# Learning to Recover 3D Human Pose and Shape from 2D Image

**Xiaowei Zhou**

**State Key Lab of CAD&CG**
**Zhejiang University**


AR
HCI
Movie
Gaming
Health
Surveillance

Cisco: the future of shopping

# 抖音尬舞机


尬舞机
253100


07:44
App
所有 App
排行榜
付费 App
免费 App
1 抖音短视频 -...
好玩的人都在这
更新
2 支付宝 - 让生...
生活
更新
3 微信
社交
更新
4 QQ
社交
更新

# Human pose estimation

# Microsoft Xbox

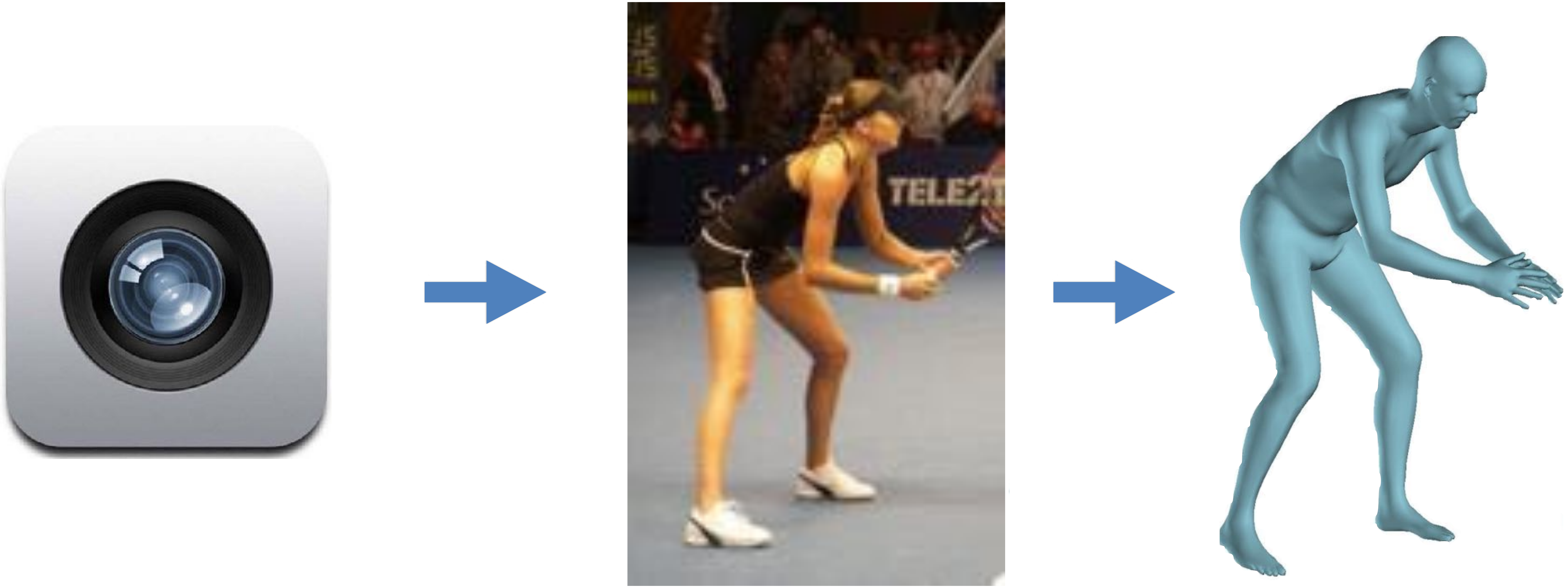
# Why Xbox is not so popular?


Expensive

Not portable

Only indoor

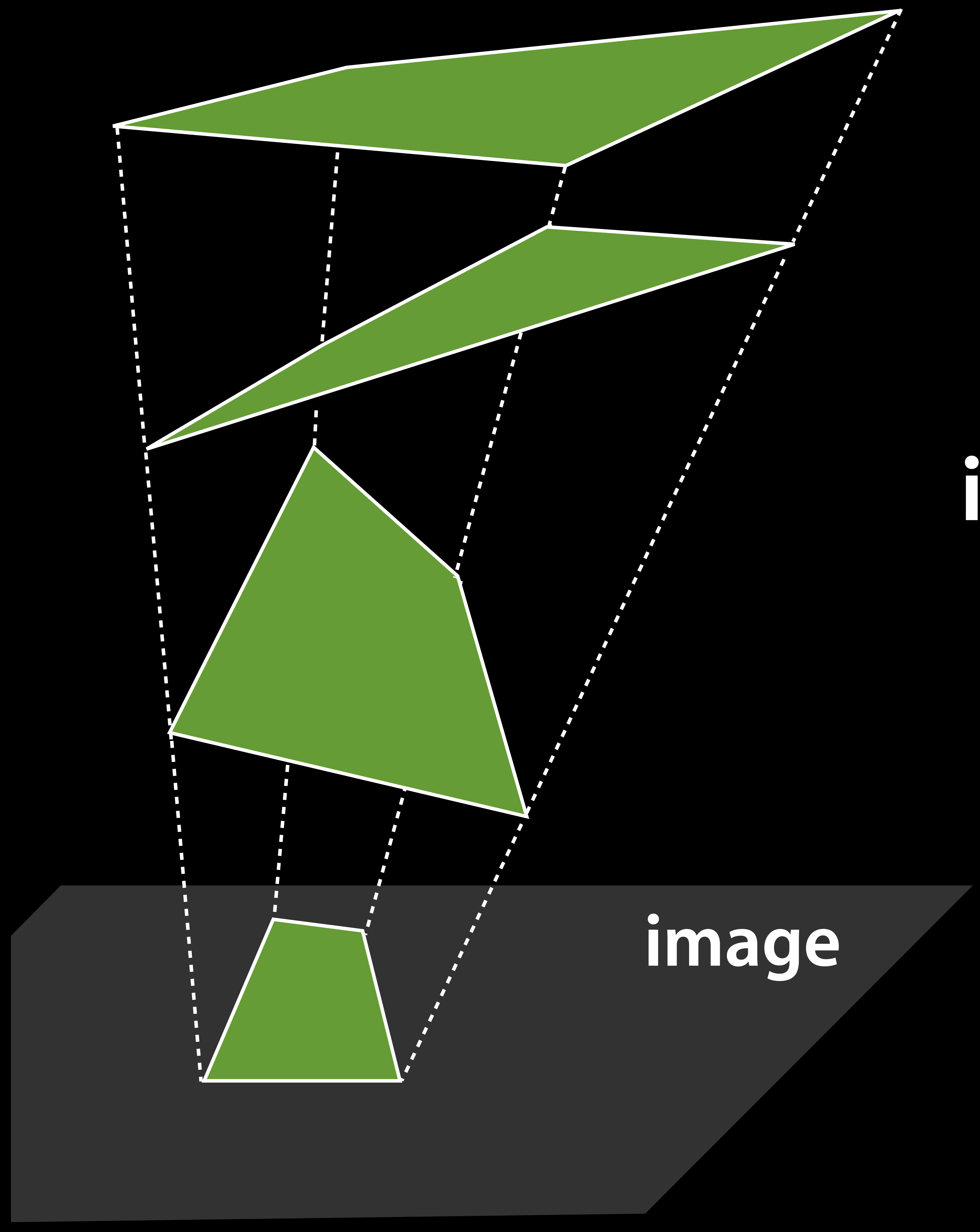
# Capture 3D pose and shape with RGB camera



# Challenge I: appearance variability

# Challenge II:  structural variability

# Challenge III: single-view ambiguity




**infinite number of possible shapes**

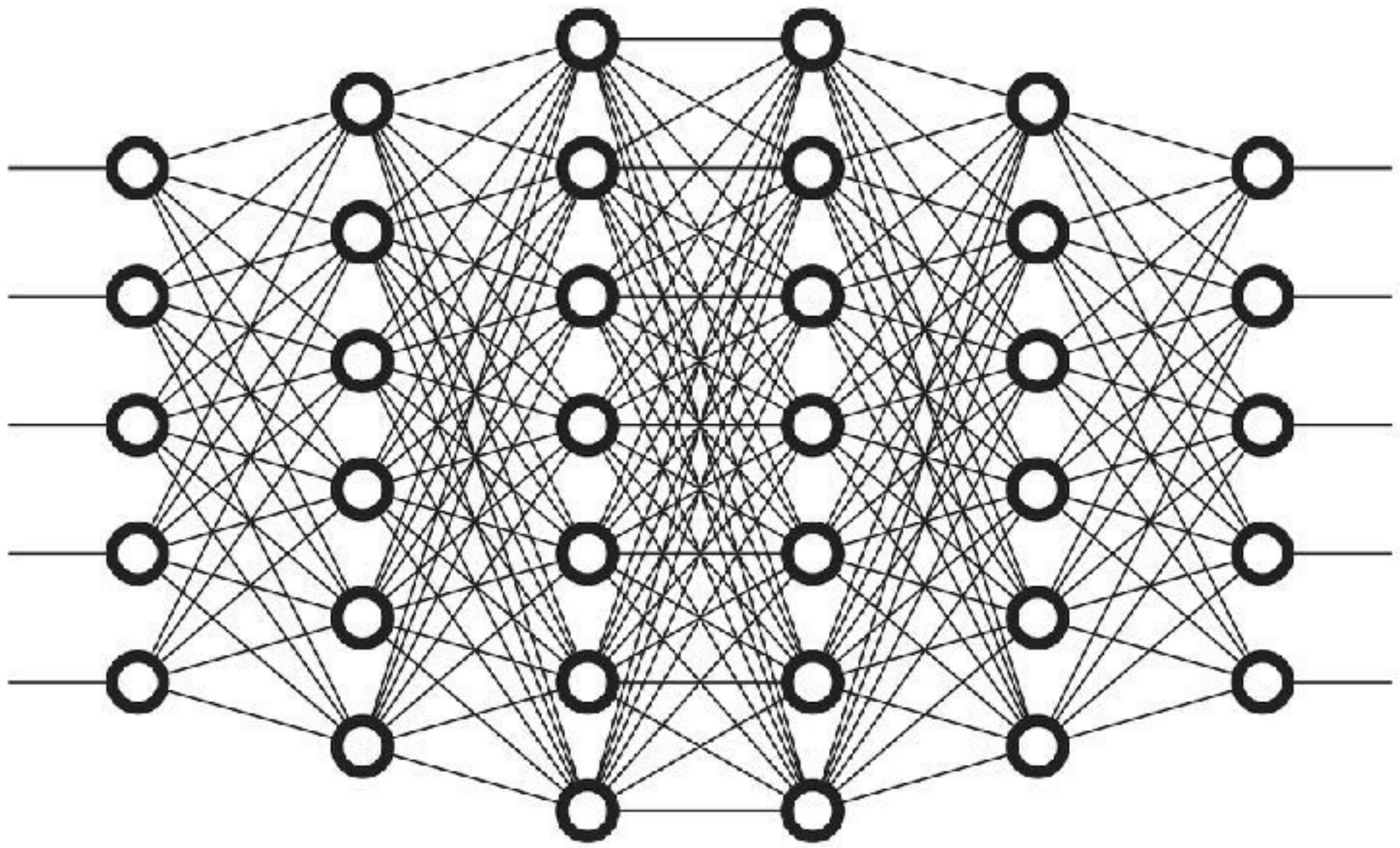
# Learning 3D geometry


MoSh datasets

3D Prior

# Deep learning

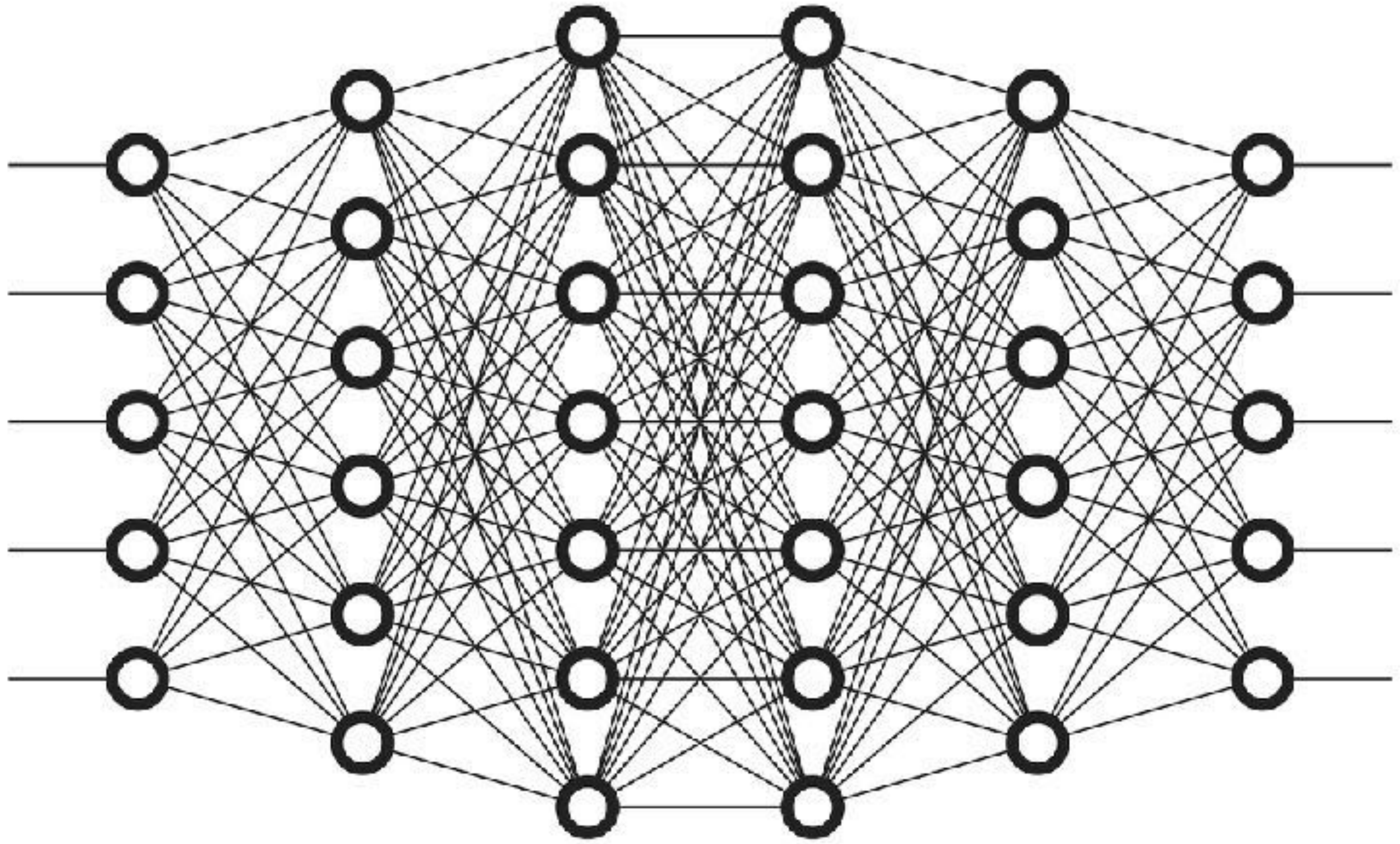
Input $X$



Output $Y$

$$\boldsymbol{Y = f(X;\theta)}$$

# Pose estimation as supervised learning



$$\begin{array}{ccc} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \\ \vdots & \vdots & \vdots \\ x_N & y_N & z_N \end{array}$$

Input image

Human pose

# Human can label 2D properties

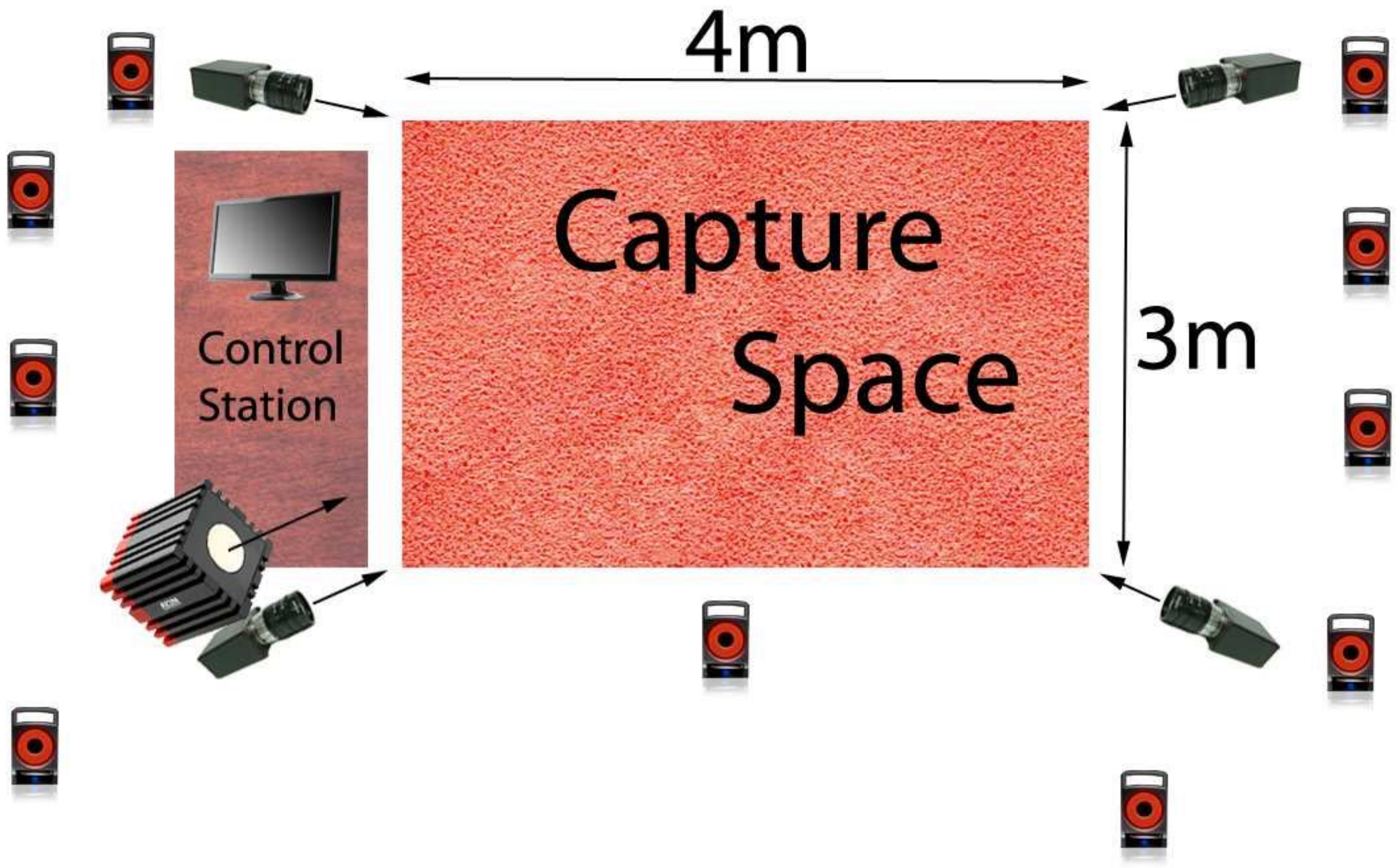


MPII Dataset (Andriluka et al., 2014)

- 25K images
- 40K poses
- 410 activities from YouTube

# Manually label 3D pose and shape ??

# Collecting training data using motion capture (MoCap)



# Domain difference

MoCap Images

In-the-wild Images

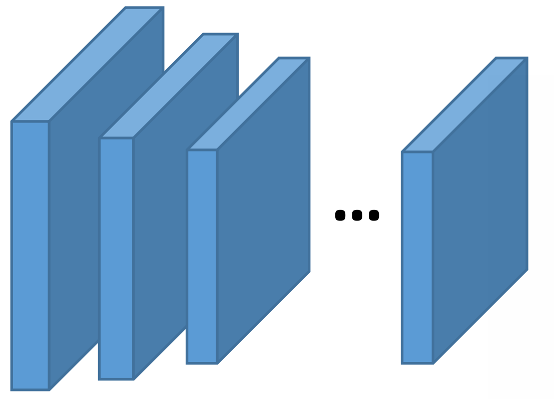
# Challenges for learning 3D pose and shape

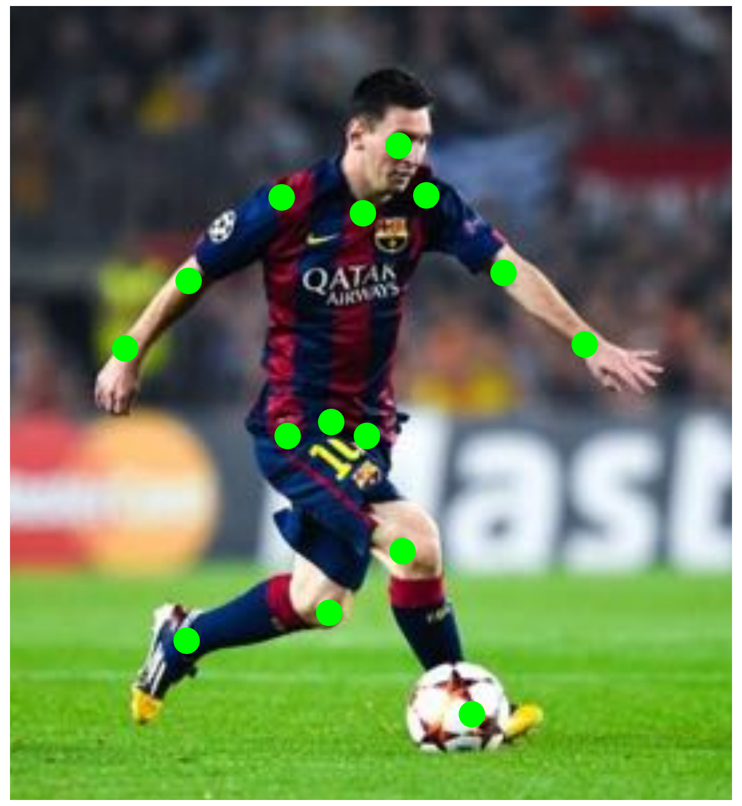
Lack of training data

Poor generalization ability

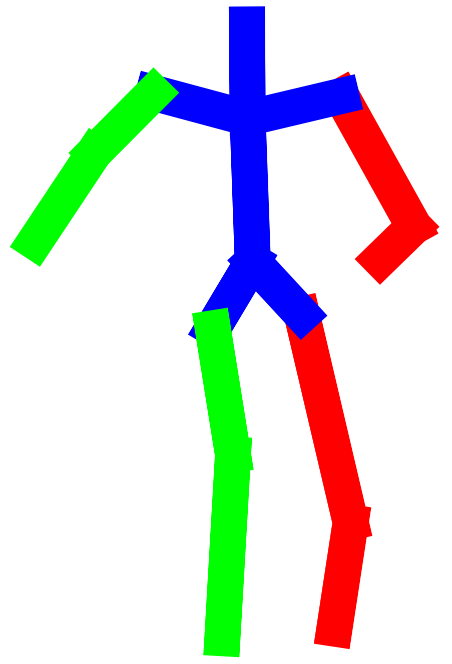
Unstructured output

# Two stage approach





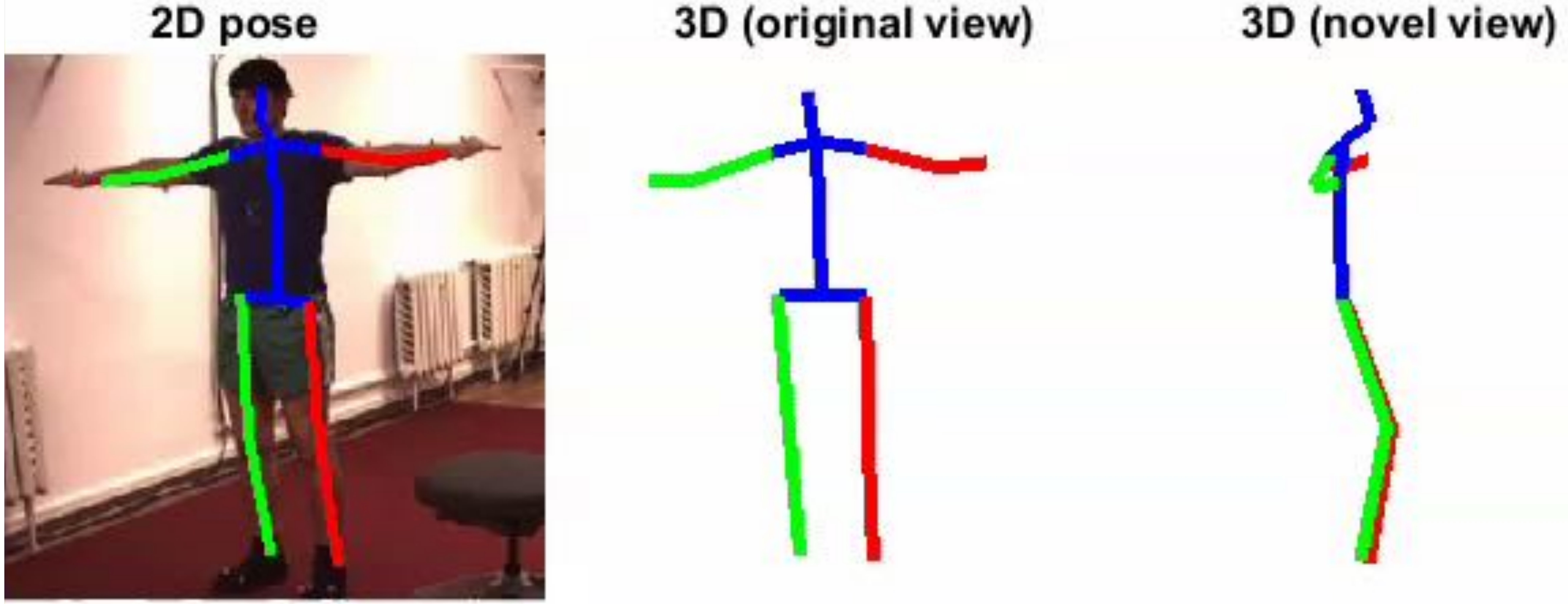
$$\min_{\boldsymbol{c},\bar{\boldsymbol{R}}} \ \frac{1}{2}\left\|\boldsymbol{W} - \bar{\boldsymbol{R}}\sum_{i=1}^{k} c_i \boldsymbol{B}_i\right\|_F^2 + \alpha\|\boldsymbol{c}\|_1$$



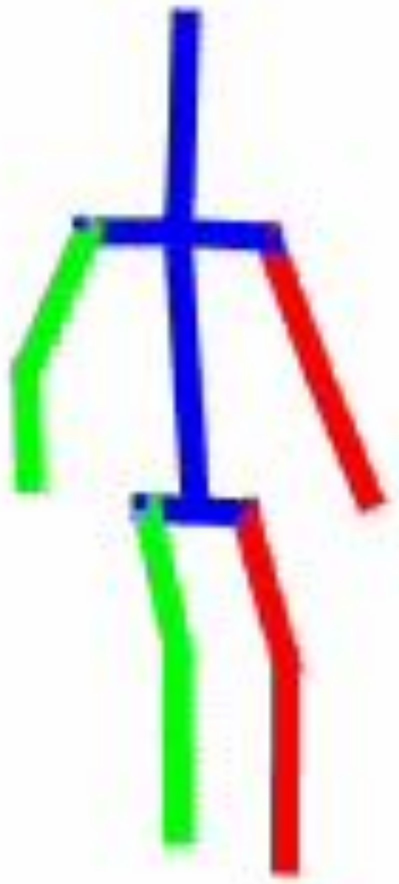
Only need 2D image-pose pairs to train 2D pose detector

Use geometric methods to lift 2D pose to 3D

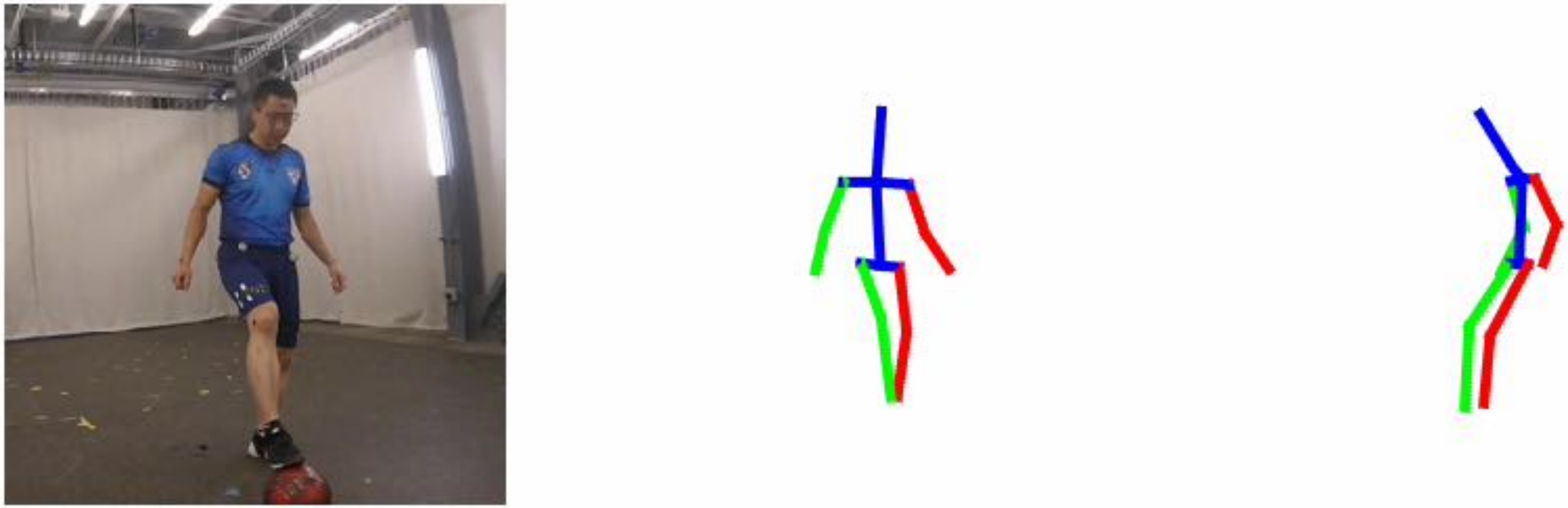
CVPR 2015
CVPR 2016


2D pose
3D (original view)
3D (novel view)

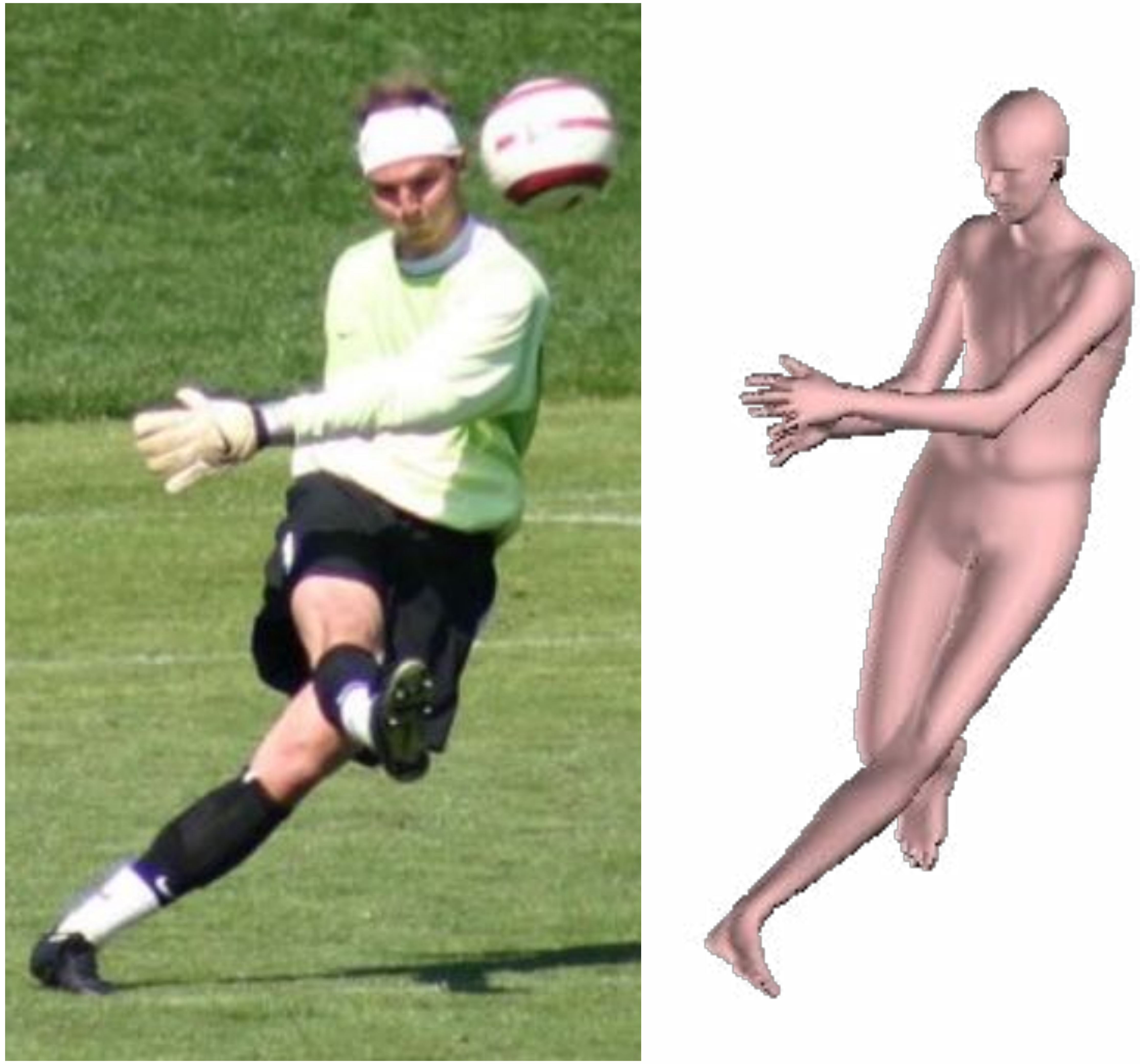

11



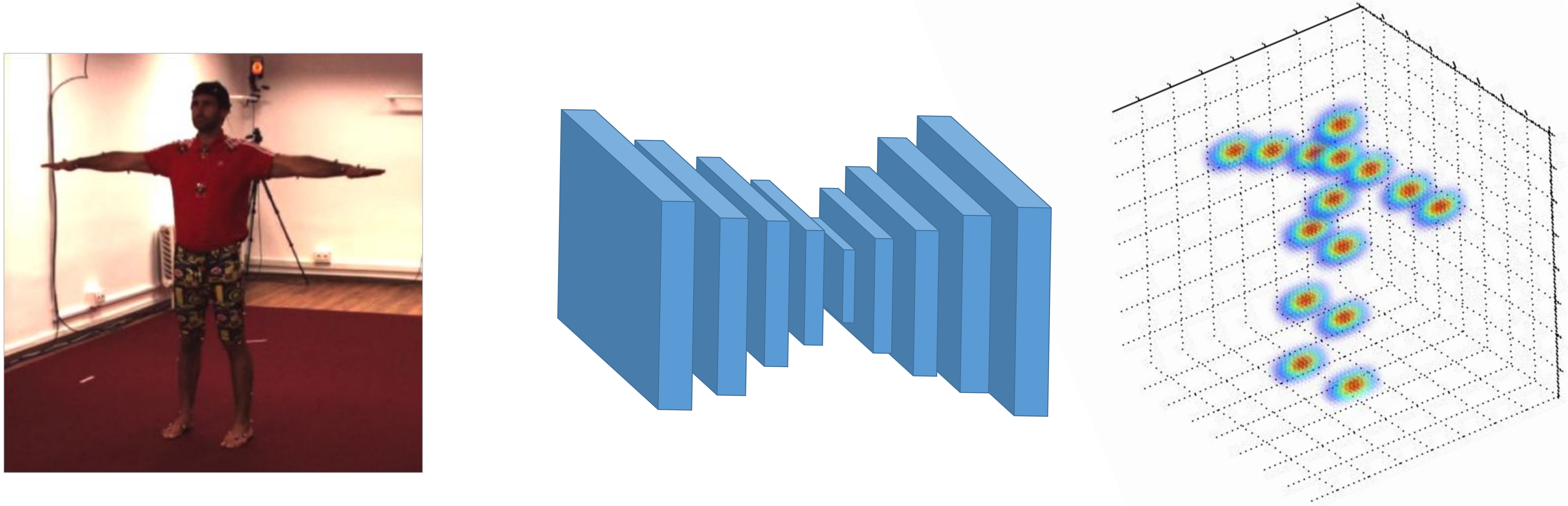
ICRA 2018



# Reconstruction ambiguity

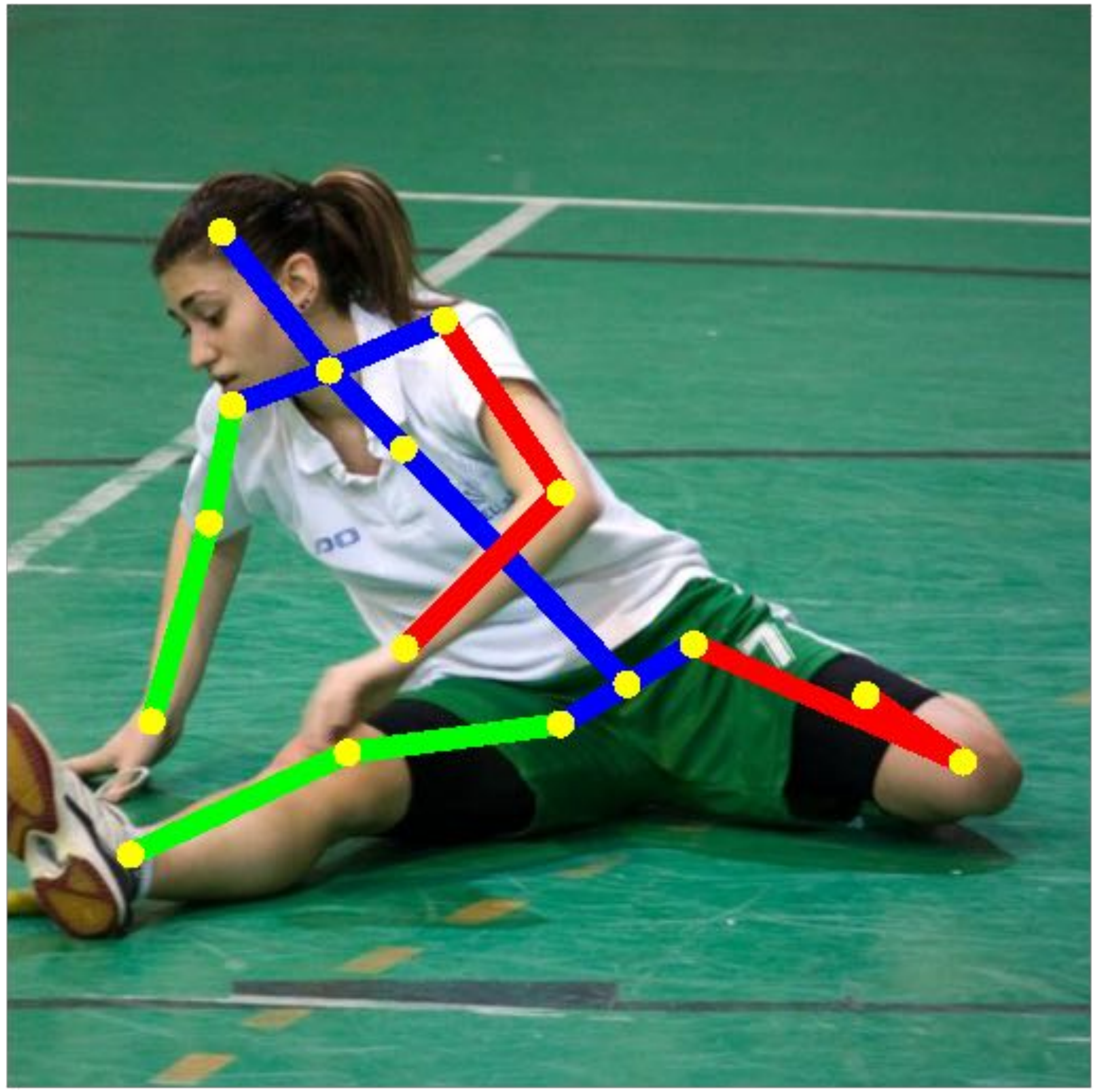


# End-to-end approach



CVPR 2017

# Using weakly annotated data



Z(left knee) > Z(right knee)

Z(right elbow) > Z(right wrist)

Z(left shoulder) < Z(right shoulder)

Z(right knee) < Z(left hip)

Z(left wrist) = Z(left elbow)

Z(head) > Z(right ankle)

Z(right hip) = Z(left hip)

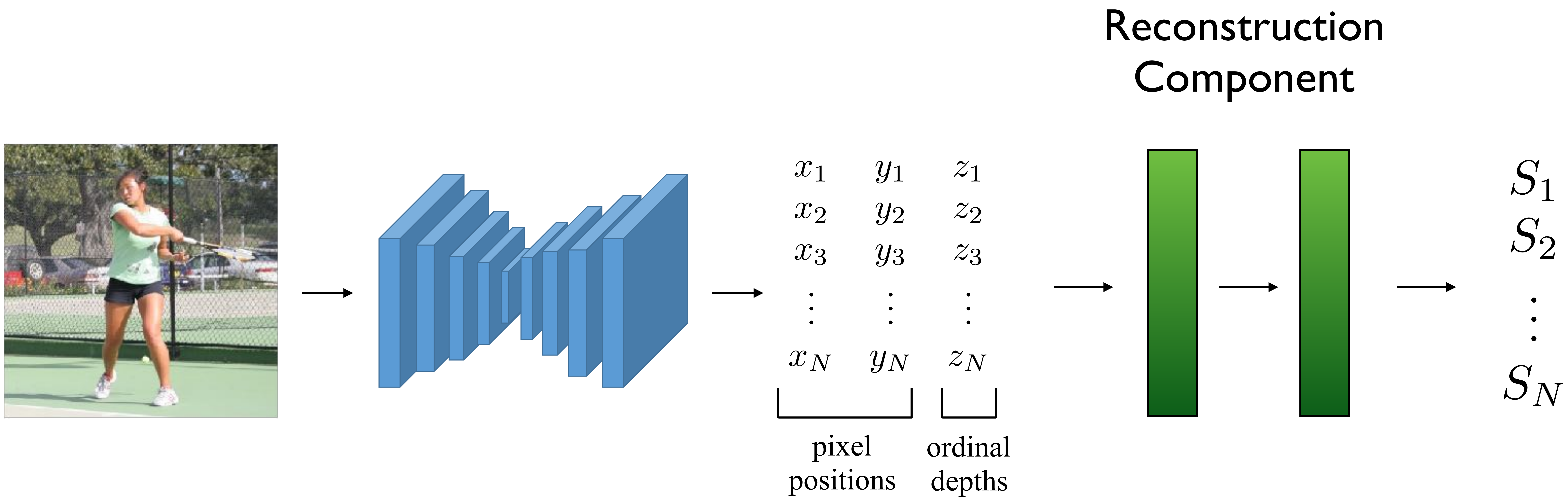
Z(right ankle) < Z(neck)

Z(left wrist) < Z(left ankle)

Humans **can** annotate ordinal depth relations.

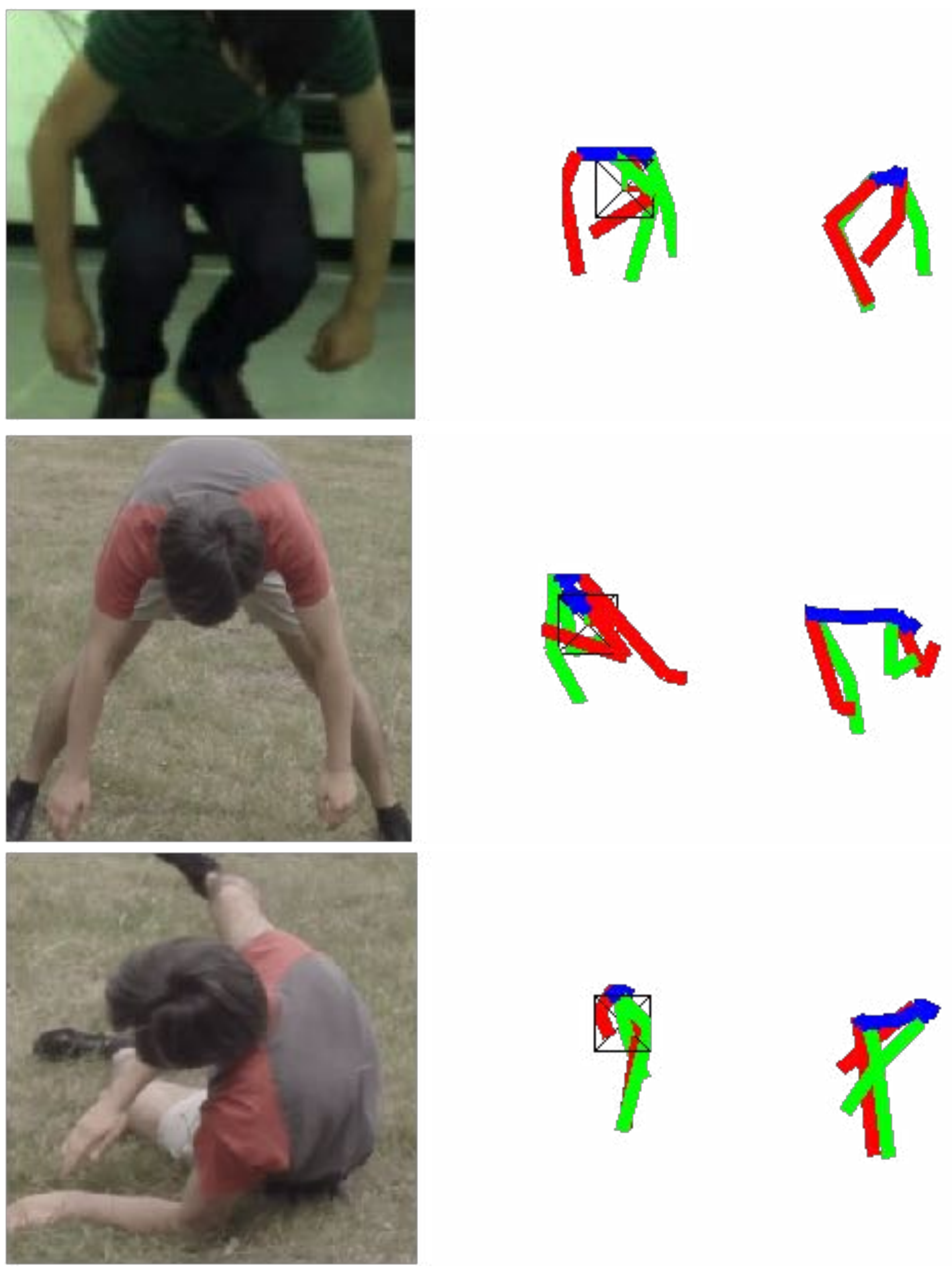
CVPR 2018

# Refinement with a reconstruction component

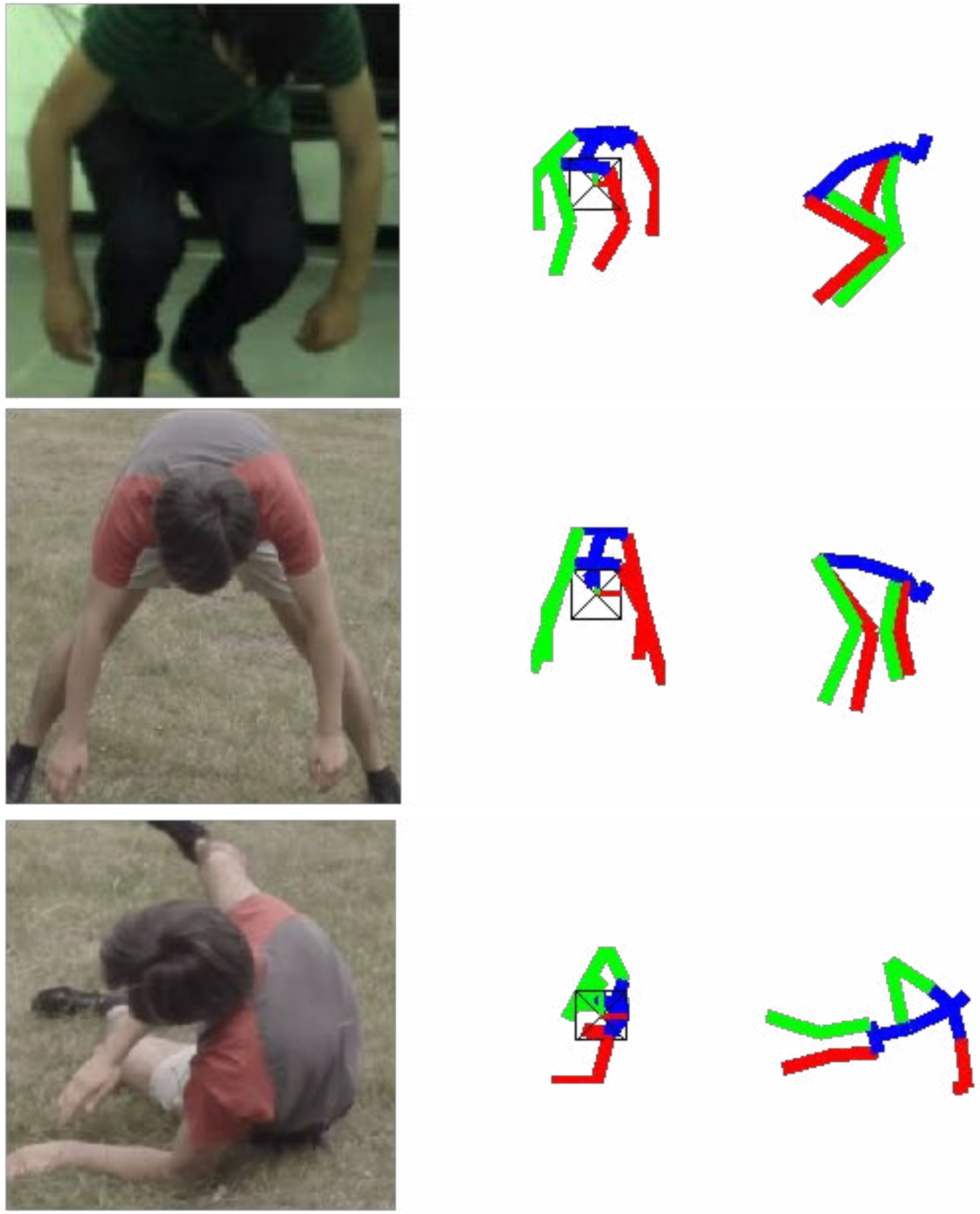



- Recovers a coherent 3D pose
- Simple multi-layer perceptron
- Trained only on MoCap data.

Only using MoCap data for training



Using MoCap + ordinal depth



# Quantitative evaluation on Human3.6M

Mean distance to ground truth per joint (mm)

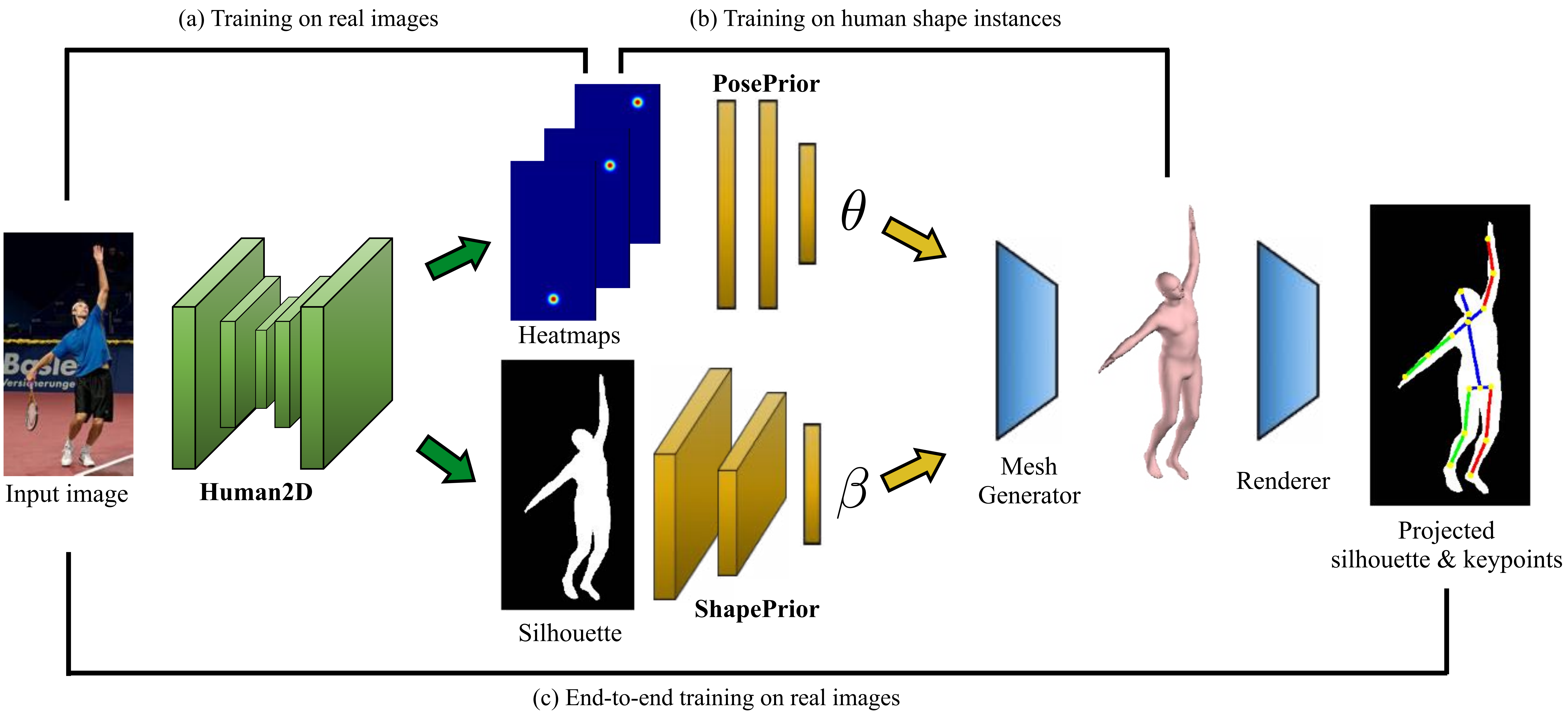
| | Direct. | Discuss | Eating | Greet | Phone | Photo | Pose | Purch. | Sitting | SitingD | Smoke | Wait | WalkD | Walk | WalkT | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tekin *et al.* [49] (CVPR'16) | 102.4 | 147.2 | 88.8 | 125.3 | 118.0 | 182.7 | 112.4 | 129.2 | 138.9 | 224.9 | 118.4 | 138.8 | 126.3 | 55.1 | 65.8 | 125.0 |
| Zhou *et al.* [68] (CVPR'16) | 87.4 | 109.3 | 87.1 | 103.2 | 116.2 | 143.3 | 106.9 | 99.8 | 124.5 | 199.2 | 107.4 | 118.1 | 114.2 | 79.4 | 97.7 | 113.0 |
| Du *et al.* [14] (ECCV'16) | 85.1 | 112.7 | 104.9 | 122.1 | 139.1 | 135.9 | 105.9 | 166.2 | 117.5 | 226.9 | 120.0 | 117.7 | 137.4 | 99.3 | 106.5 | 126.5 |
| Zhou *et al.* [66] (ECCVW'16) | 91.8 | 102.4 | 96.7 | 98.8 | 113.4 | 125.2 | 90.0 | 93.8 | 132.2 | 159.0 | 107.0 | 94.4 | 126.0 | 79.0 | 99.0 | 107.3 |
| Chen *et al.* [10] (CVPR'17) | 89.9 | 97.6 | 90.0 | 107.9 | 107.3 | 139.2 | 93.6 | 136.1 | 133.1 | 240.1 | 106.7 | 106.2 | 114.1 | 87.0 | 90.6 | 114.2 |
| Tome *et al.* [51] (CVPR'17) | 65.0 | 73.5 | 76.8 | 86.4 | 86.3 | 110.7 | 68.9 | 74.8 | 110.2 | 173.9 | 85.0 | 85.8 | 86.3 | 71.4 | 73.1 | 88.4 |
| Rogez *et al.* [40] (CVPR'17) | 76.2 | 80.2 | 75.8 | 83.3 | 92.2 | 105.7 | 79.0 | 71.7 | 105.9 | 127.1 | 88.0 | 83.7 | 86.6 | 64.9 | 84.0 | 87.7 |
| Pavlakos *et al.* [32] (CVPR'17) | 67.4 | 71.9 | 66.7 | 69.1 | 72.0 | 77.0 | 65.0 | 68.3 | 83.7 | 96.5 | 71.7 | 65.8 | 74.9 | 59.1 | 63.2 | 71.9 |
| Nie *et al.* [60] (ICCV'17) | 90.1 | 88.2 | 85.7 | 95.6 | 103.9 | 103.0 | 92.4 | 90.4 | 117.9 | 136.4 | 98.5 | 94.4 | 90.6 | 86.0 | 89.5 | 97.5 |
| Tekin *et al.* [48] (ICCV'17) | 54.2 | 61.4 | 60.2 | 61.2 | 79.4 | 78.3 | 63.1 | 81.6 | 70.1 | 107.3 | 69.3 | 70.3 | 74.3 | 51.8 | 74.3 | 69.7 |
| Zhou *et al.* [64] (ICCV'17) | 54.8 | 60.7 | 58.2 | 71.4 | 62.0 | 65.5 | 53.8 | 55.6 | 75.2 | 111.6 | 64.2 | 66.1 | 51.4 | 63.2 | 55.3 | 64.9 |
| Martinez *et al.* [25] (ICCV'17) | 51.8 | 56.2 | 58.1 | 59.0 | 69.5 | 78.4 | 55.2 | 58.1 | 74.0 | 94.6 | 62.3 | 59.1 | 65.1 | 49.5 | 52.4 | 62.9 |
| Ours | **48.5** | **54.4** | **54.4** | **52.0** | **59.4** | **65.3** | **49.9** | **52.9** | **65.8** | **71.1** | **56.6** | **52.9** | **60.9** | **44.7** | **47.8** | **56.2** |

# Predicting pose & shape

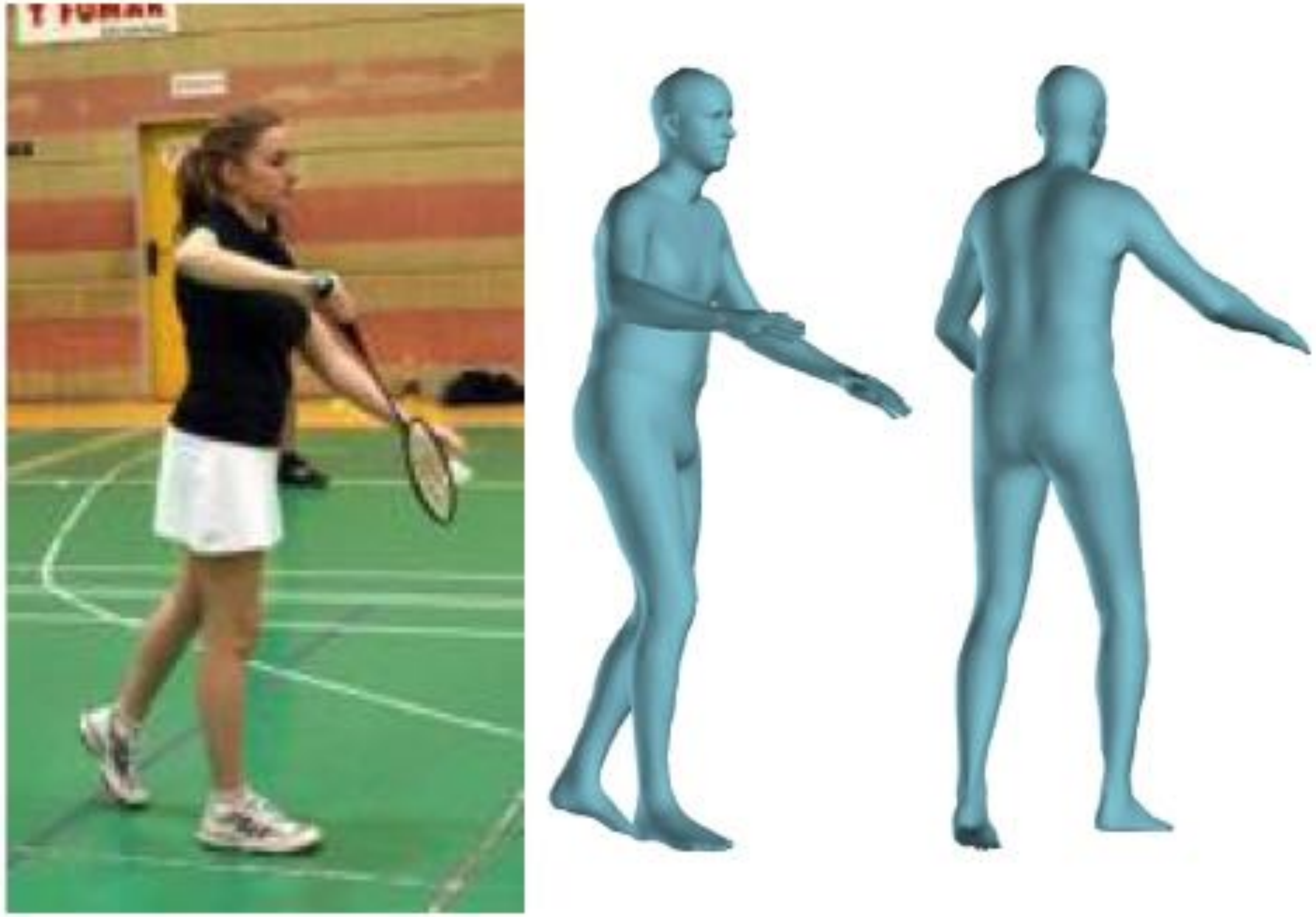
Stickman figures are nice…

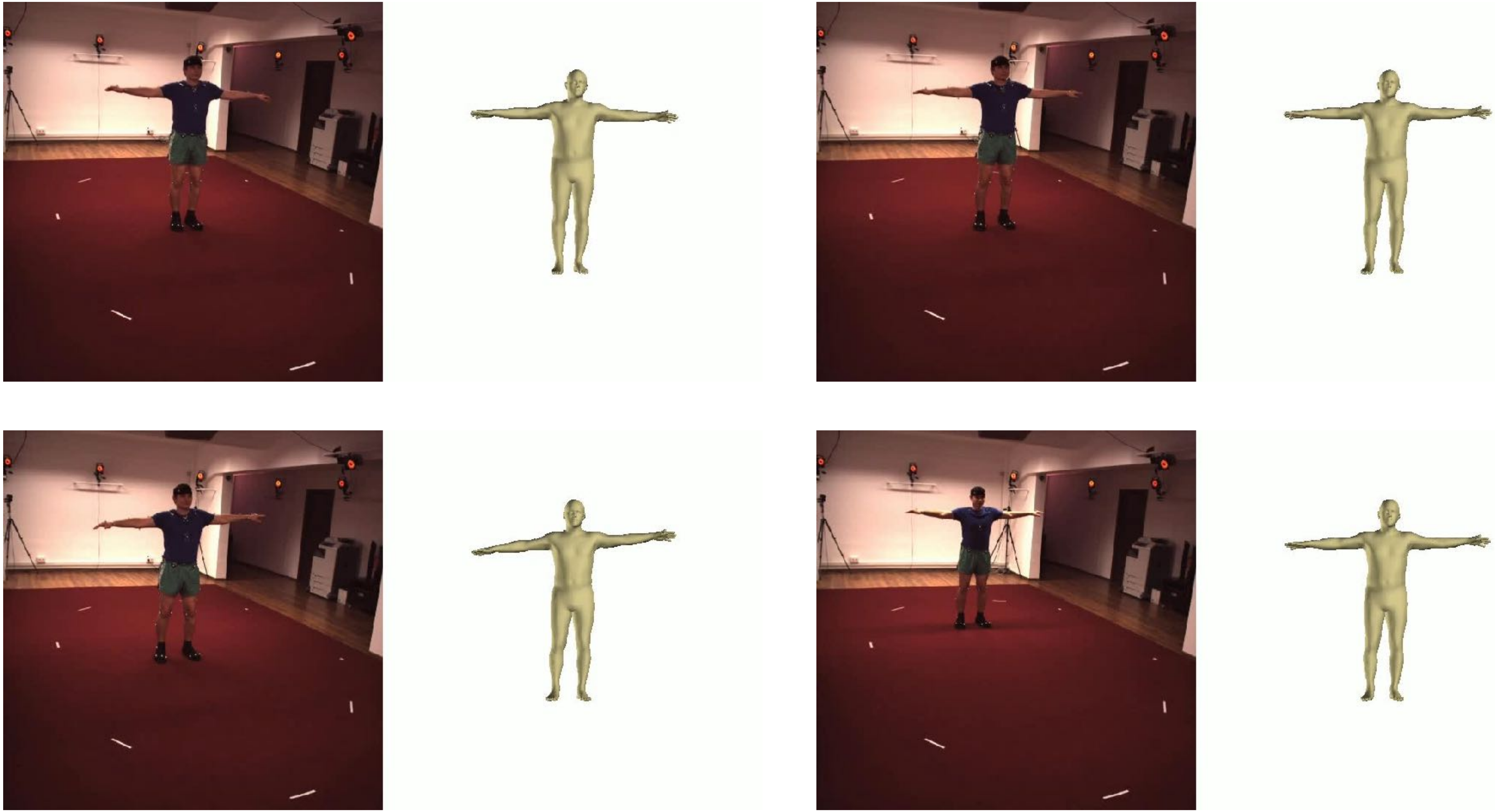
…but full pose and shape is even better.

# Integrating a statistical shape model into CNNs




CVPR 2018





# Summary

3D human sensing is important, interesting and challenging

3D from single view is possible with learning-based methods

But deep learning cannot solve everything and we still need geometry

欢迎硕士、博士、博士后加入浙大CAD实验室三维视觉小组