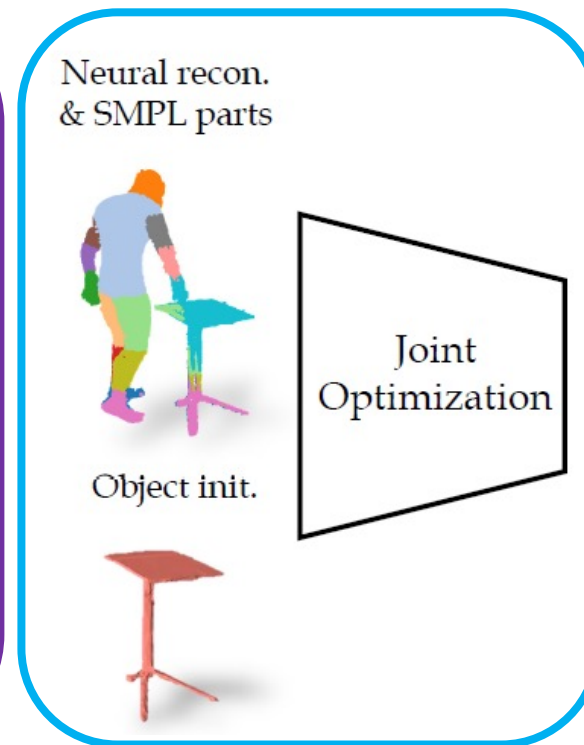
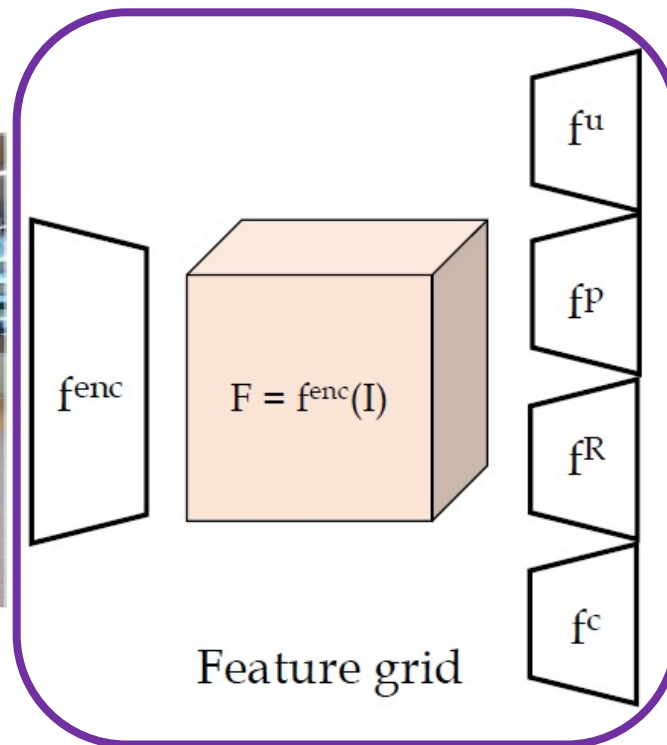
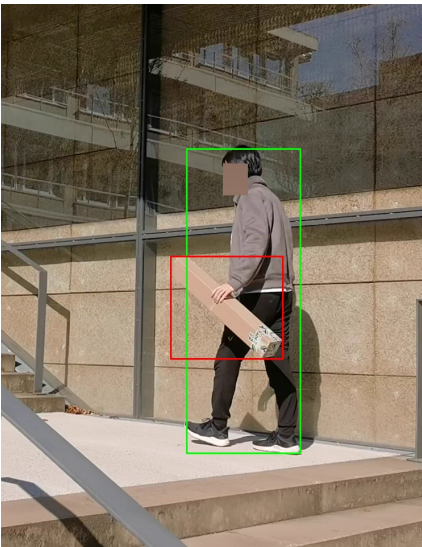


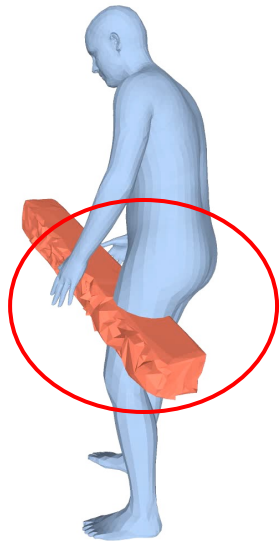
Image overlay

CHORE outperforms PHOSA both indoor and outdoor

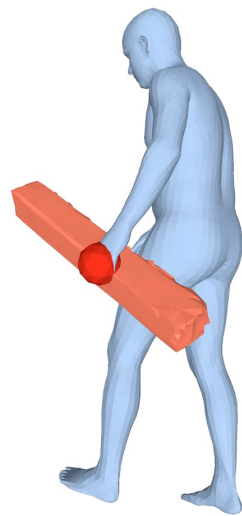
Input Image



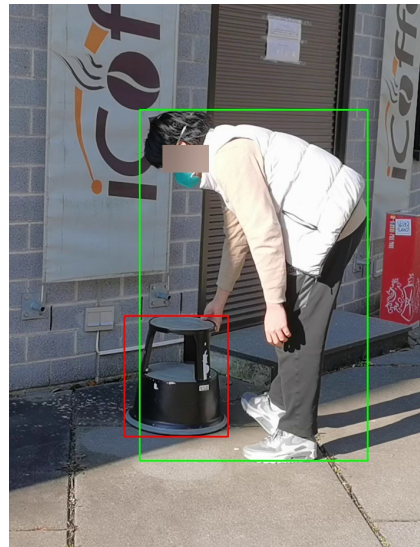
PHOSA



Ours



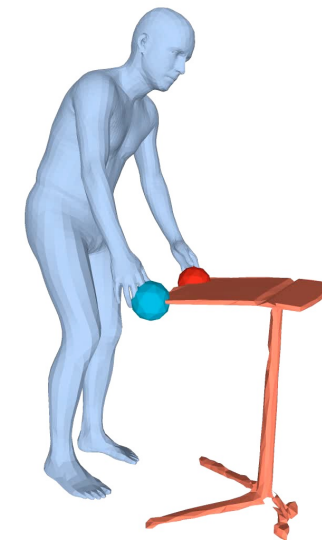
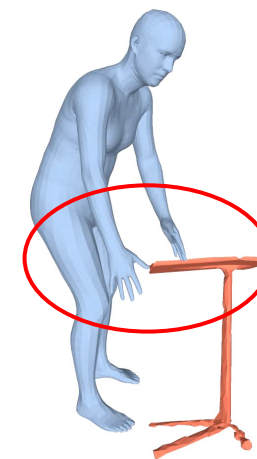
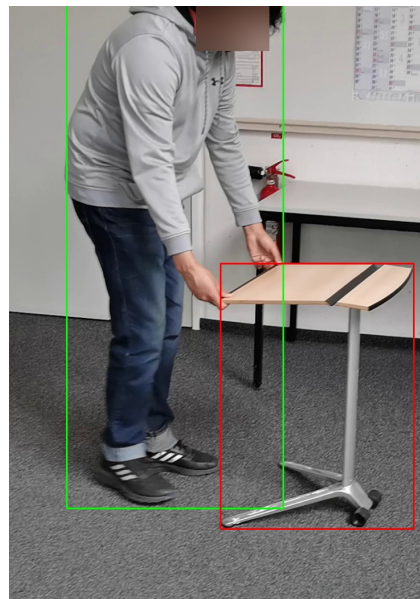
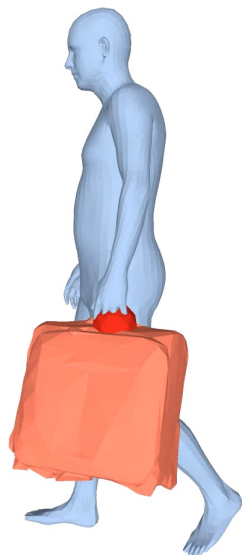
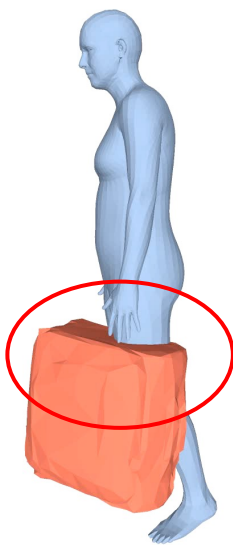
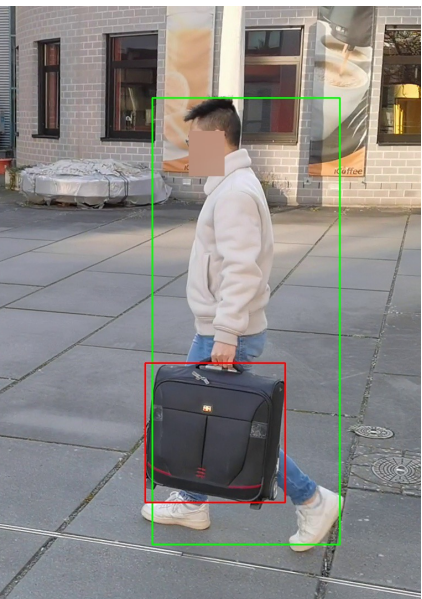
Input Image



PHOSA



Ours



PHOSA: Zhang et al. ECCV'20

CHORE produces accurate and stable results over time

Input Image

PHOSA

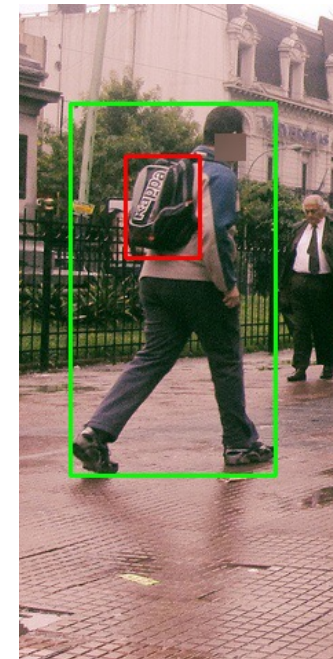
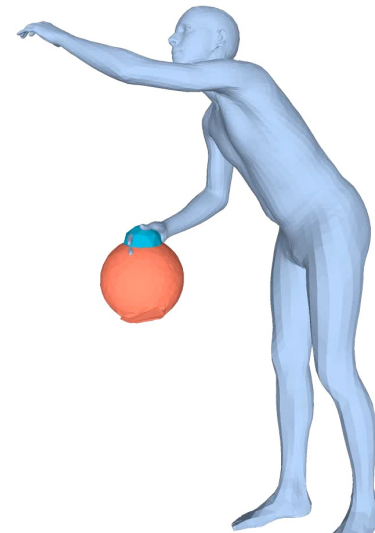
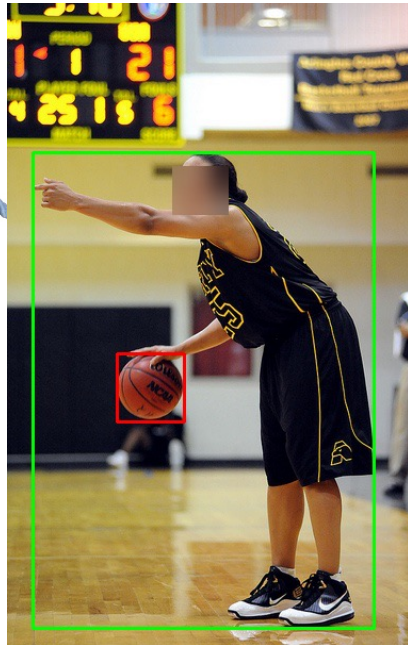
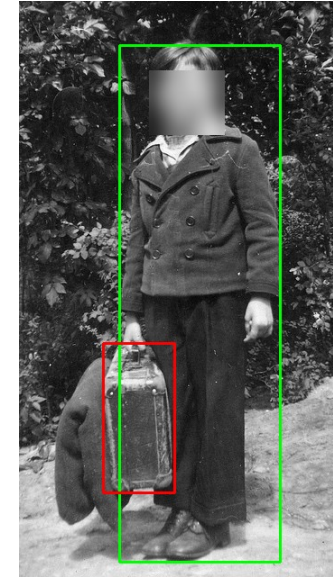
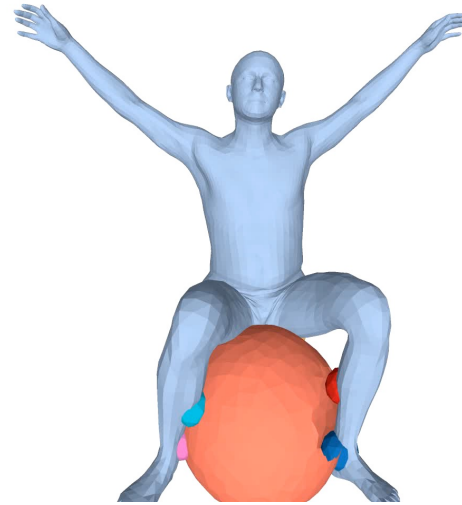
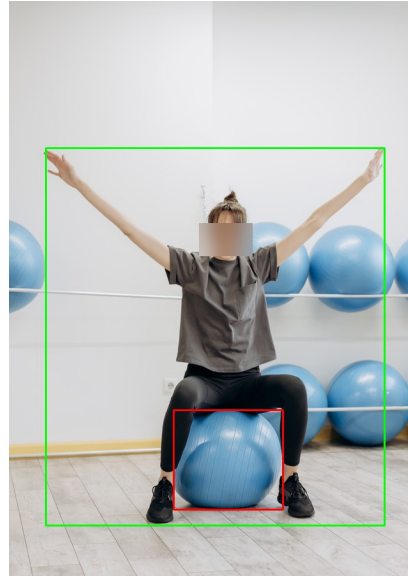
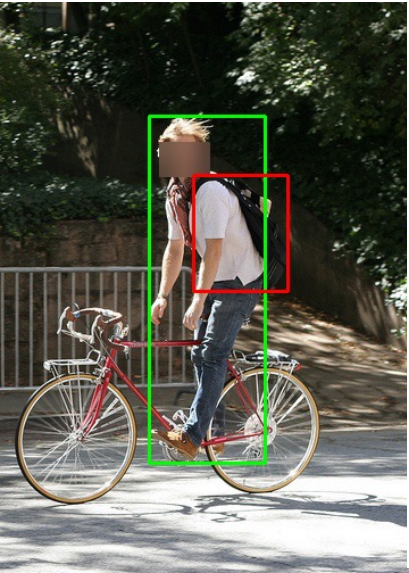
Ours



From sequence captured with phone camera. Frame by frame estimation, no temporal smoothing.

PHOSA: Zhang et al. ECCV'20

CHORE can reconstruct images in the wild



Key takeaways

- PHOSA relied on heuristics, does not scale.
- Pixel aligned learning key for reconstruction.
- Neural networks can learn a strong prior on HOI (BEHAVE, CHORE).
- Depth aware scaling can be used to avoid "depth-scale" ambiguity.