Deep Learning for Human Sensing from Visual Data

Grégory Rogez
CV team – NAVER LABS Europe

NAVER

CONTENTS

1. The problem, its applications and challenges

Human sensing from images and videos

1 person: 1 girl

Interaction: Waving & smiling

Activity: Greeting

8 persons: 8 boys

Interaction: Kicking football

Activity: Playing football

8 persons: 4 male, 4 female

Interaction: Holding glasses

Activity: Toasting

3D human pose estimation

DEVIEW 2019

The goal is to localize human body keypoints (joints) in 3D space.

Figure: courtesy of Gyeongsik Moon et al, ICCV 2019

Why is it interesting?

- A large proportion of visual content on the web contains humans
- Many possible applications including:

-HCI, gaming, AR / VR

-Dancing / sport analysis

Why is it interesting?

- A large proportion of visual content on the web contains humans
- Many possible applications including:

- Human - Robot interactions
- Learning from demonstration
- Surveillance, safety

© 2019 NAVER LABS. All rights reserved.

Why is it difficult?

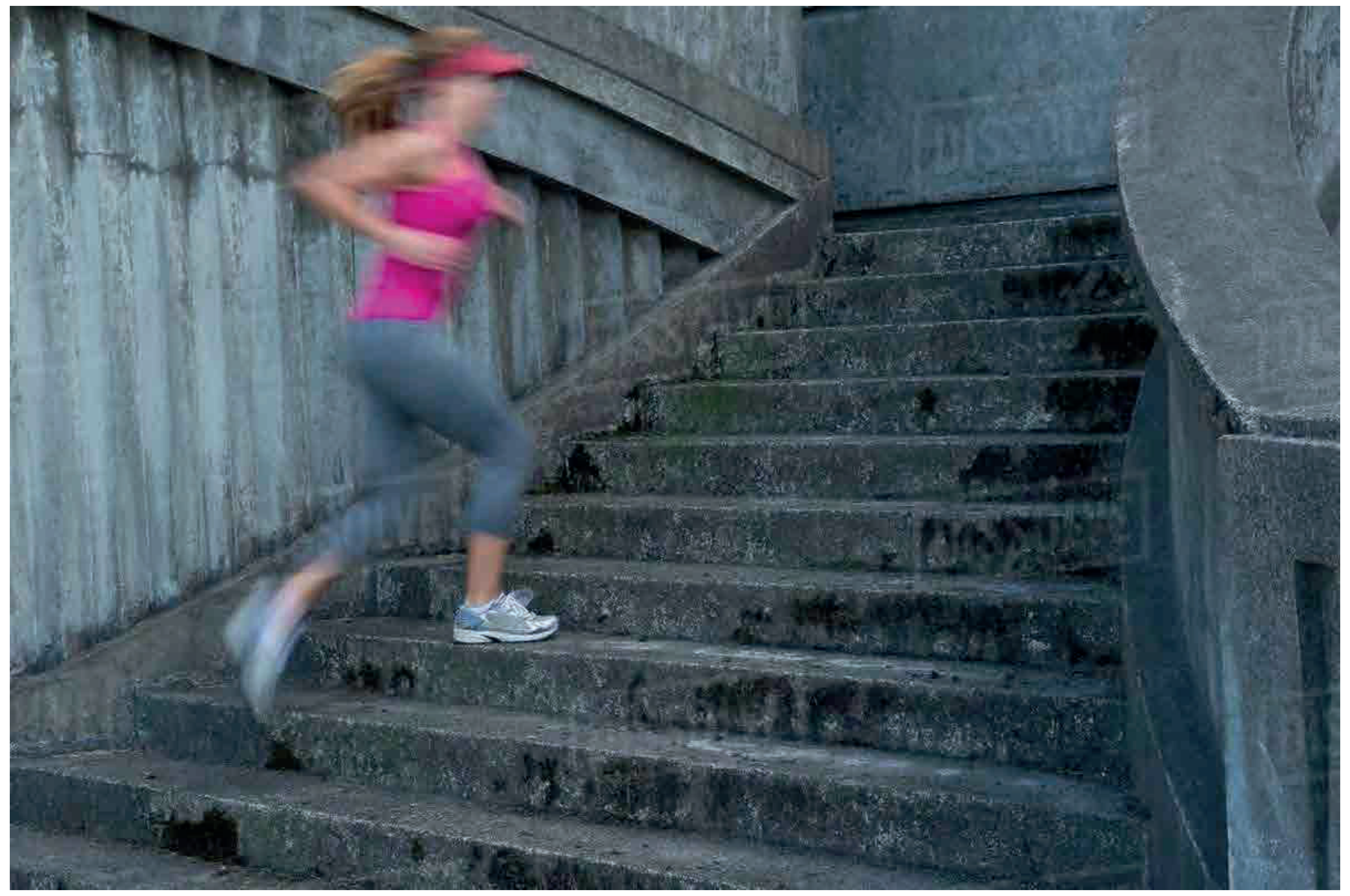
variation in illumination

variation in appearance

variation in pose, viewpoint

body part foreshortening

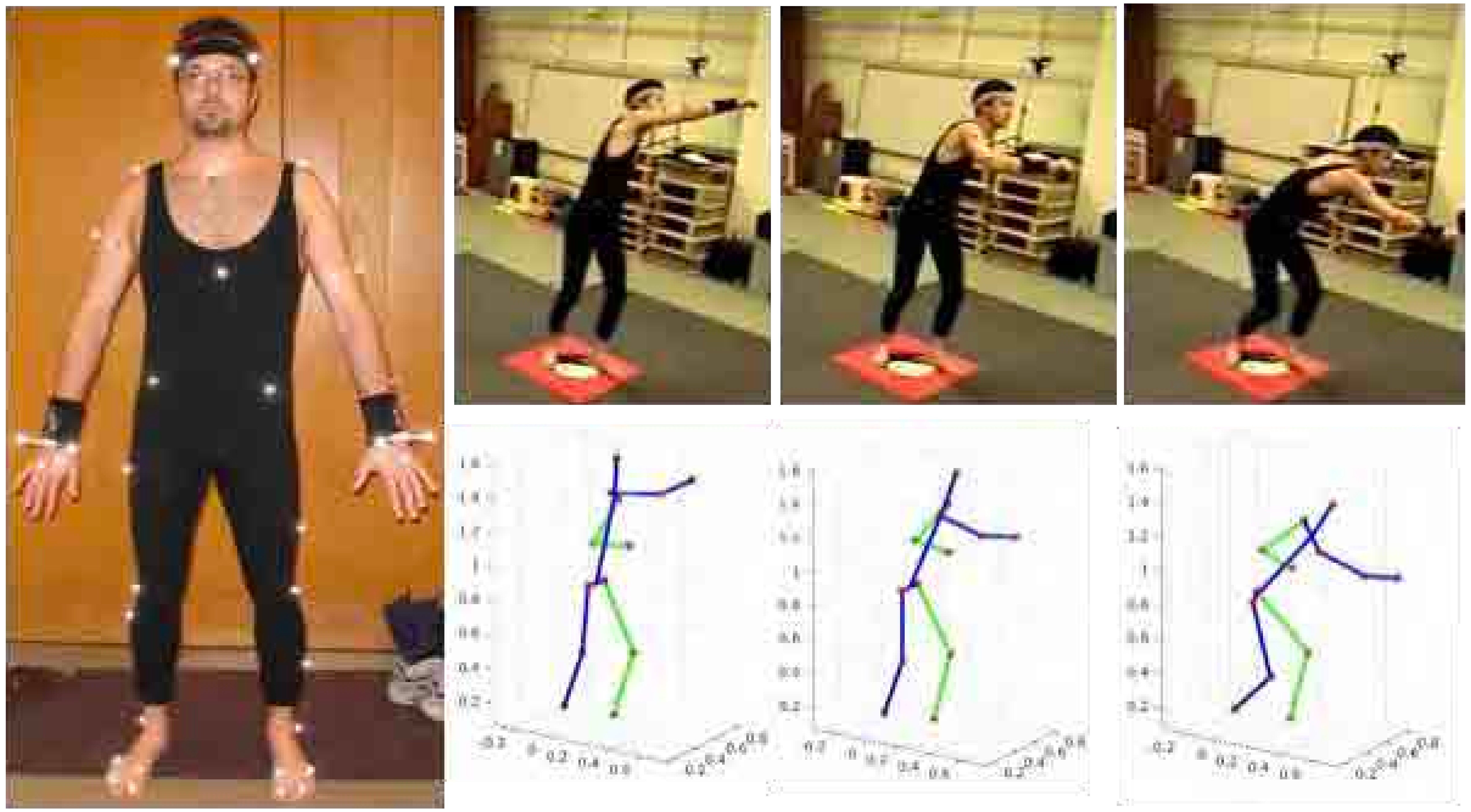
occlusion & clutter



motion blur

Why is it difficult?

Lack of data, i.e. images with 3D annotations

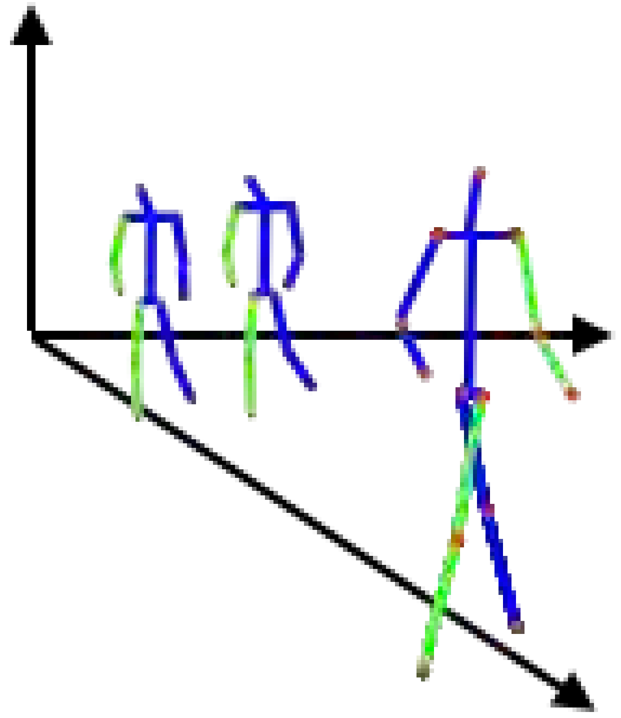


Accurate 3D data in constrained environments
(Motion Capture Room)

In real-world images, only manually
labelled 2D data

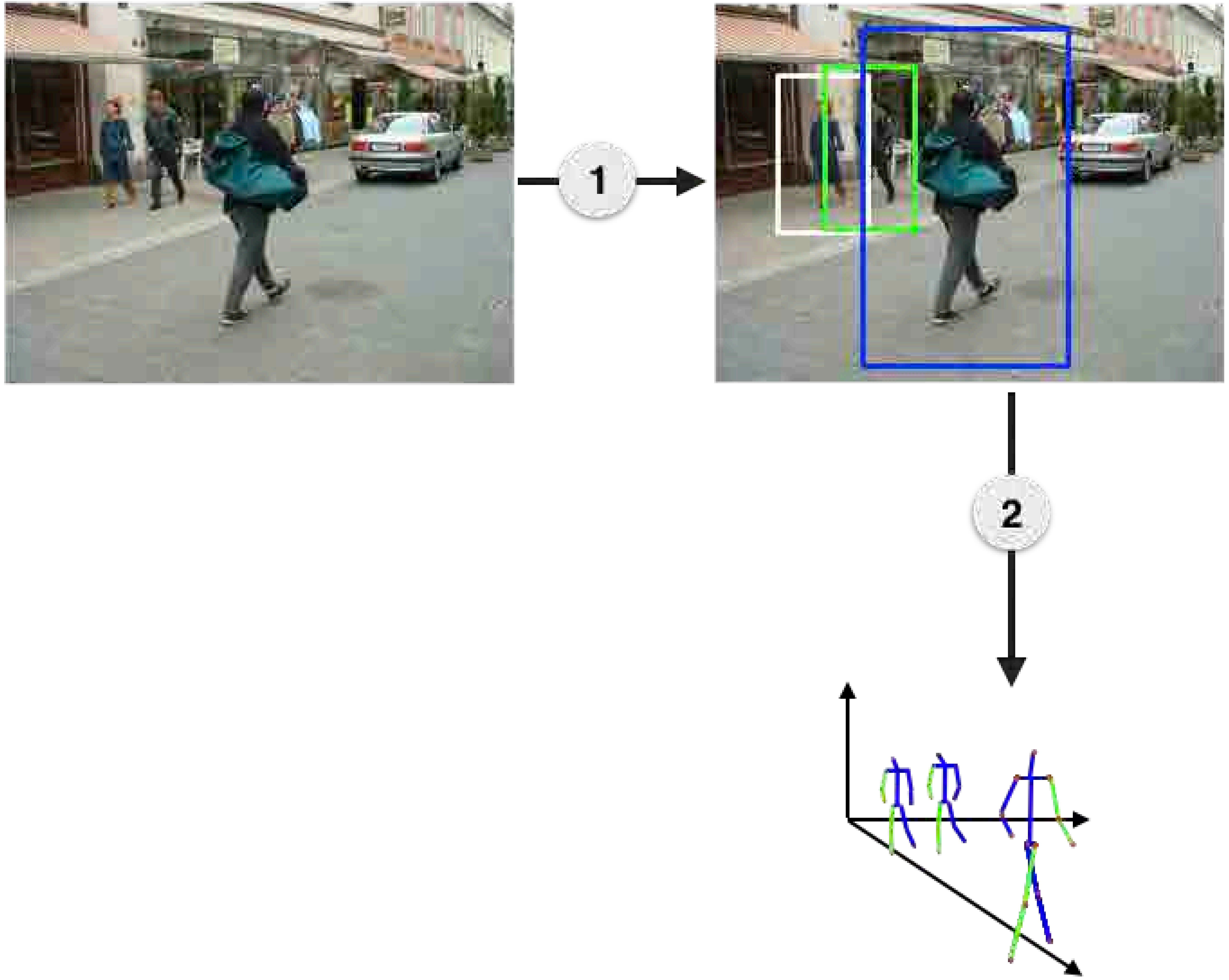
2. The state-of-the-art

Several possible paths from an input image to 3D pose...



3D pose

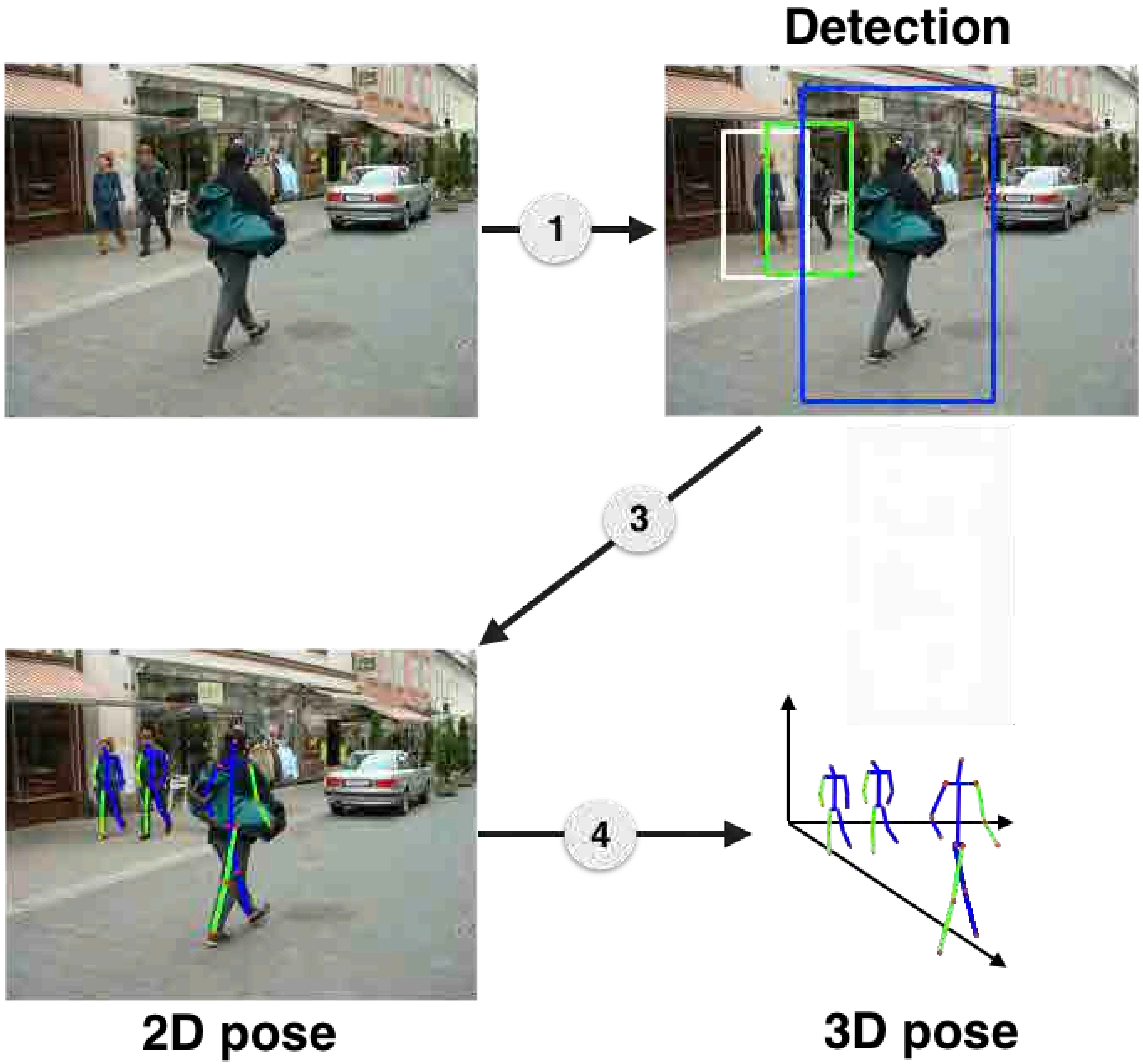
Detection



3D pose

1 [Dalal & Triggs, CVPR’05]

2 [Li et al, ICCV’15, Tekin et al, Zhou et al, CVPR’16]

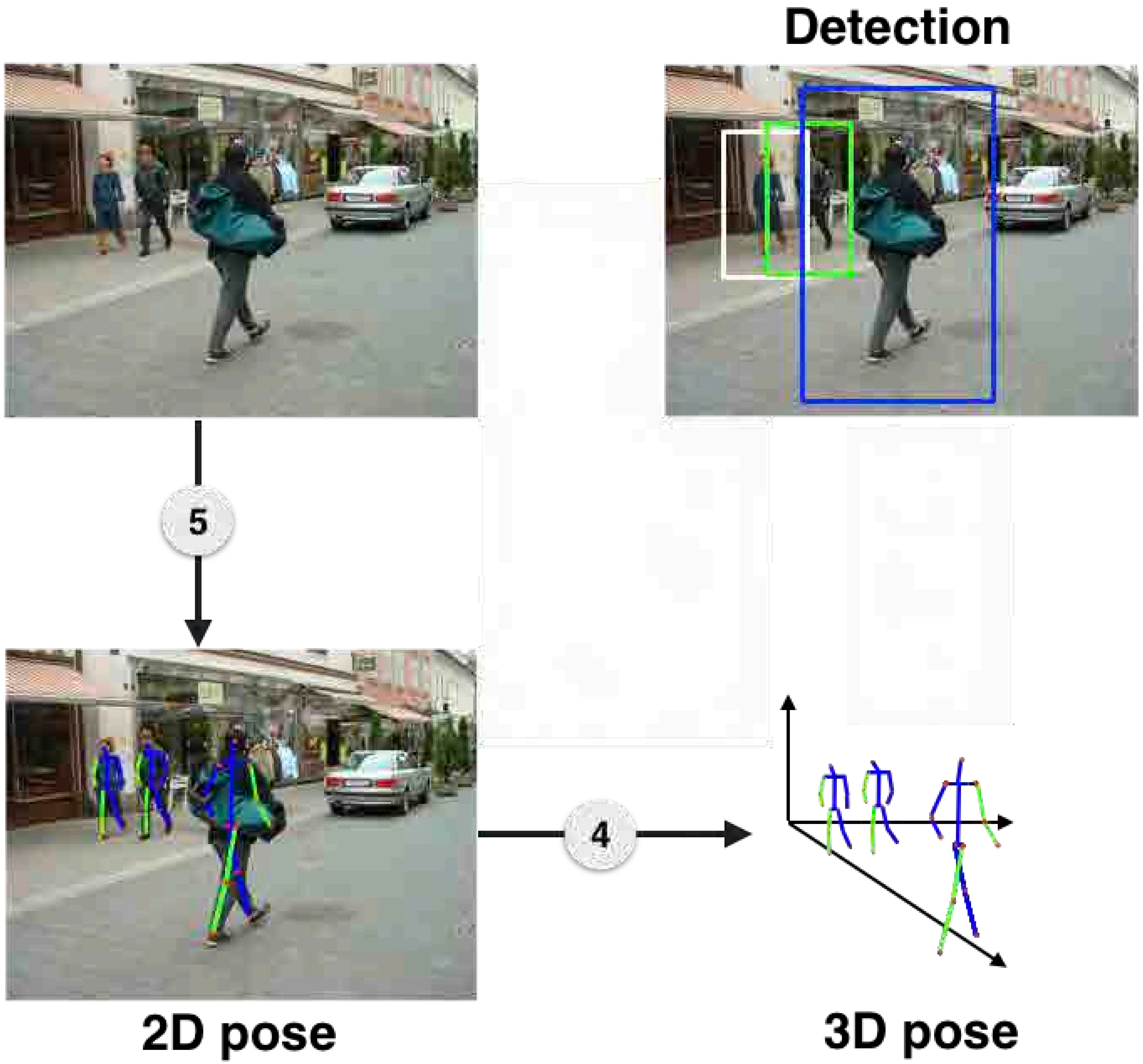


1 [Dalal & Triggs, CVPR’05]

2 [Li et al, ICCV’15, Tekin et al, Zhou et al, CVPR’16]

3 [**ECCV’16:** Newell et al., Insafutdinov et al., Gkioxary et al., Lifshitz et al., Bulat &Tzimiropoulos
CVPR’16: Wei et al, Yang et al, Pishchulin et al, Hu& Ramanan, Carreira et al.,]

4 [Akhter & Black, CVPR’15, Zhou et al., CVPR’15, Bogo et al., ECCV’16, Martinez et al., CVPR’17]



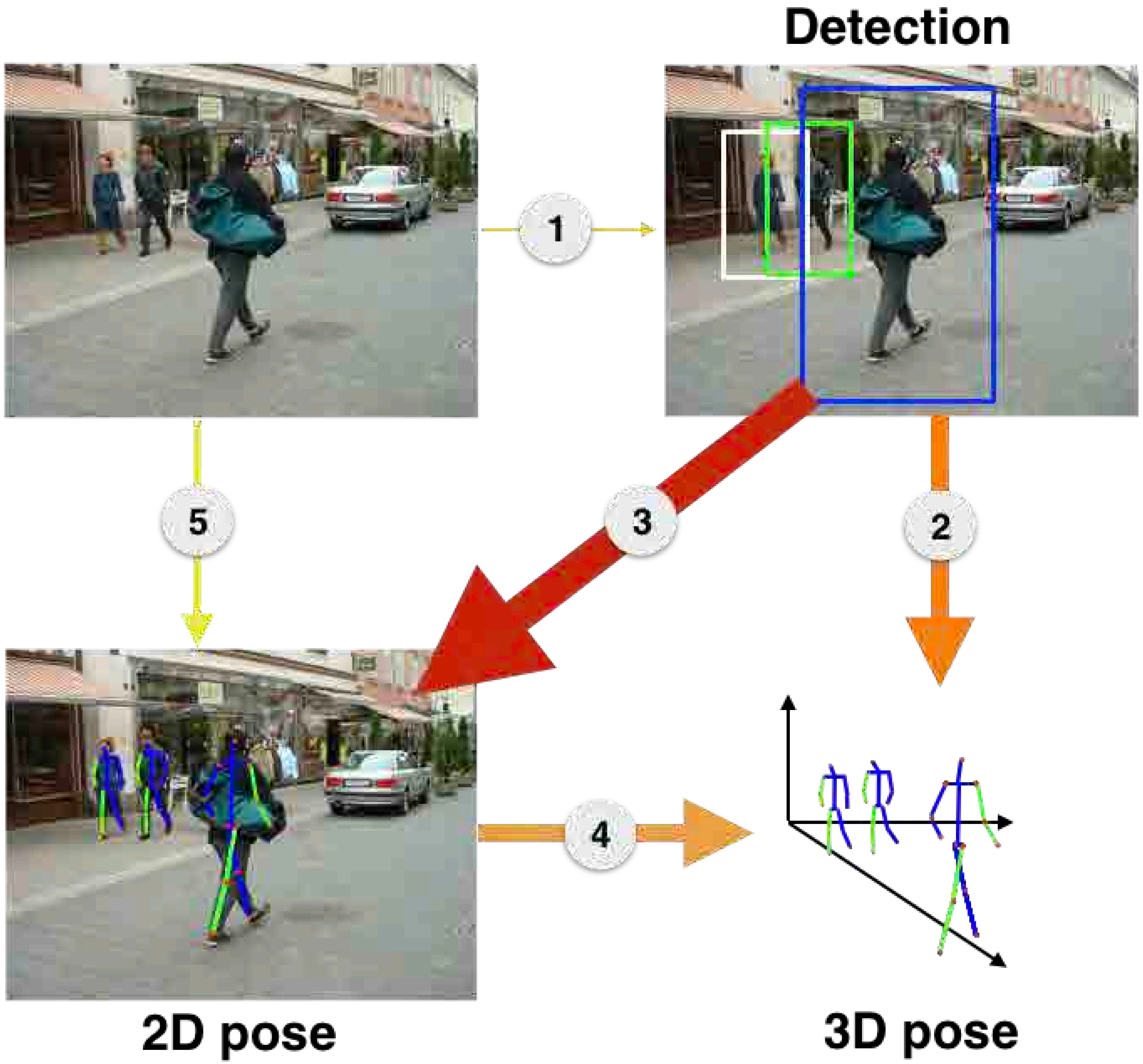
1 [Dalal & Triggs, CVPR'05]

2 [Li et al, ICCV'15, Tekin et al, Zhou et al, CVPR'16]

3 [**ECCV'16:** Newell et al., Insafutdinov et al., Gkioxary et al., Lifshitz et al., Bulat &Tzimiropoulos
CVPR'16: Wei et al, Yang et al, Pishchulin et al, Hu& Ramanan, Carreira et al.,]

4 [Akhter & Black, CVPR'15, Zhou et al., CVPR'15, Bogo et al., ECCV'16, Martinez et al., CVPR'17]

5 [Pishchulin et al, CVPR'16, Cao et al., CVPR'17]



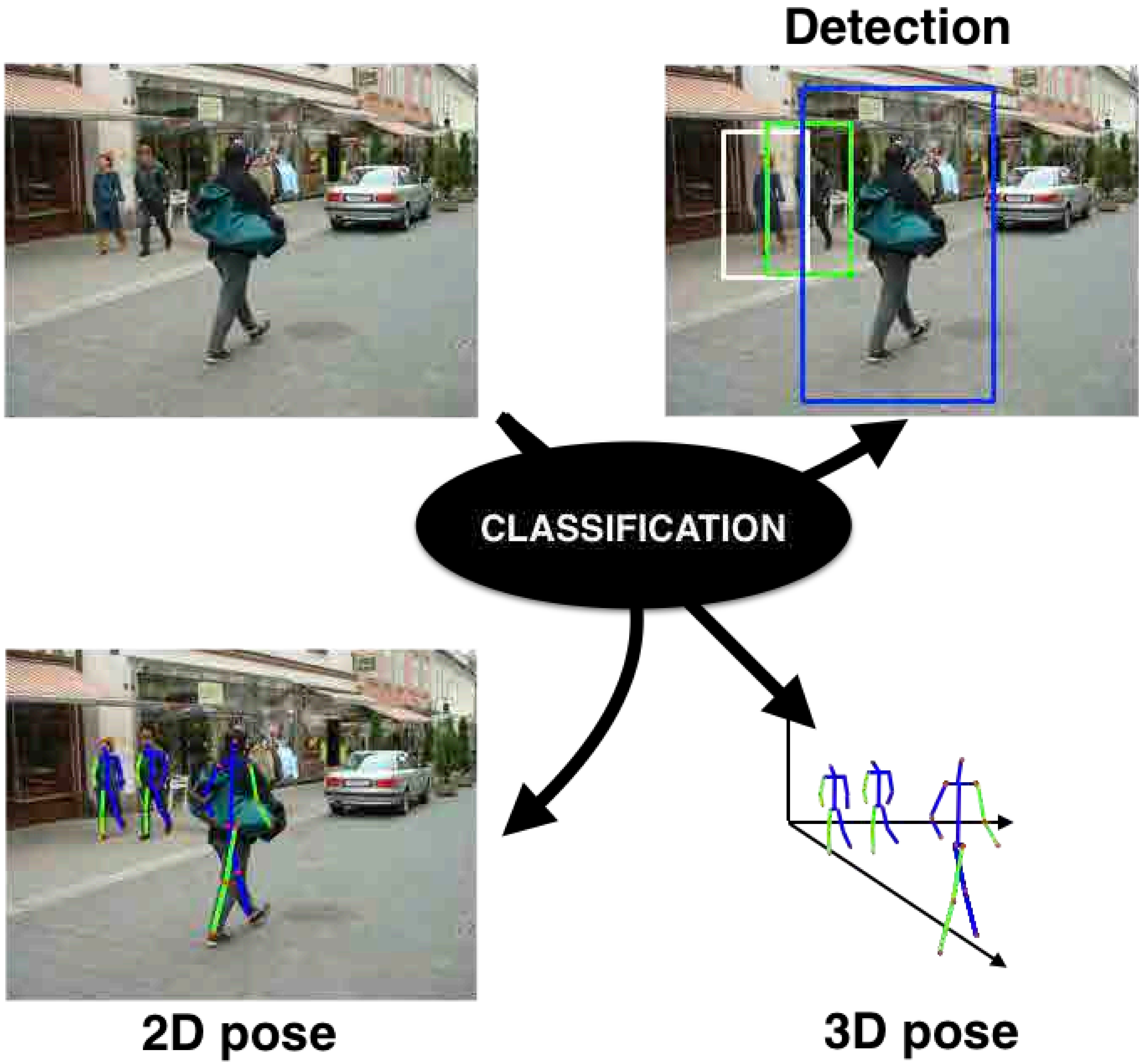
1 [Dalal & Triggs, CVPR'05]

2 [Li et al, ICCV'15, Tekin et al, Zhou et al, CVPR'16]

3 [**ECCV'16:** Newell et al., Insafutdinov et al., Gkioxary et al., Lifshitz et al., Bulat &Tzimiropoulos
CVPR'16: Wei et al, Yang et al, Pishchulin et al, Hu& Ramanan, Carreira et al.,]

4 [Akhter & Black, CVPR'15, Zhou et al., CVPR'15, Bogo et al., ECCV'16, Martinez et al., CVPR'17]

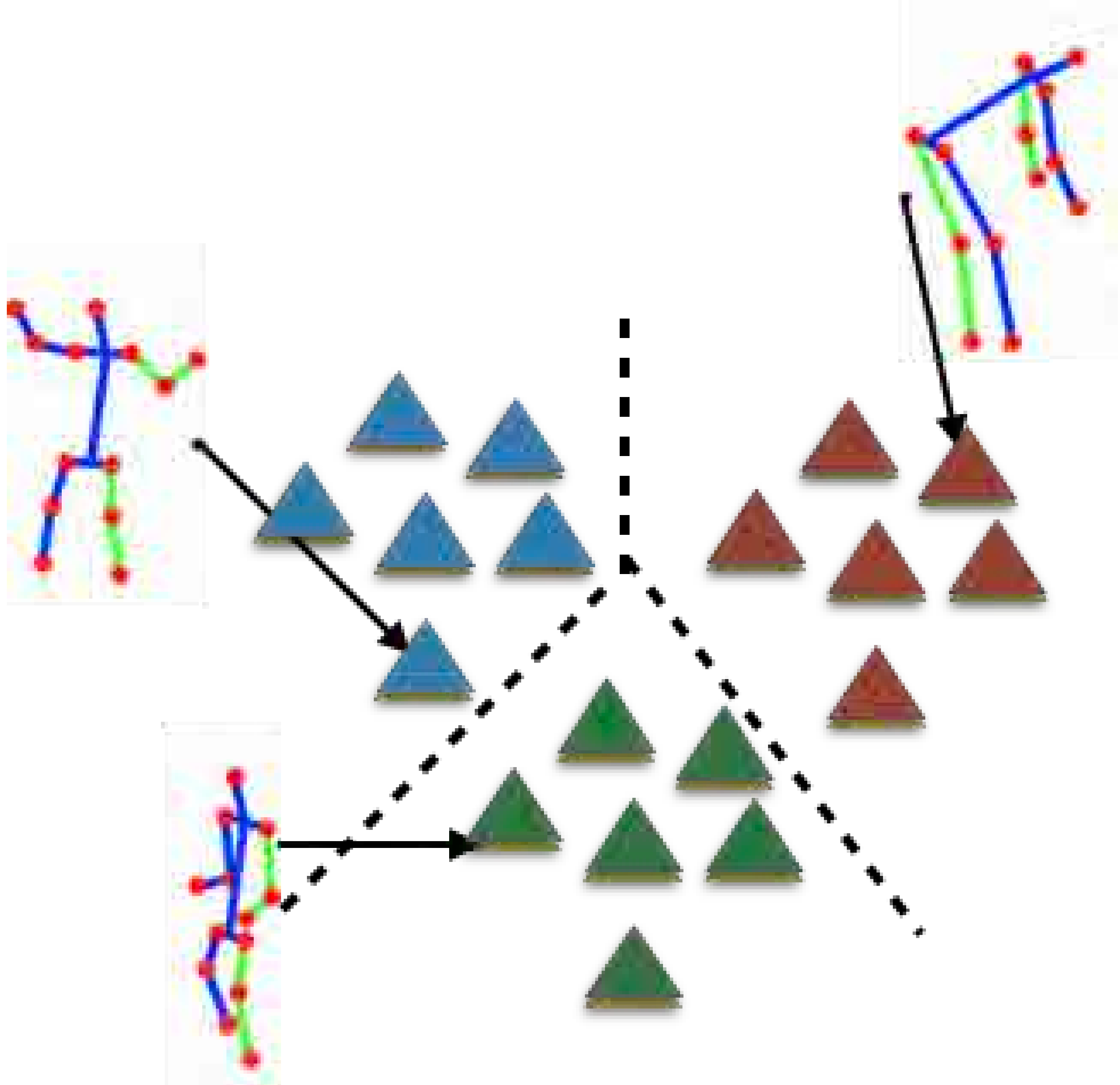
5 [Pishchulin et al, CVPR'16, Cao et al., CVPR'17]

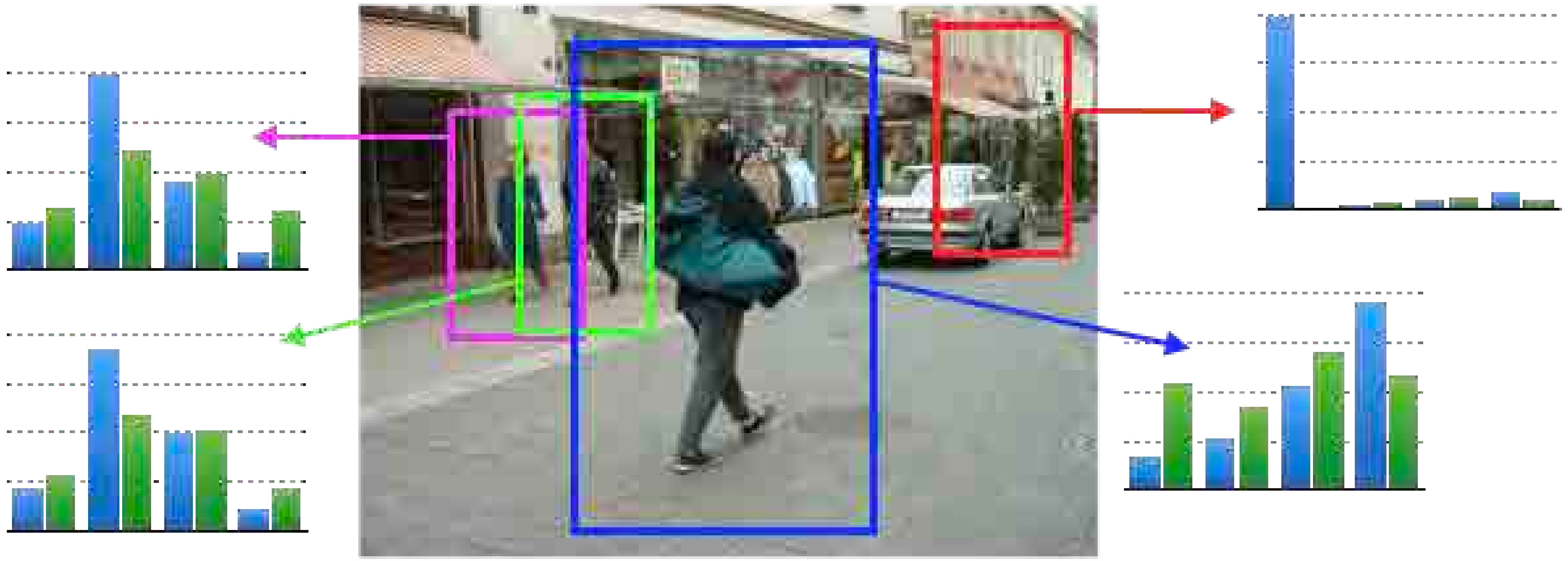

Detection
CLASSIFICATION
2D pose
3D pose

3.Our approaches

Human pose detection

- Partition the space of body poses into K classes
- Train a K-way classifier (K pose classes + bgd)
- Joint localization and pose estimation

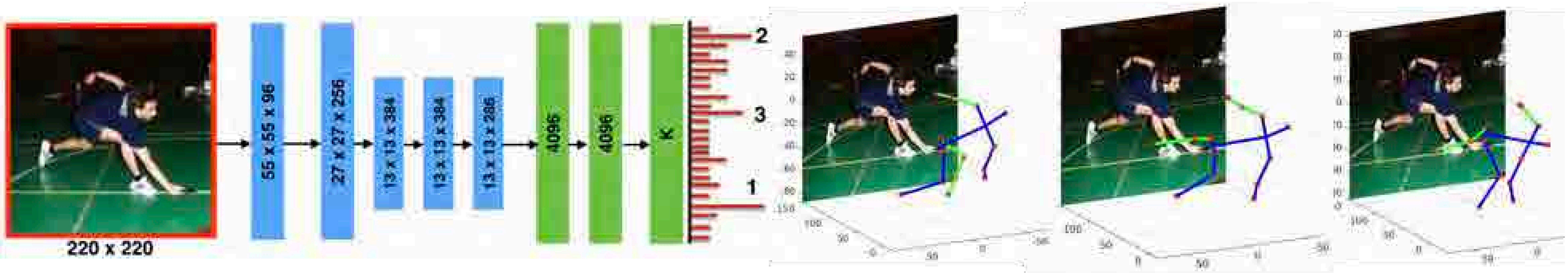




- Return pose (eg, center of top class)

CNN for human pose classication

- 3D pose space partitioned into K clusters (K=5000)
- AlexNet adapted to output a probability distribution over pose classes.



Average 2D/3D poses of top scoring class returned for evaluation.

[G. Rogez and C. Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS 2016]

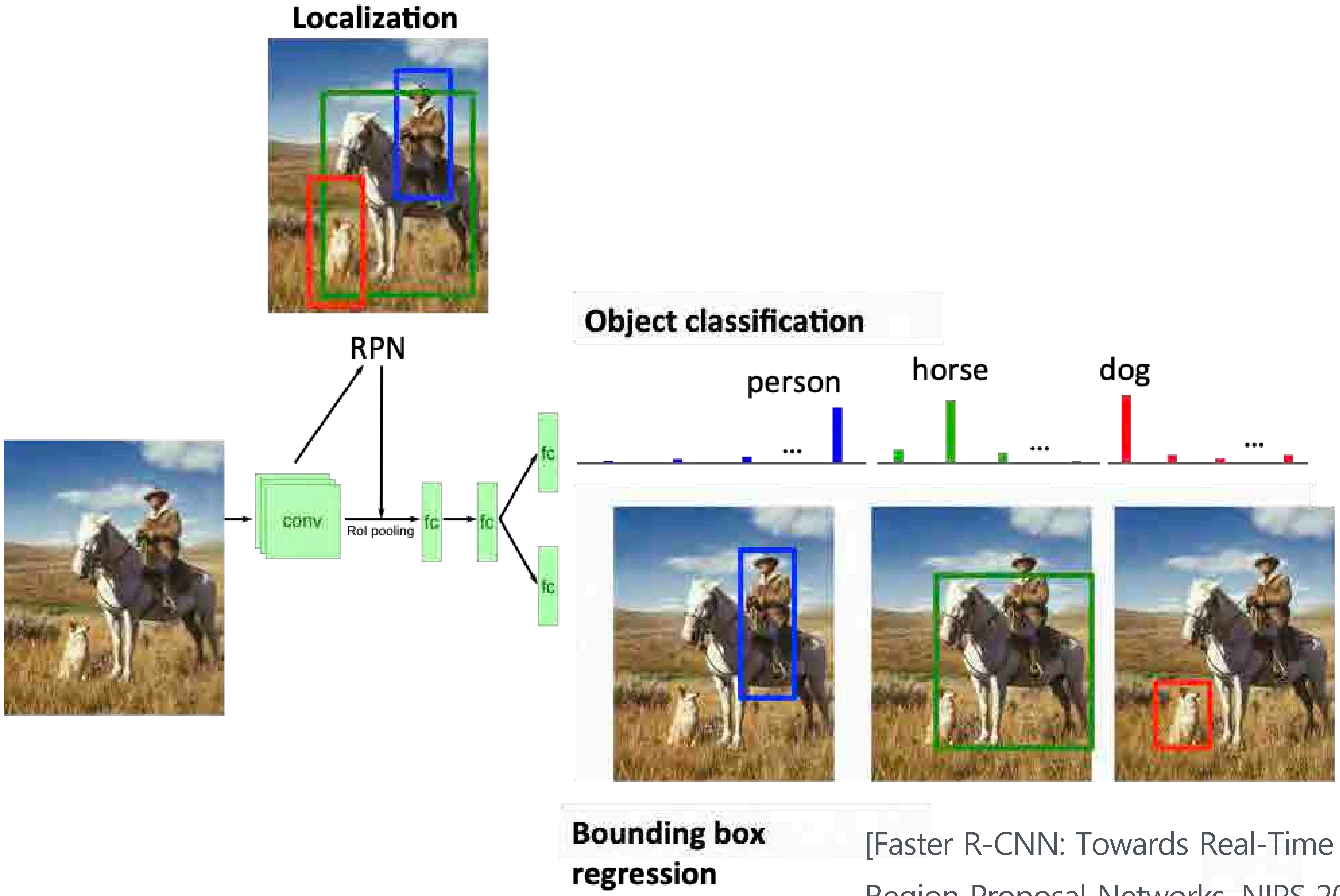
CNN for human pose detection

- Problem: Requires a well-centered bounding box
  A large number of cluster is required (K=5000)

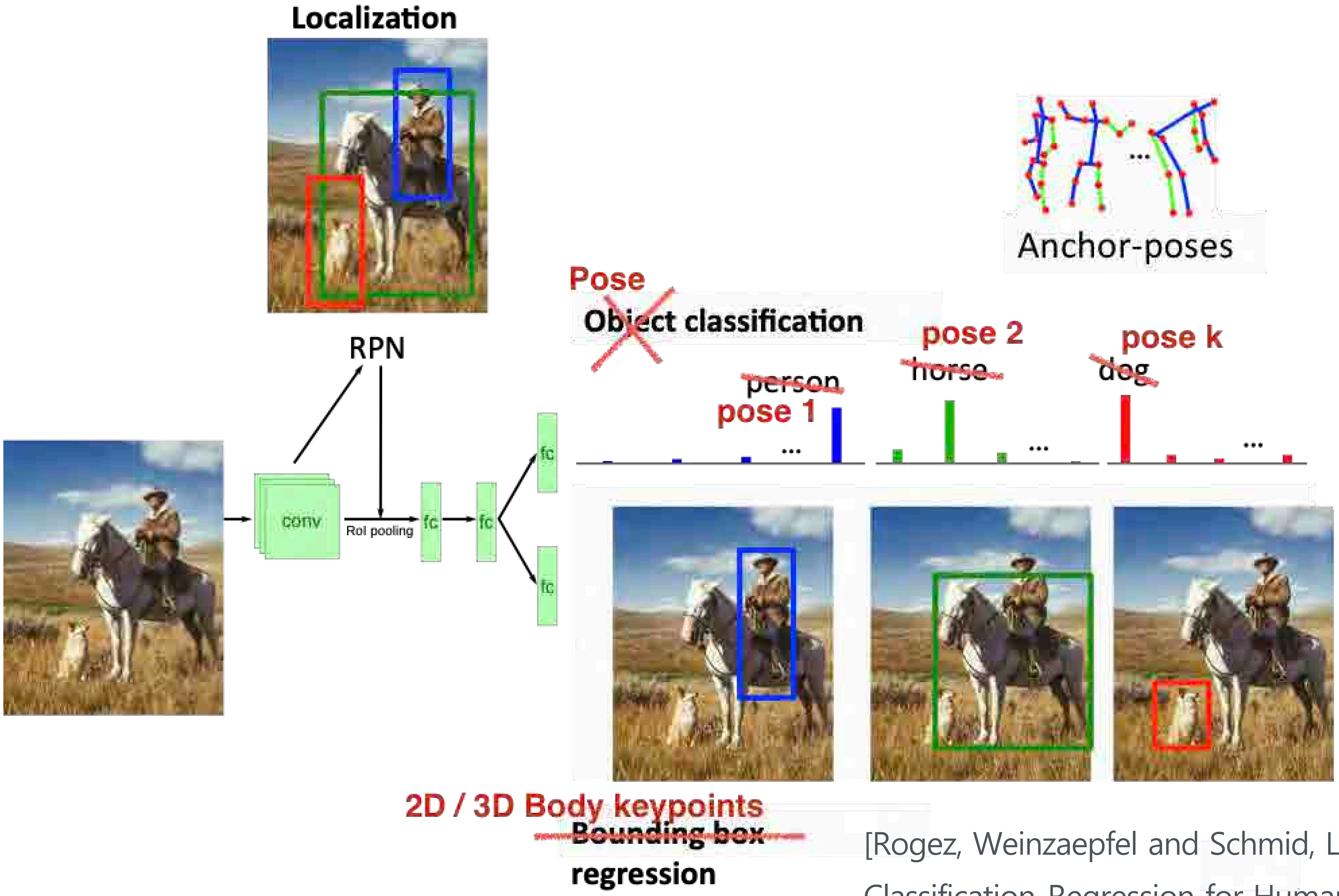

220 x 220

- Solution: Integrate Localisation, Classification and Regression into an end-to-end deep network, **LCR-Net.**

[G. Rogez, P. Weinzaepfel and C. Schmid, LCR-Net: Localization-Classification-Regression for Human Pose. CVPR 2017]



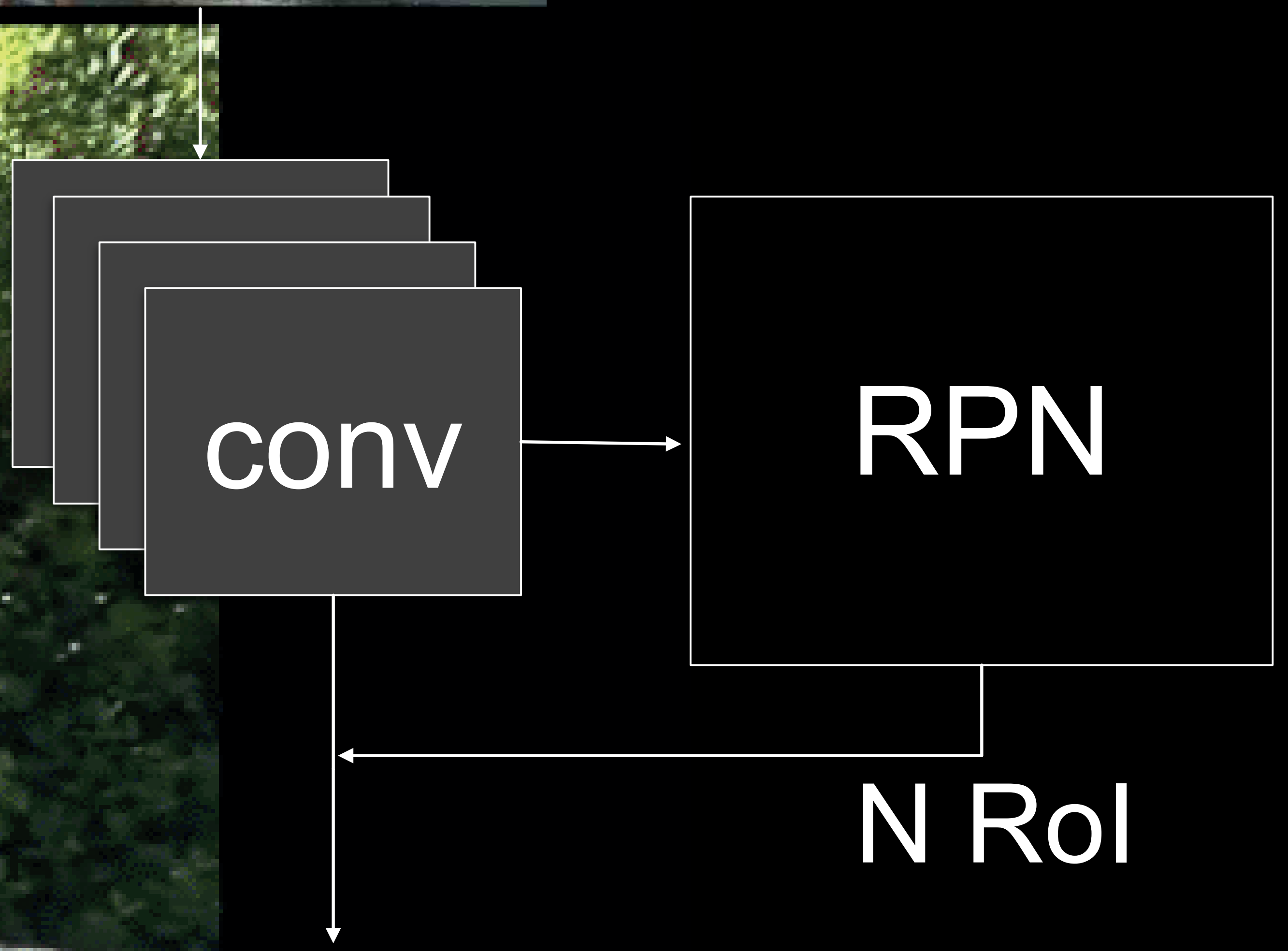
[Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS 2015]



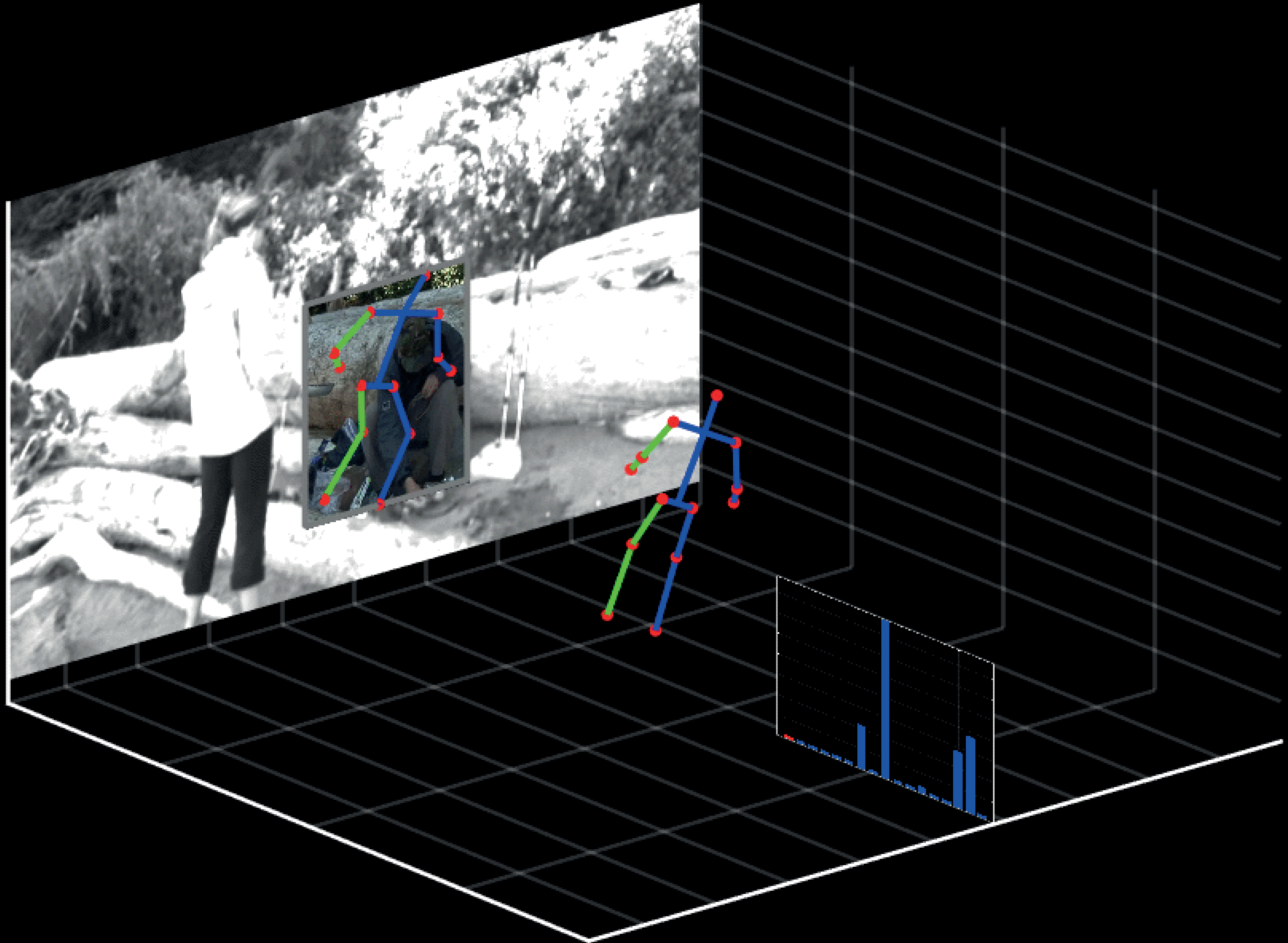
[Rogez, Weinzaepfel and Schmid, LCR-Net: Localization-Classification-Regression for Human Pose. CVPR 2017]

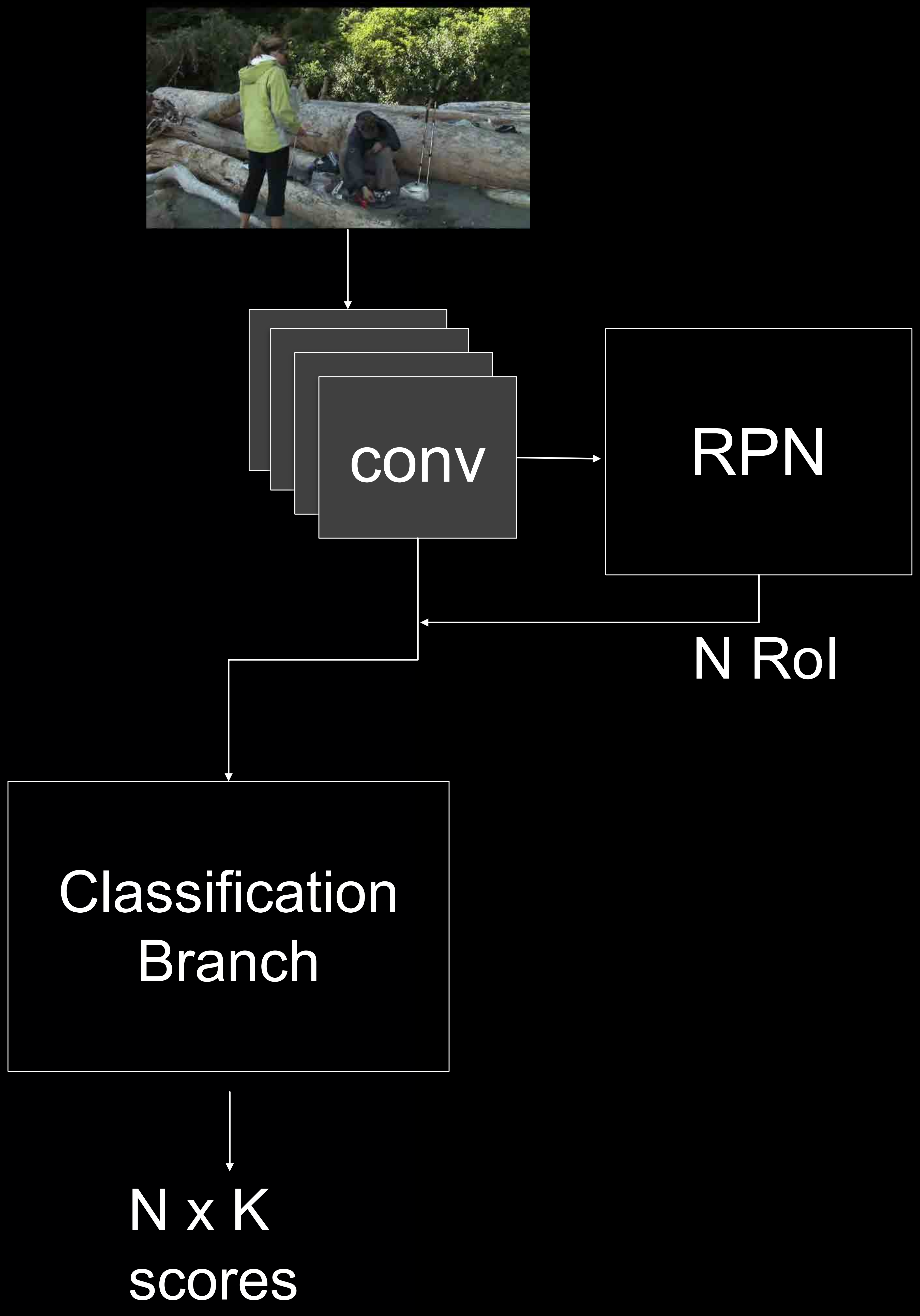
LCR-Net: End-to-End Architecture

LCR-Net: Localization

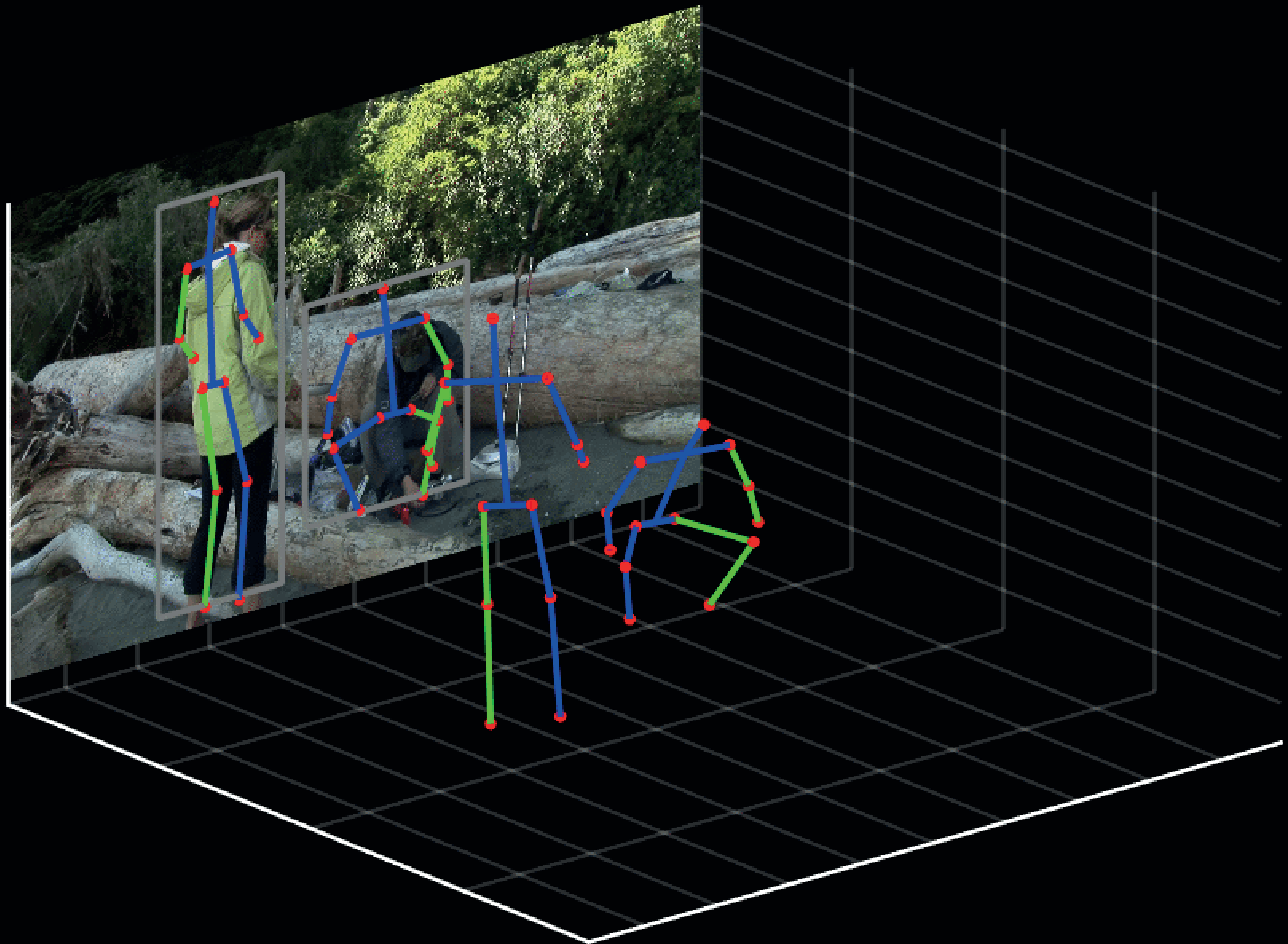


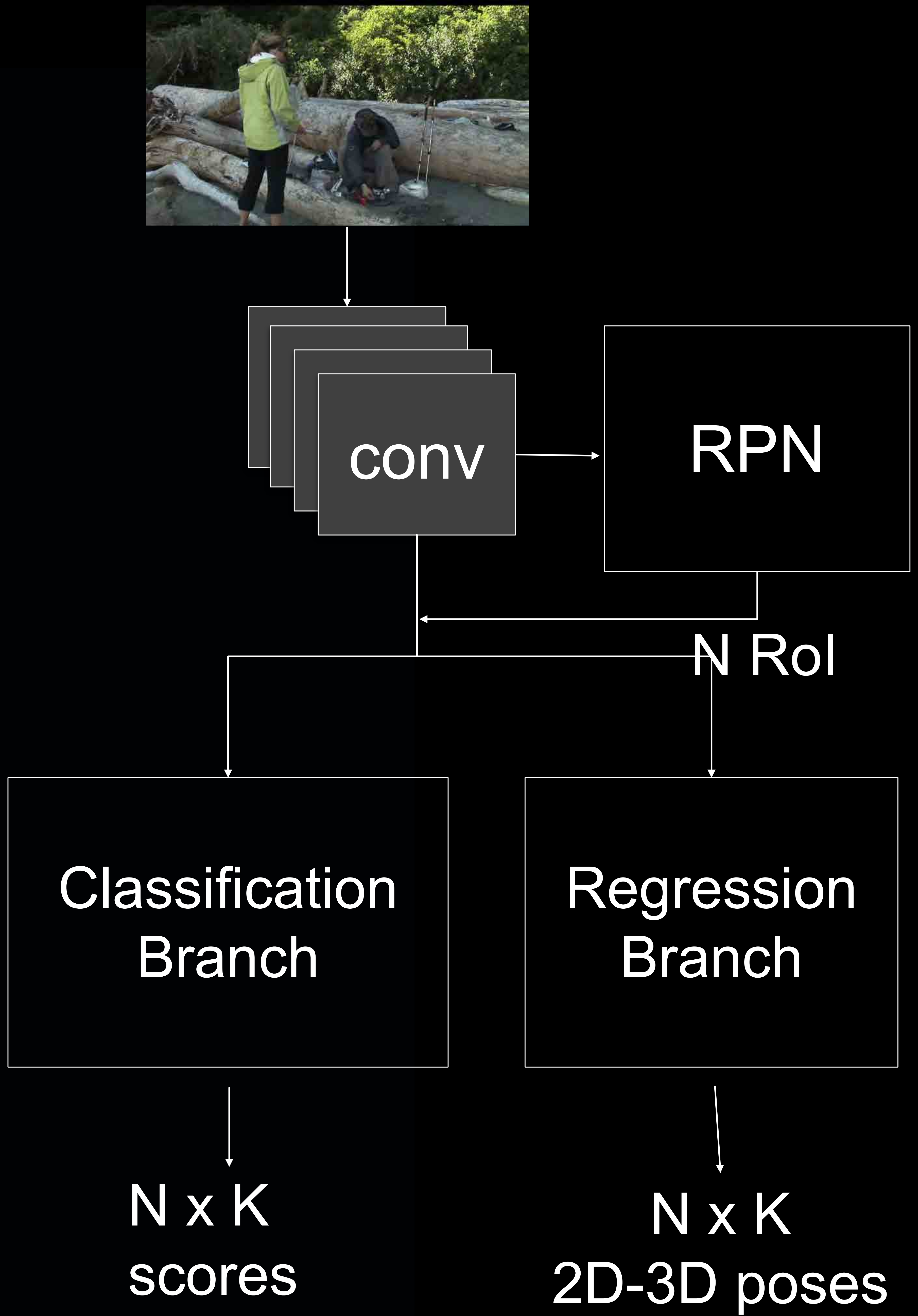
LCR-Net: Classification





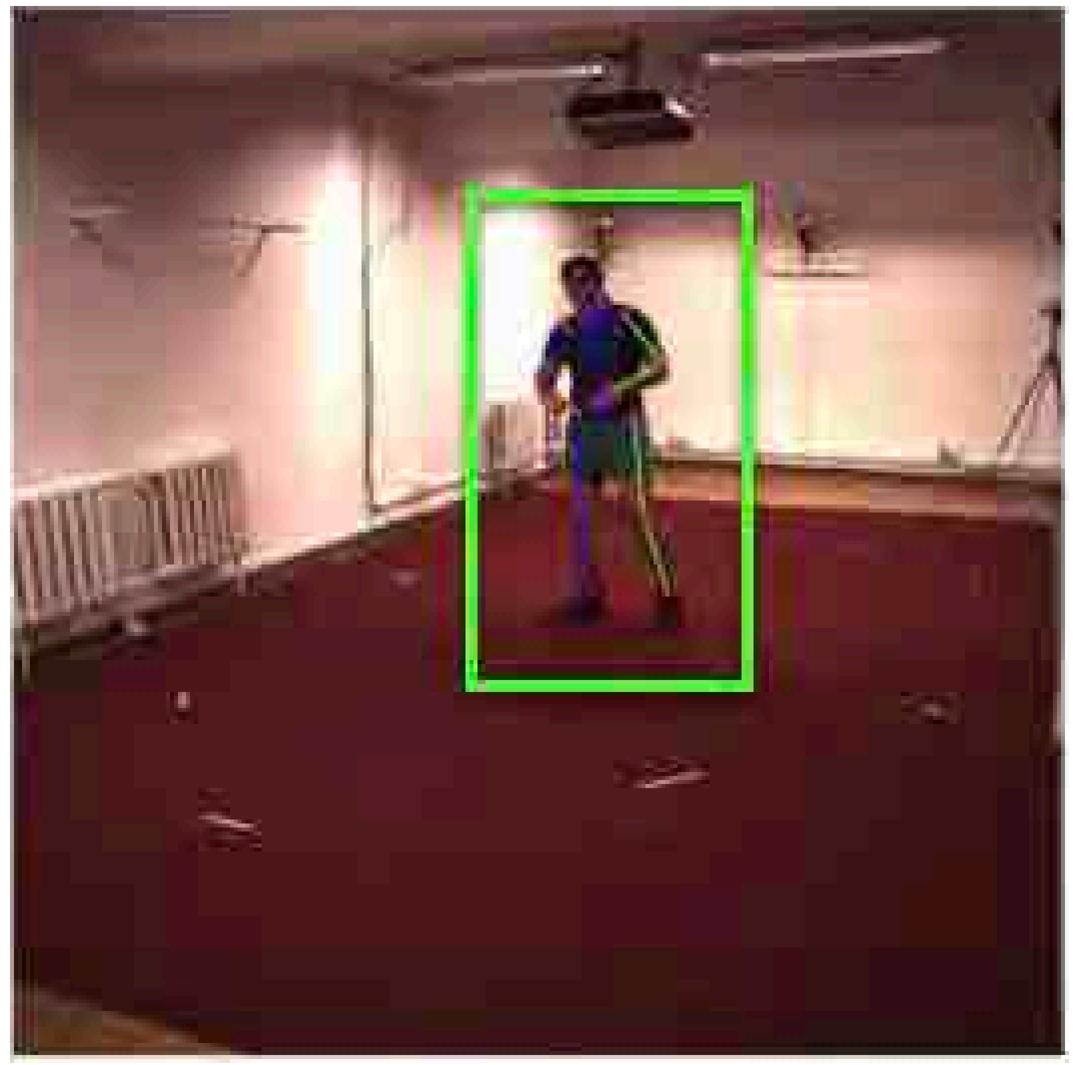
LCR-Net: Regression





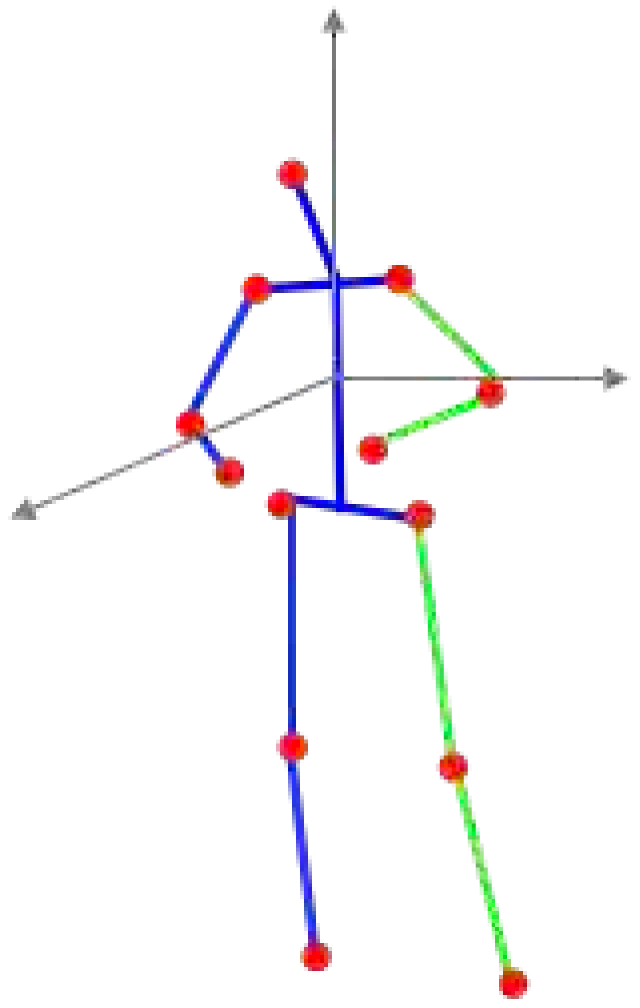
LCR-Net training

DEVIEW 2019



Bounding box + class label

normalized 2D poses
(w.r.t bounding box)



normalized 3D poses
(aligned+orientated)

LCR-Net Loss: $\mathcal{L}$

RPN loss
(cf FasterRCNN)

log loss of
the true class

L1-smooth loss

Evaluation on Human 3.6M

DEVIEW 2019

300k training images with 2D and 3D poses

5 subjects for training

2 subjects for test

Method	Error (mm)
AlexNet (K=5000)	87.3
LCR-Net with VGG16 backbone (K=100)	71.6
+ Synth training data	**59.3**
+ ResNet50 backbone (LCR-Net++)	**54.3**

Best performance achieved for **K=100** anchor poses.

Boost with Synth. data

Small improvement with ResNet50

[Rogez, Weinzaepfel & Schmid, LCR-Net++, IEEE T. PAMI 2019]

Qualitative results on Human 3.6M

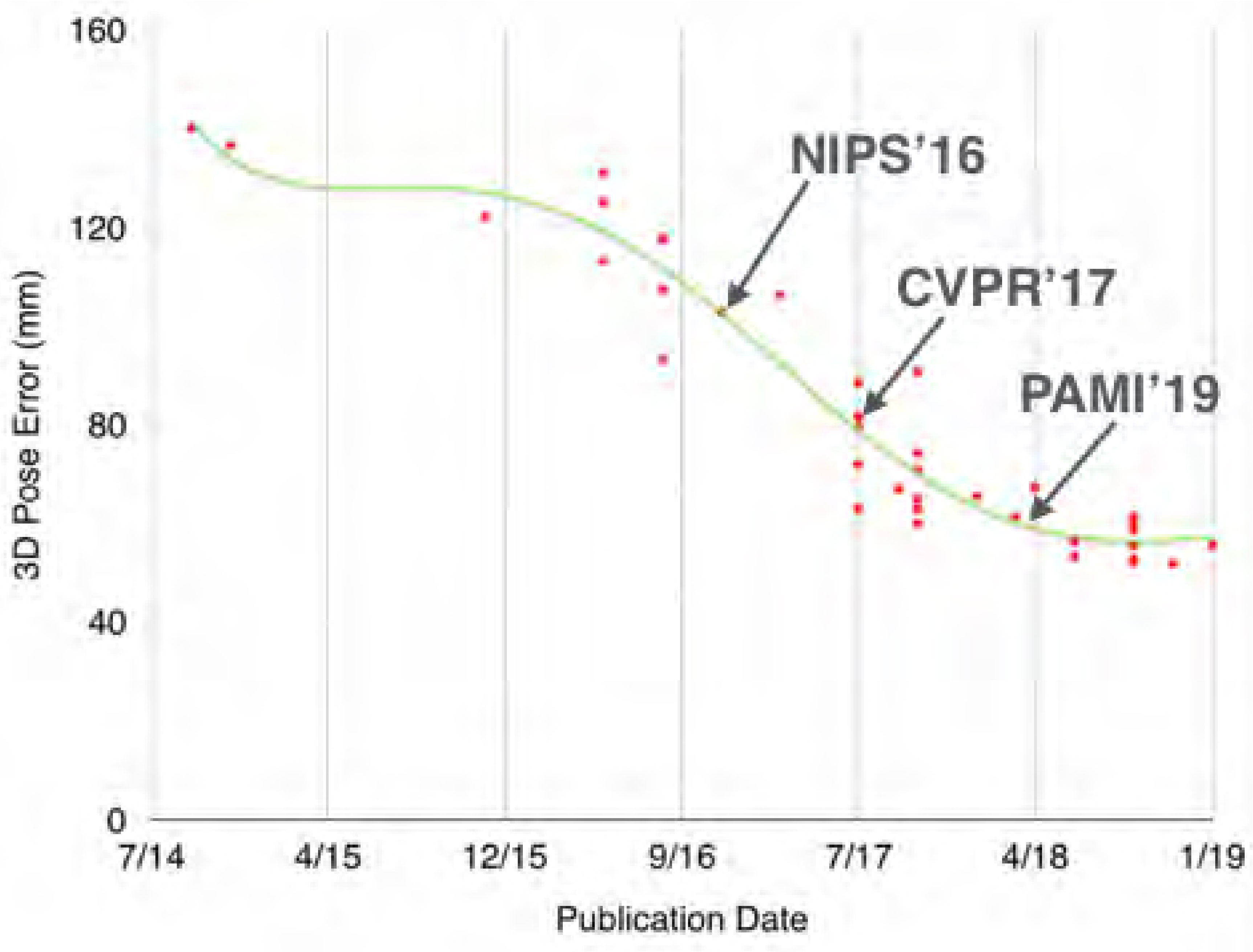
Comparaison with state of the art

DEVIEW 2019

300k training images with 2D and 3D poses

5 subjects for training

2 subjects for test

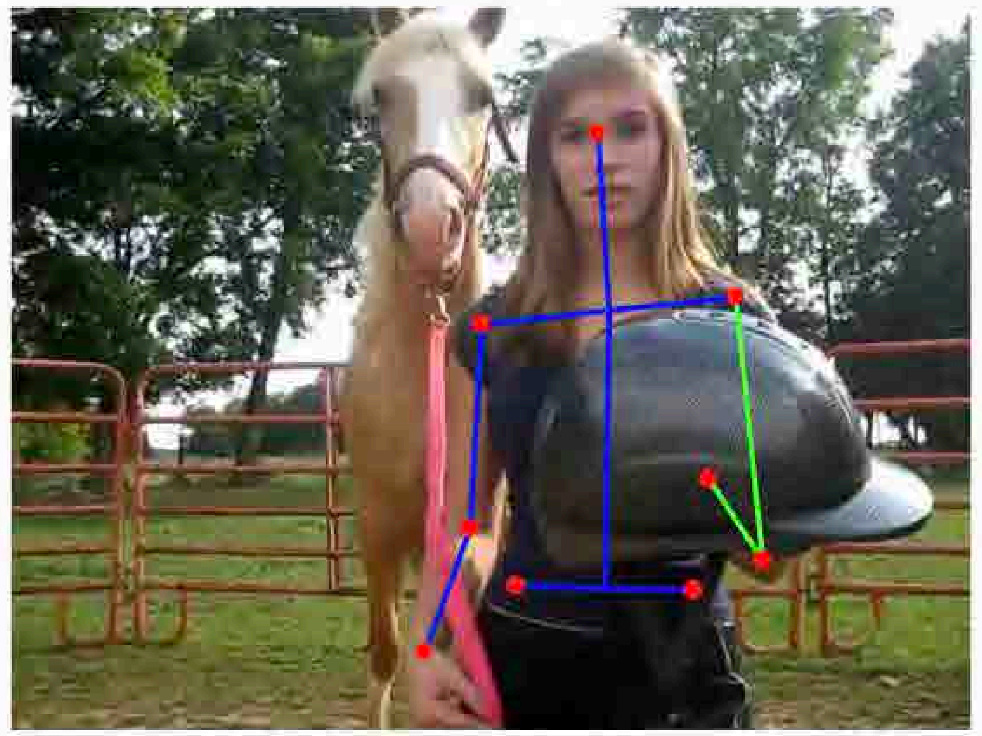


- The 3 presented papers **on par with state of the art**
- Performance on H3.6M is saturating
- **Error ~precision** of sensor used to capture groundtruth

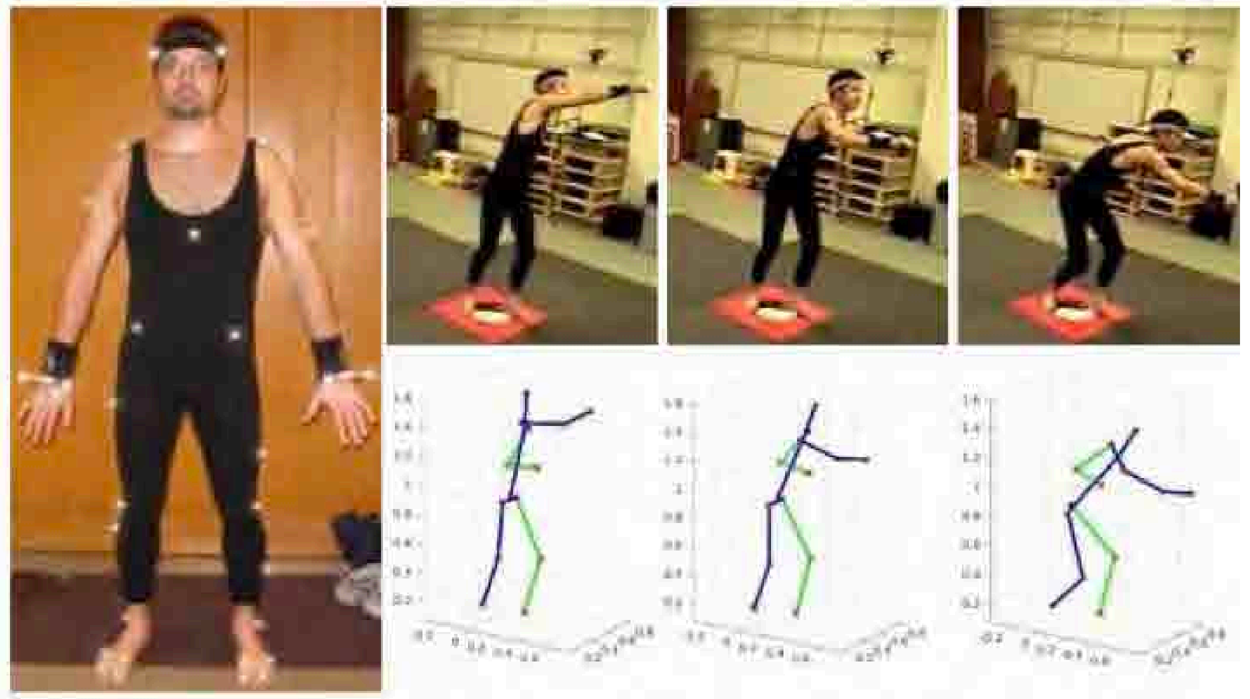
  ->Problem solved or data not hard enough?

  Training and testset distribution overlap

In the wild training data



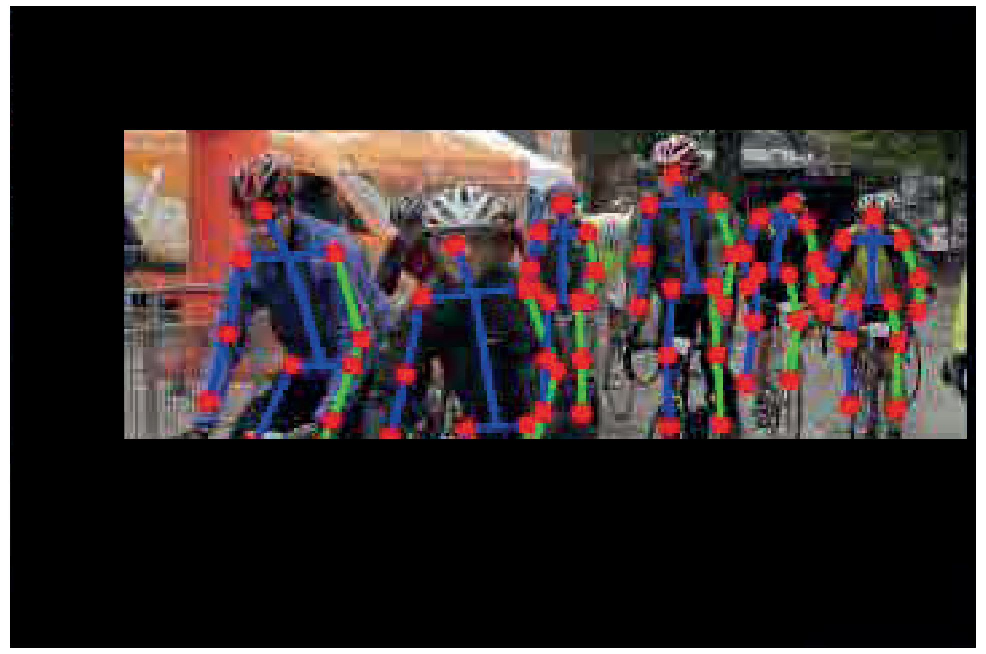
2D pose annotations



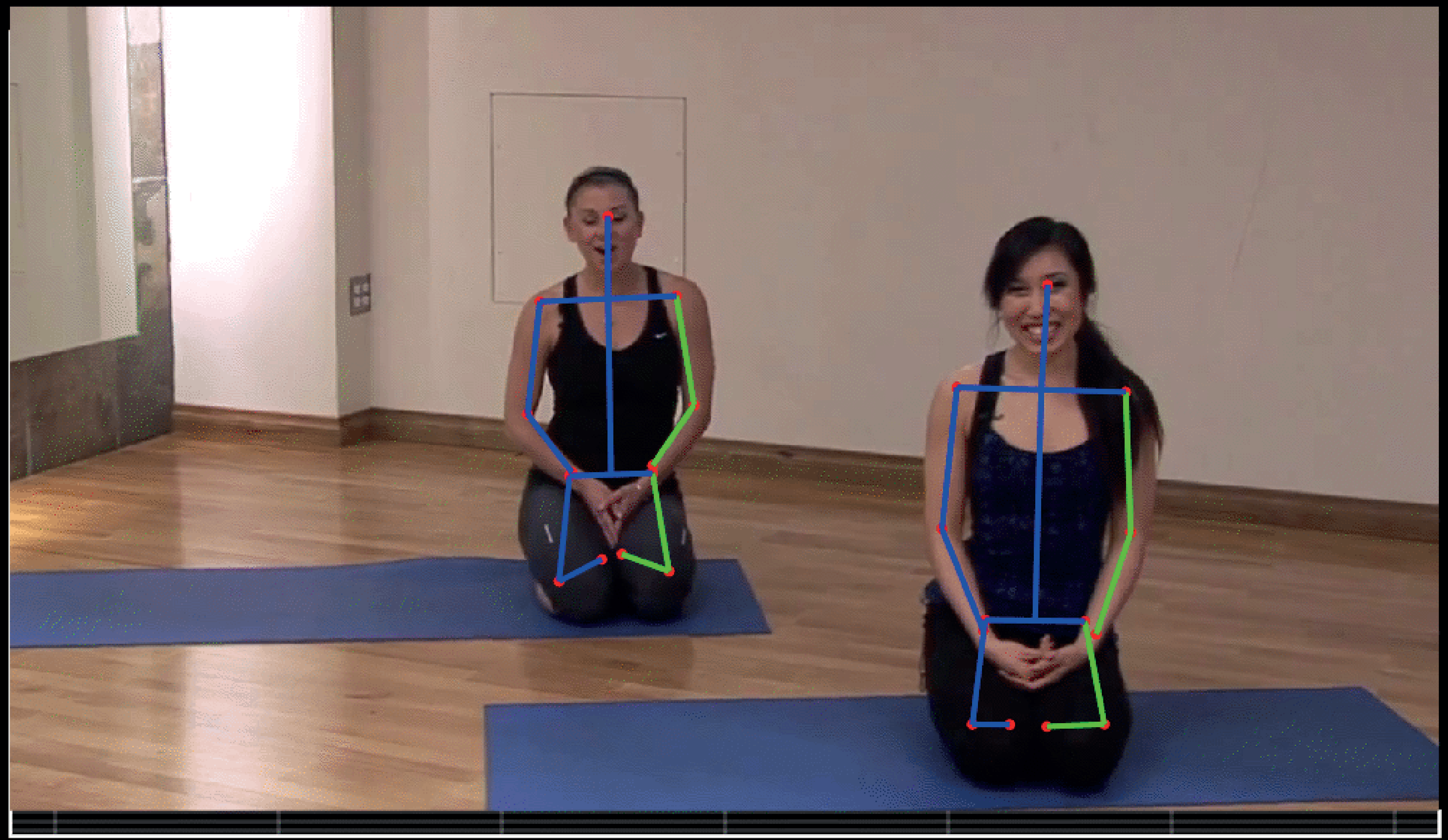
MoCap 3D data

Results in the wild

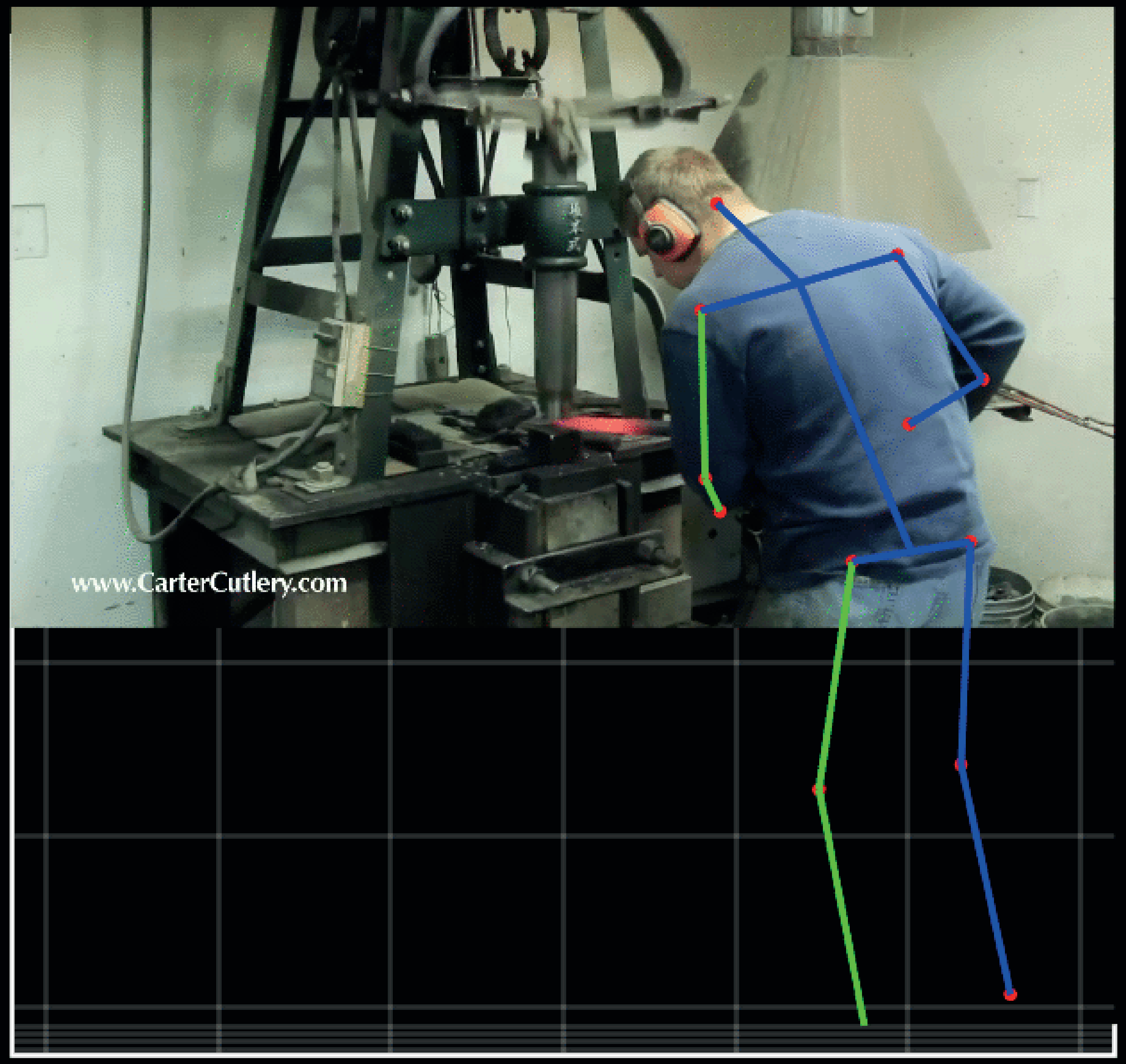
We are the first to evaluate in 4 regimes with competitive results in all 4:

	Single person	Multi-person
2D pose estimation	 MPII	 MPII
3D pose estimation	 H3.6M	 MuPoTS

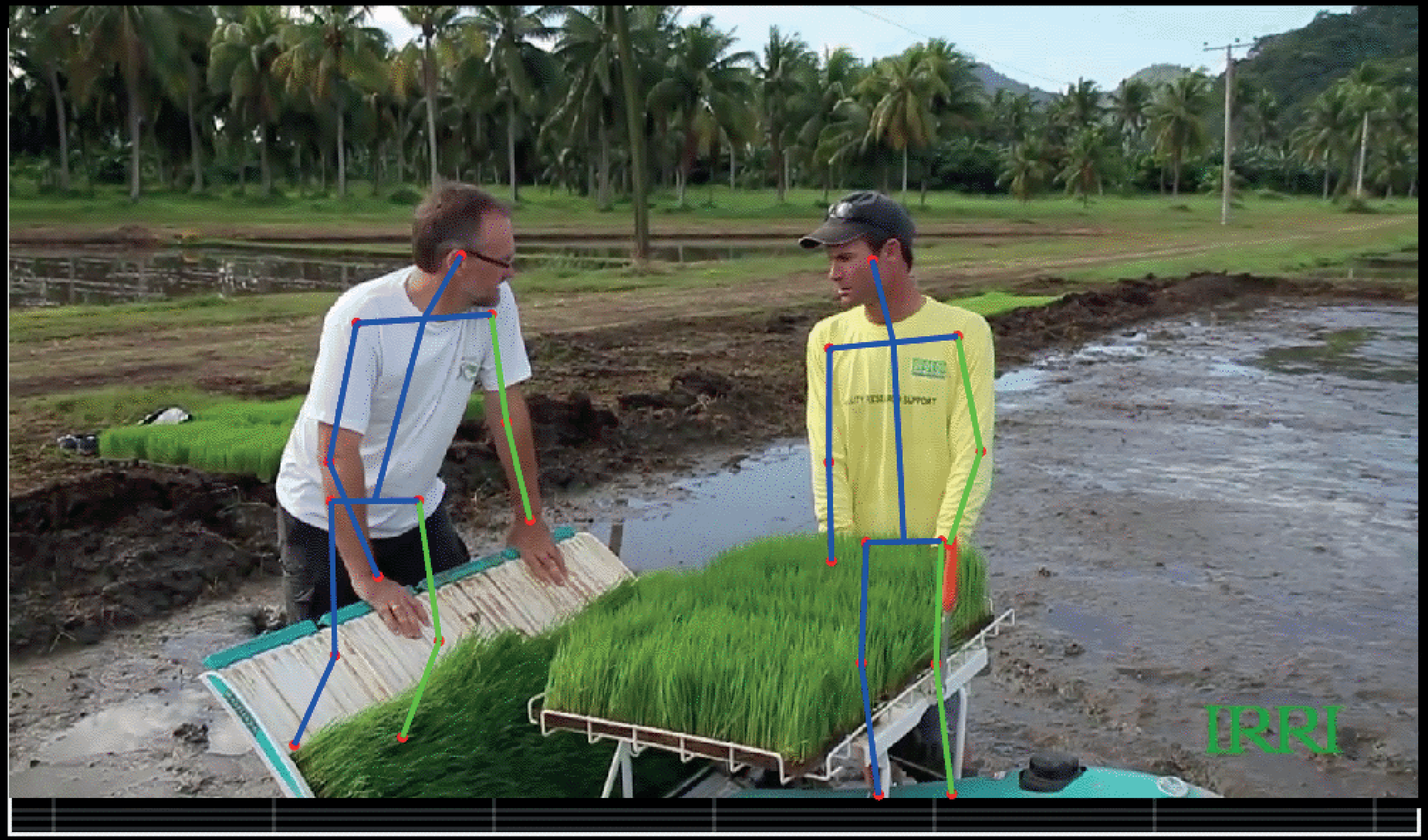

LCR-Net can handle varied poses …

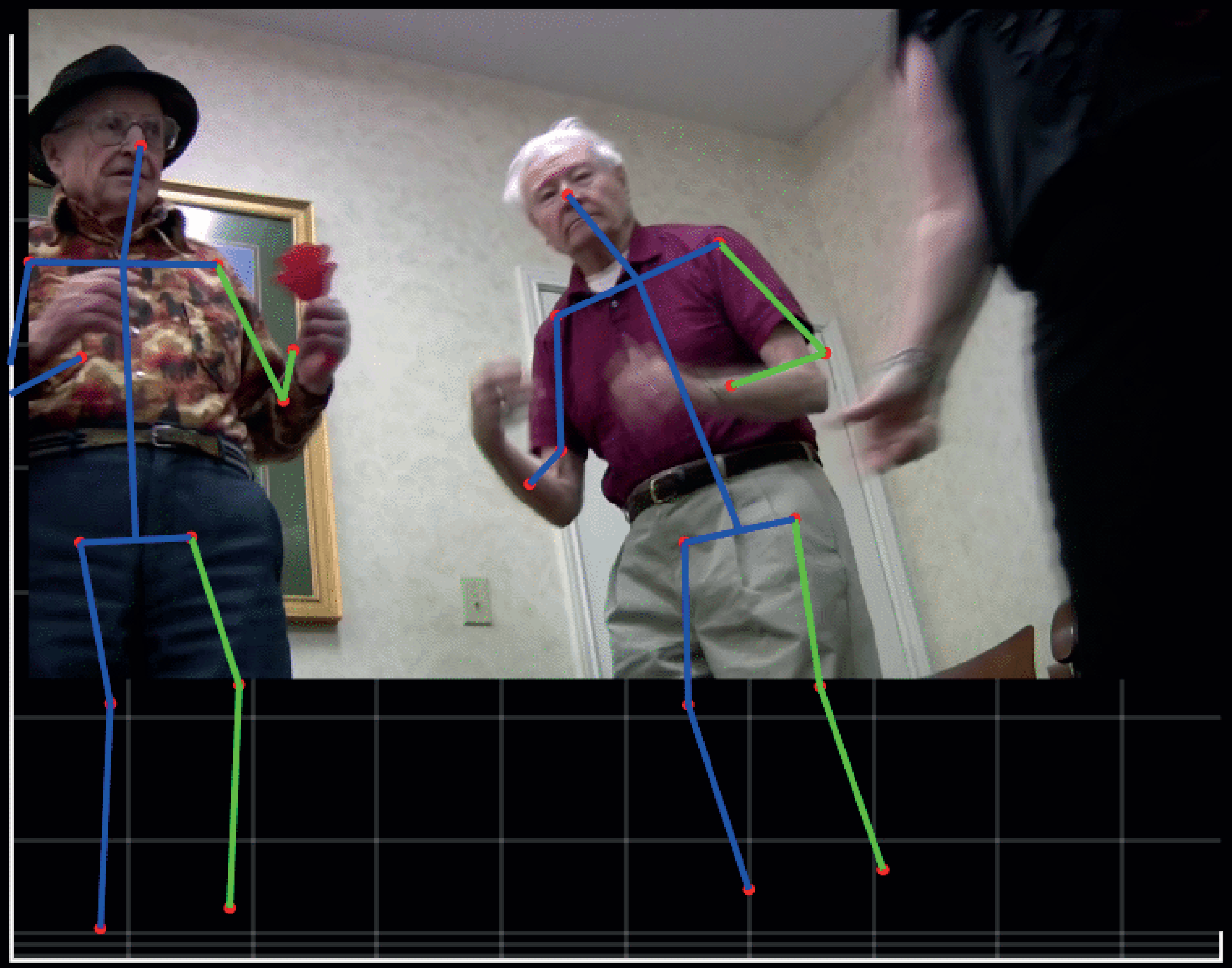



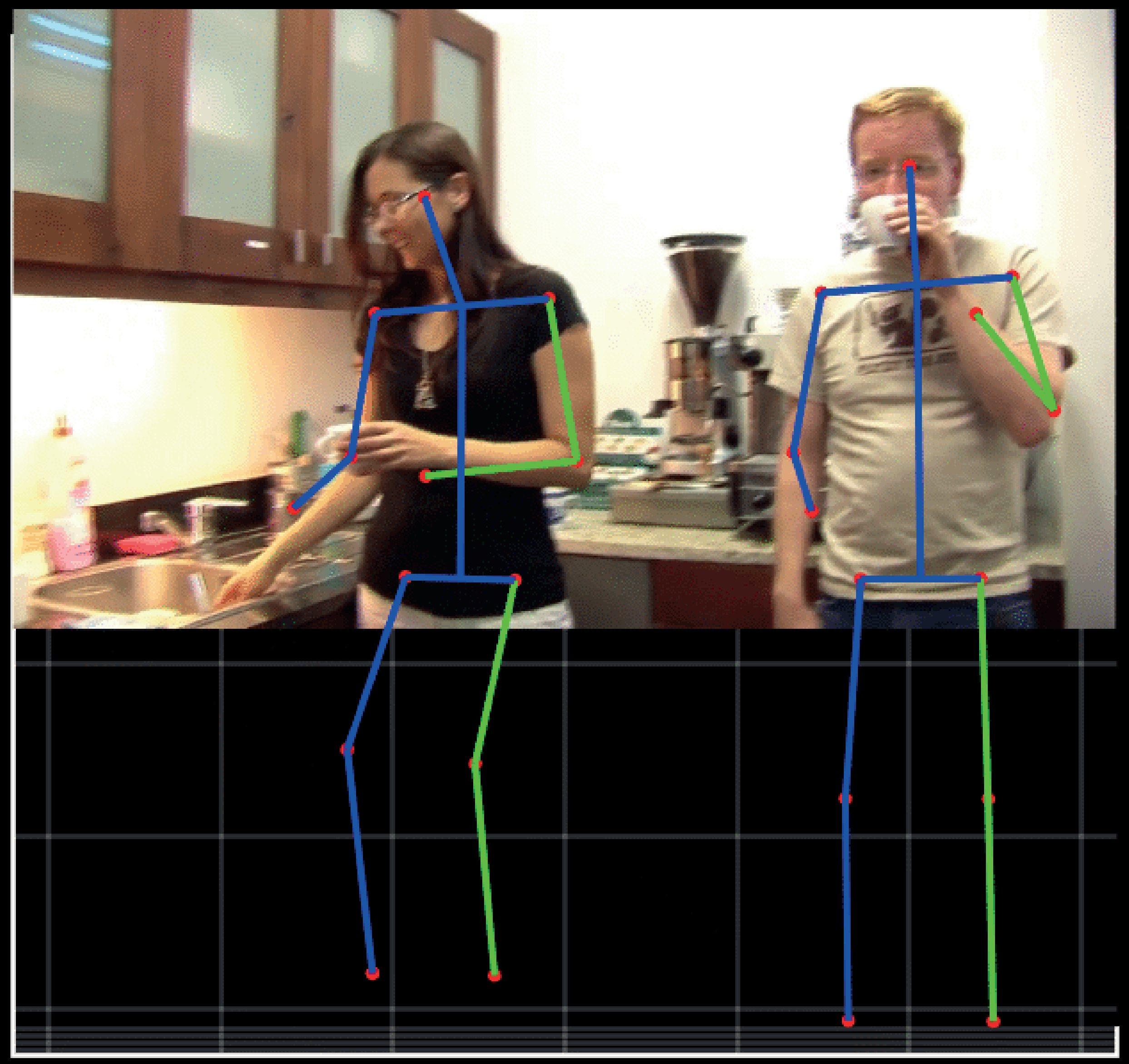
Self-occlusions



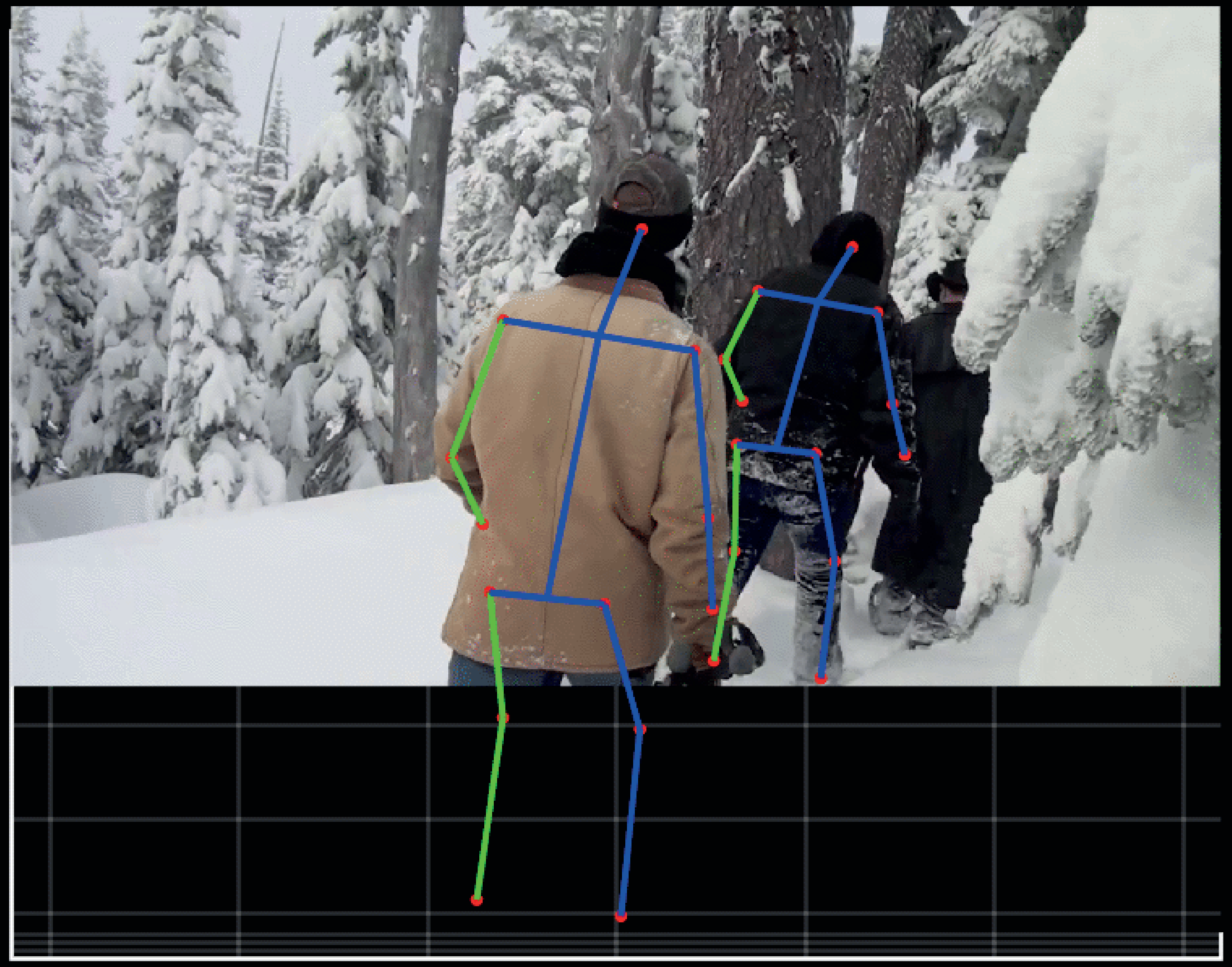
Occlusions

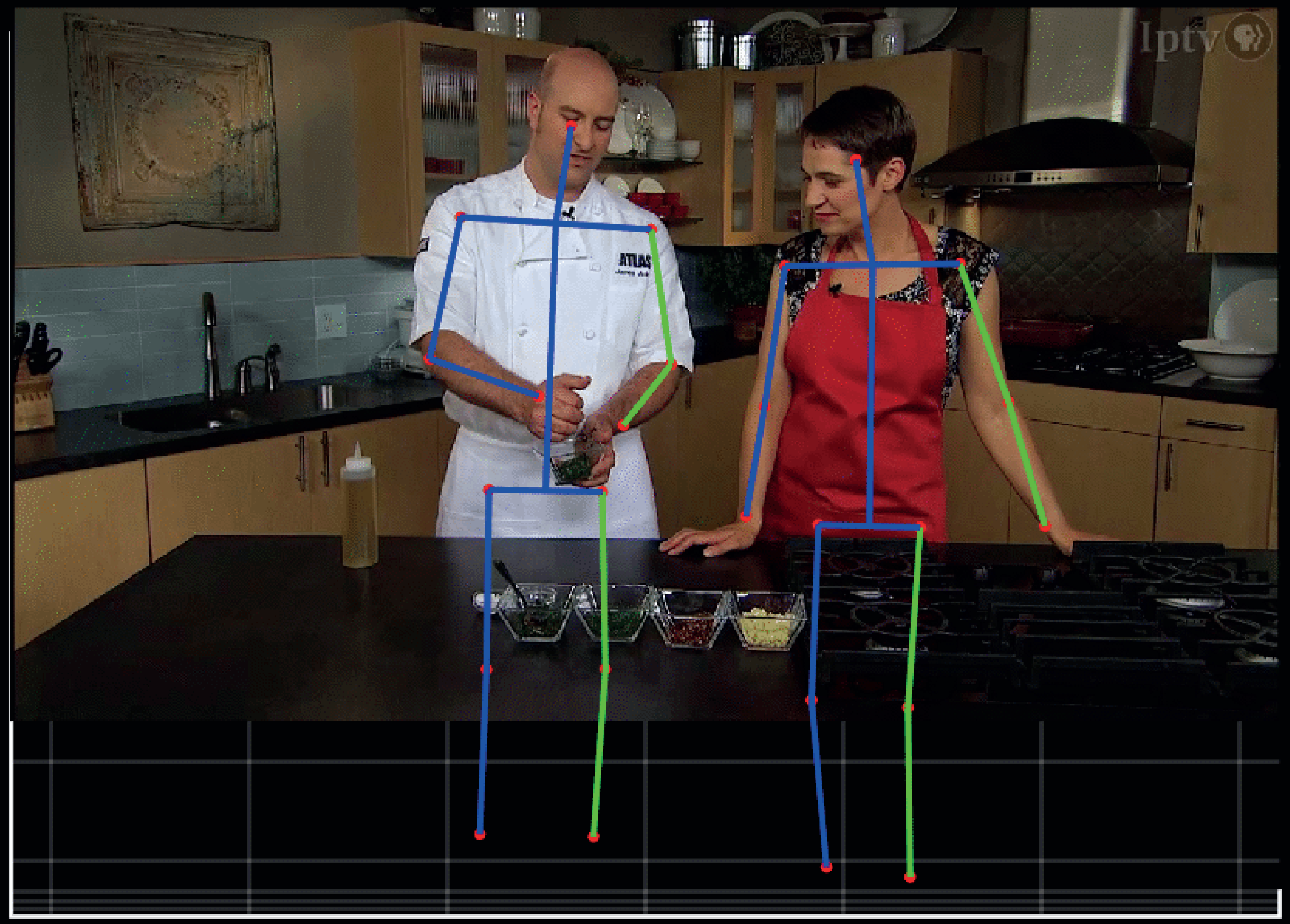






Truncations

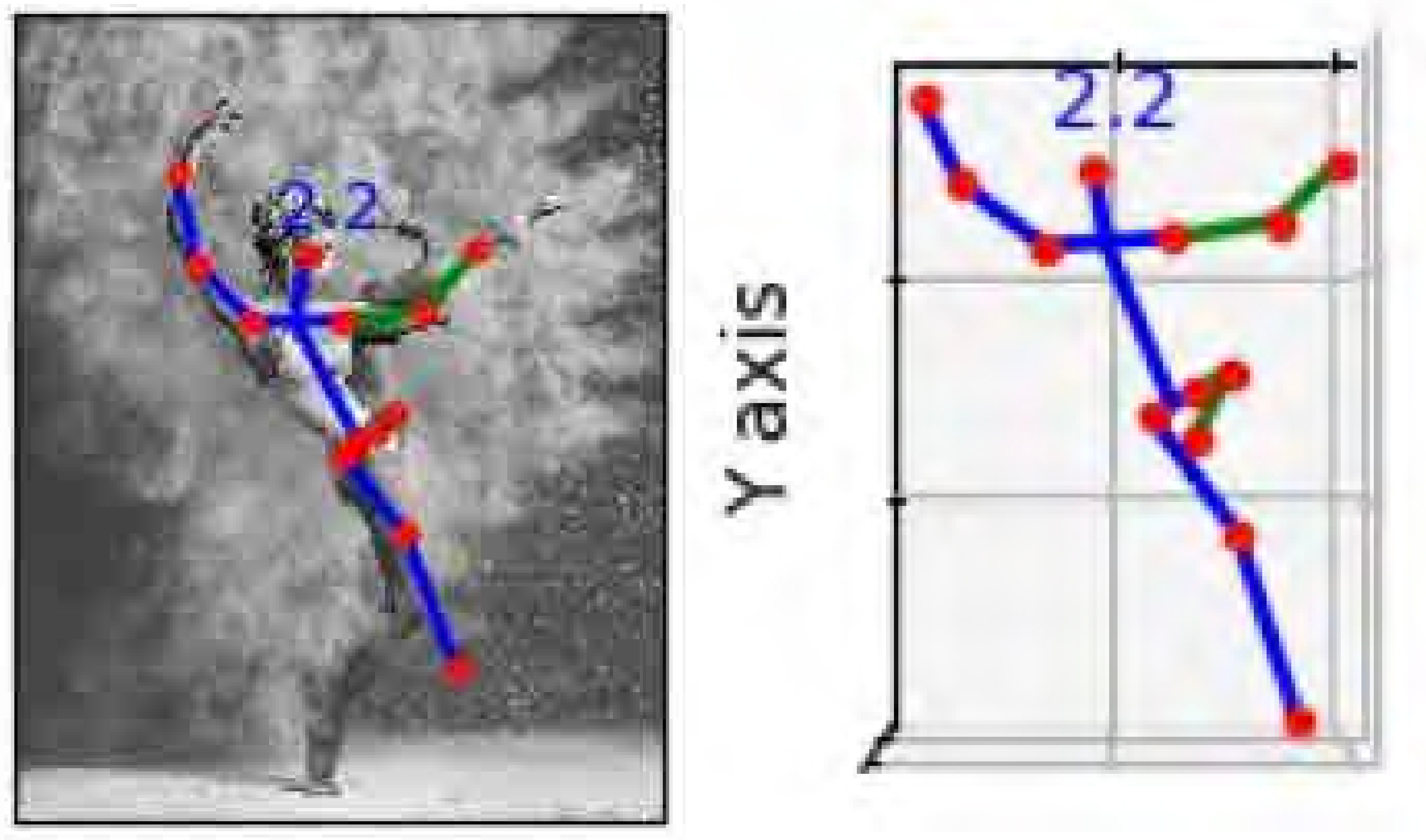


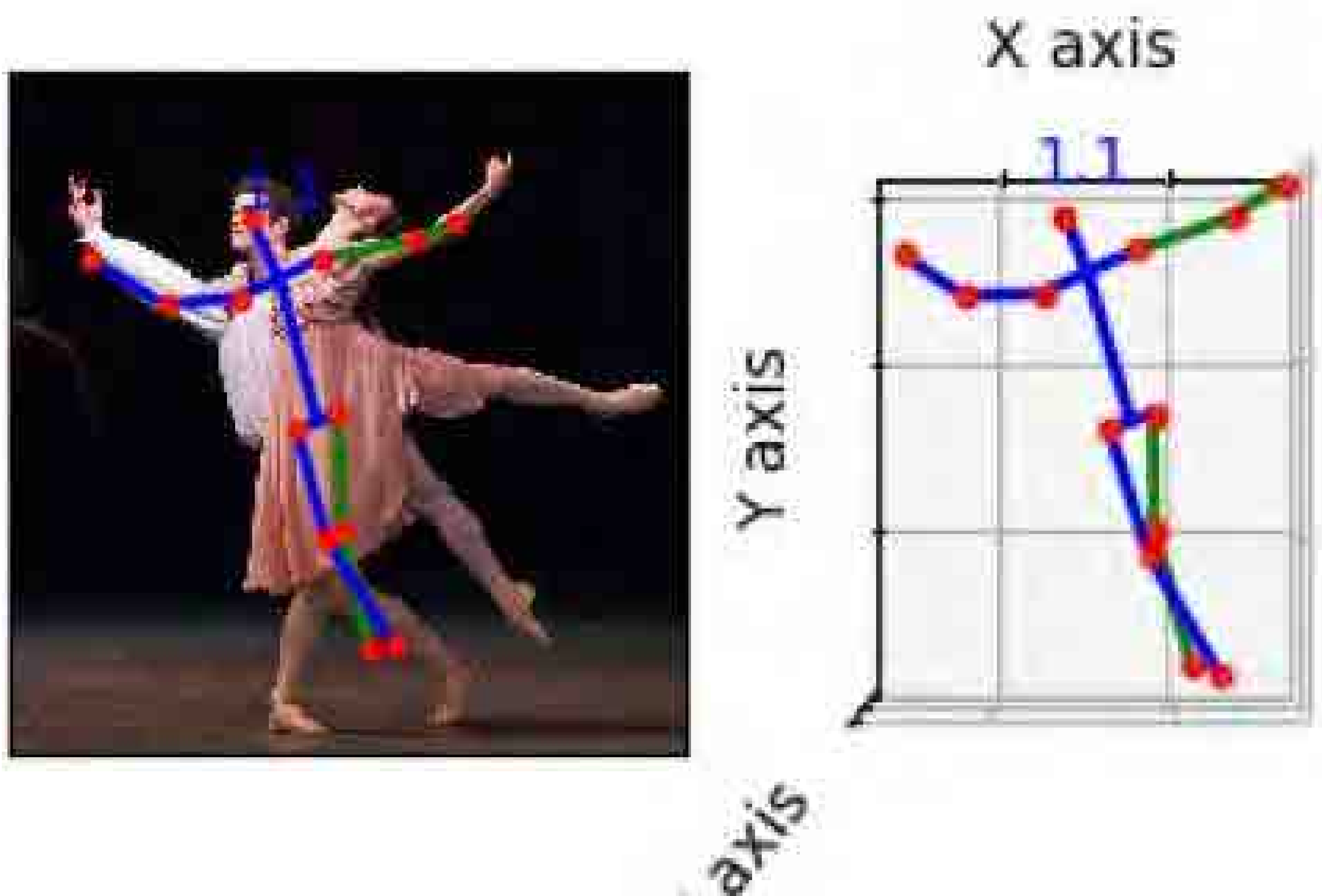


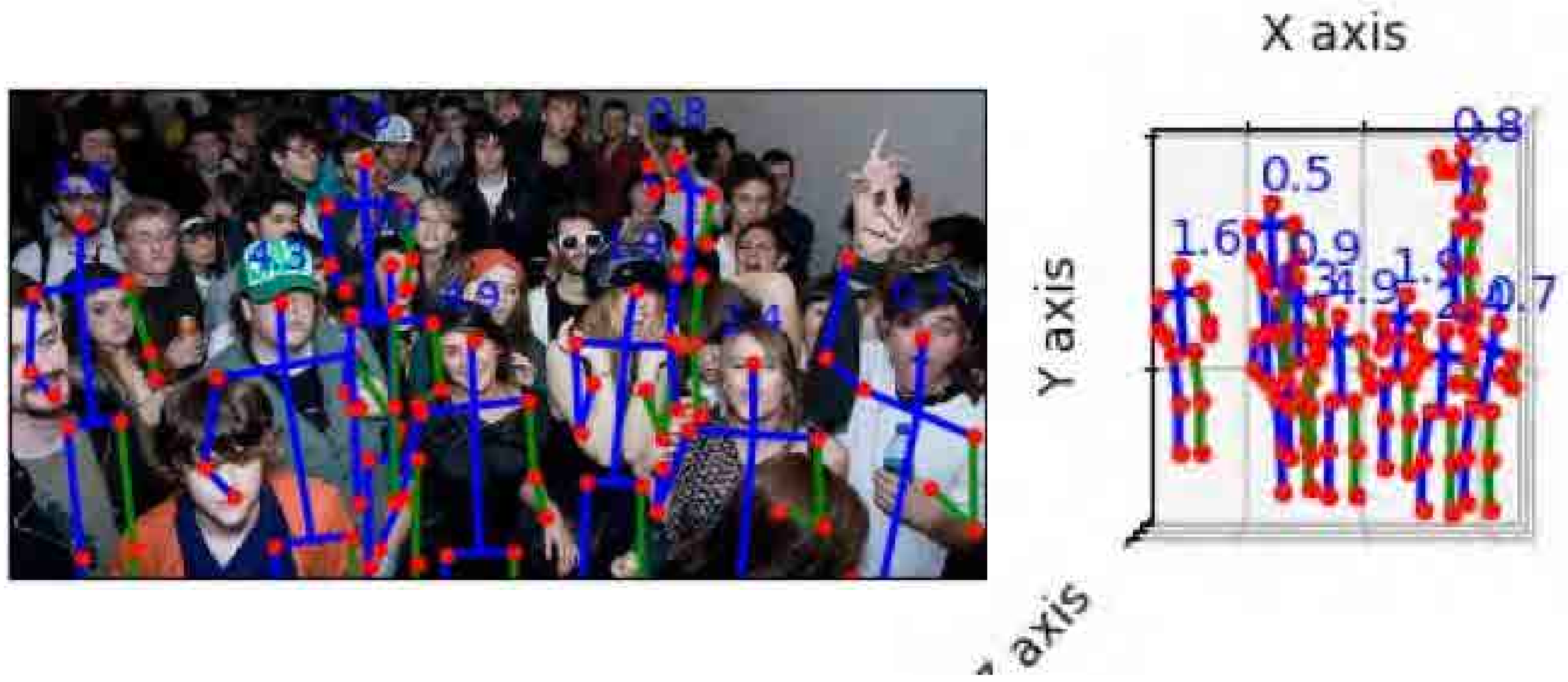
Failure cases

- Complex and/or unseen poses (danse, gymnastic)

- Misdetections when people are **overlapping** too much







3D surface prediction

DEVIEW 2019

What if we need more than just 3D keypoints?

- To create his own 3D full-body avatar (Gaming or AR/VR) applications
- For virtual **try-on for e-commerce**

A detailed 3D shape from a single image would be better.

3D representation

Most state-of-the-art methods rely on parametric models such as **SMPL**...

SMPL 3D shape

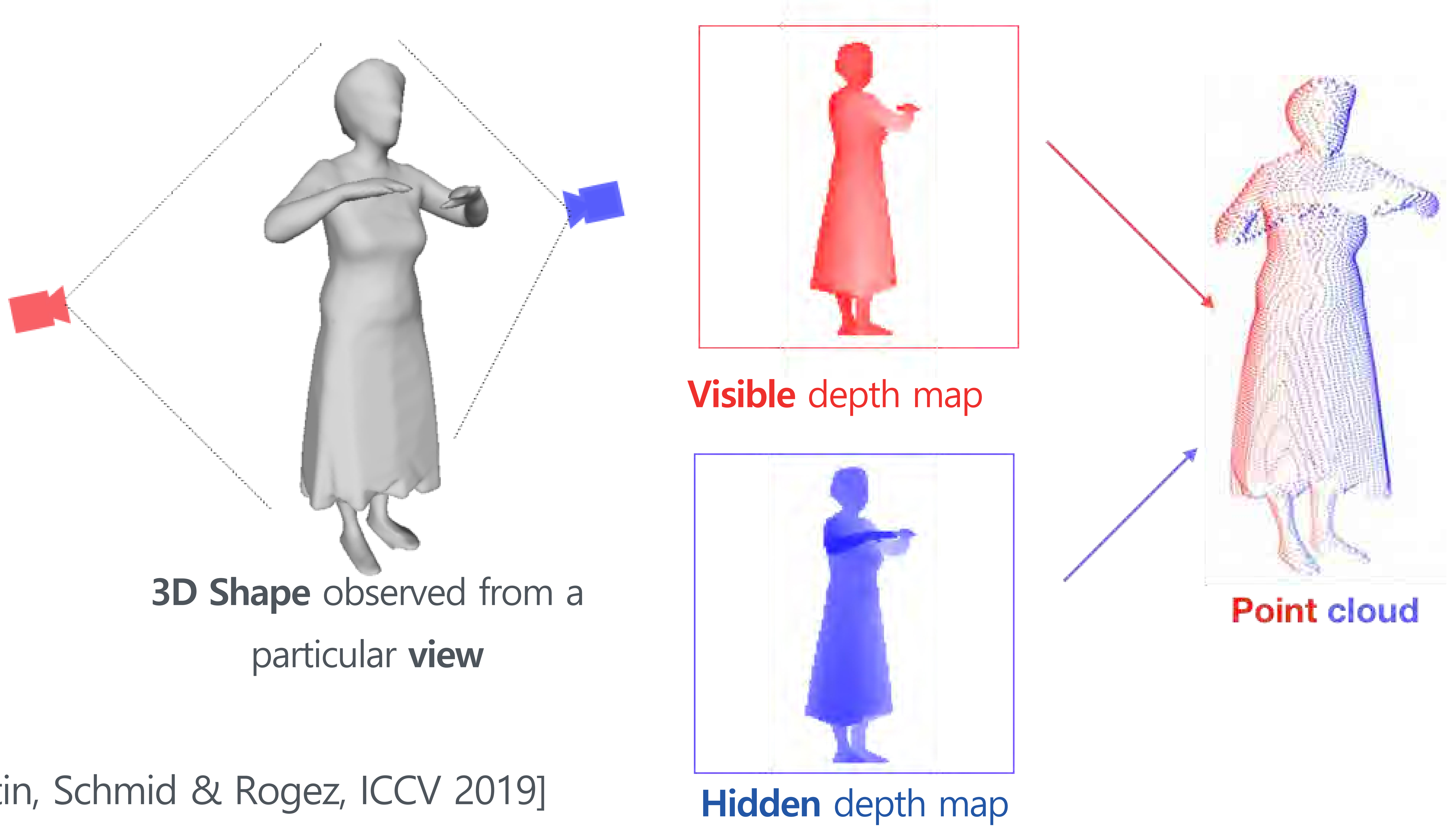
Detailed 3D shape

but these are limited to **naked body shapes** and cannot represent hair and clothes.

[SMPL: A Skinned Multi-Person Linear Model, Loper et al, ACM Trans. Graphics 2015]

3D representation

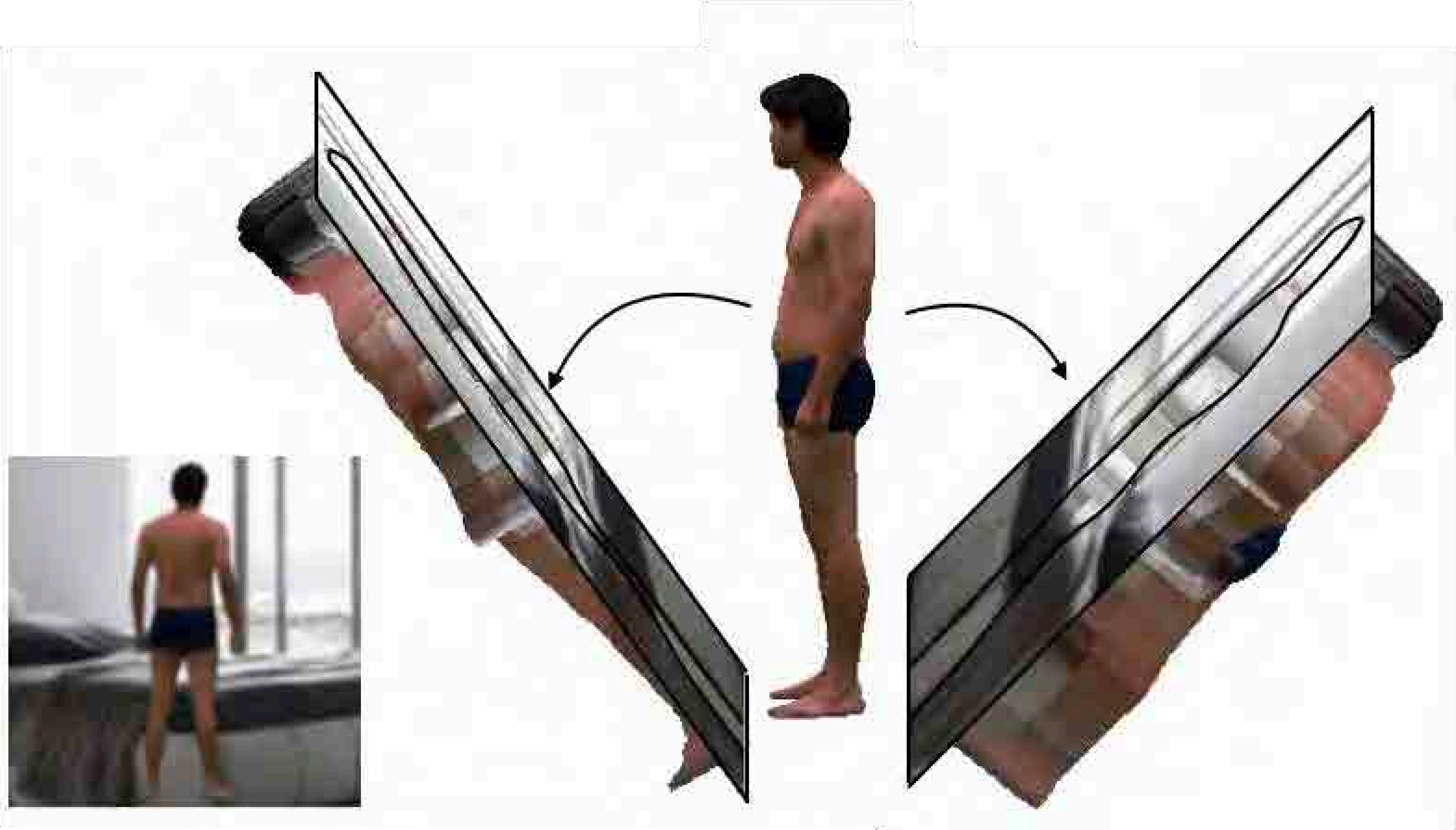
We propose a **non-parametric** approach that represents the 3D surface as the combination of 2 depth maps: the **visible** and the **hidden** depth map.



[Gabeur, Franco, Martin, Schmid & Rogez, ICCV 2019]

3D representation

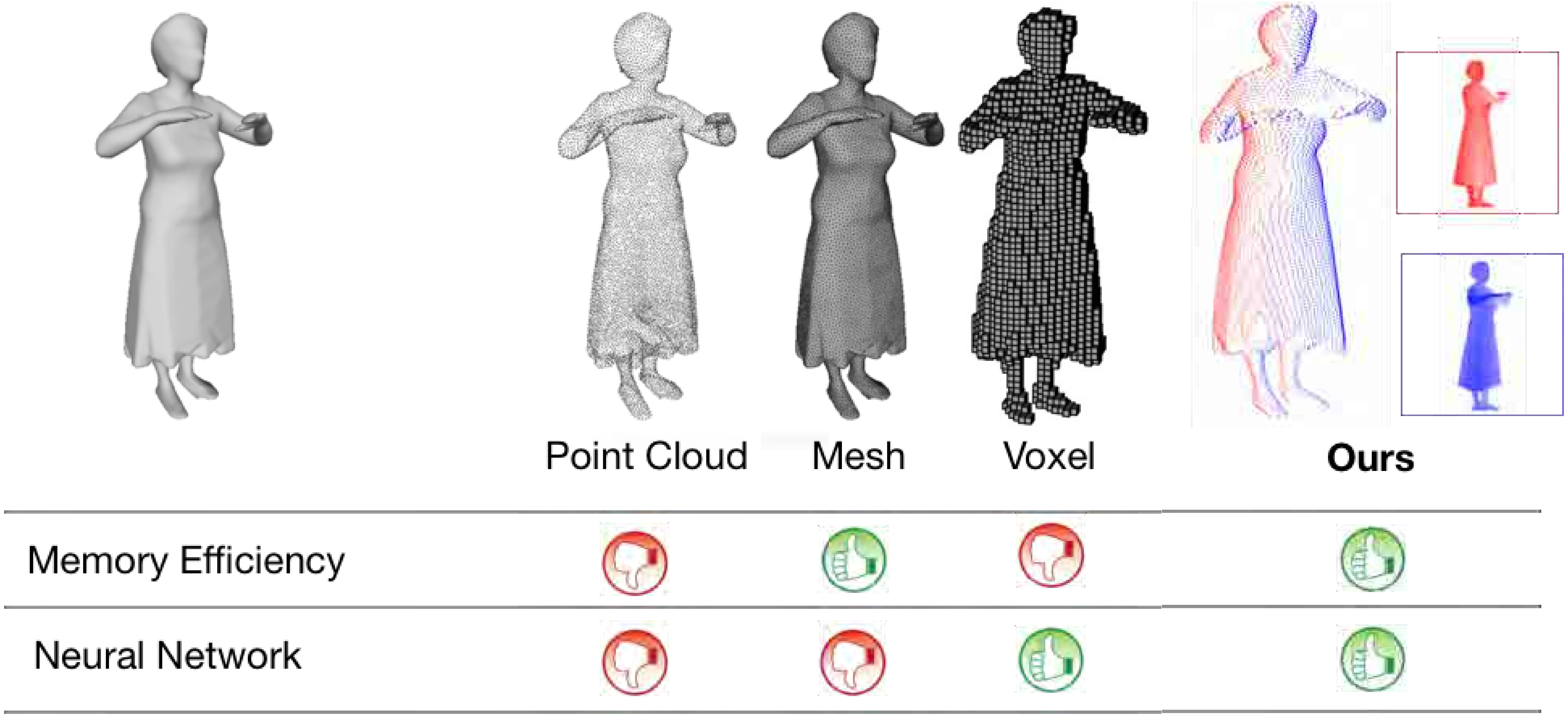
Our representation can be seen as the two halves of a mould…



[Gabeur, Franco, Martin, Schmid & Rogez, ICCV 2019]

3D representation

Other **non-parametric** representations include point clouds, meshes and voxel grids

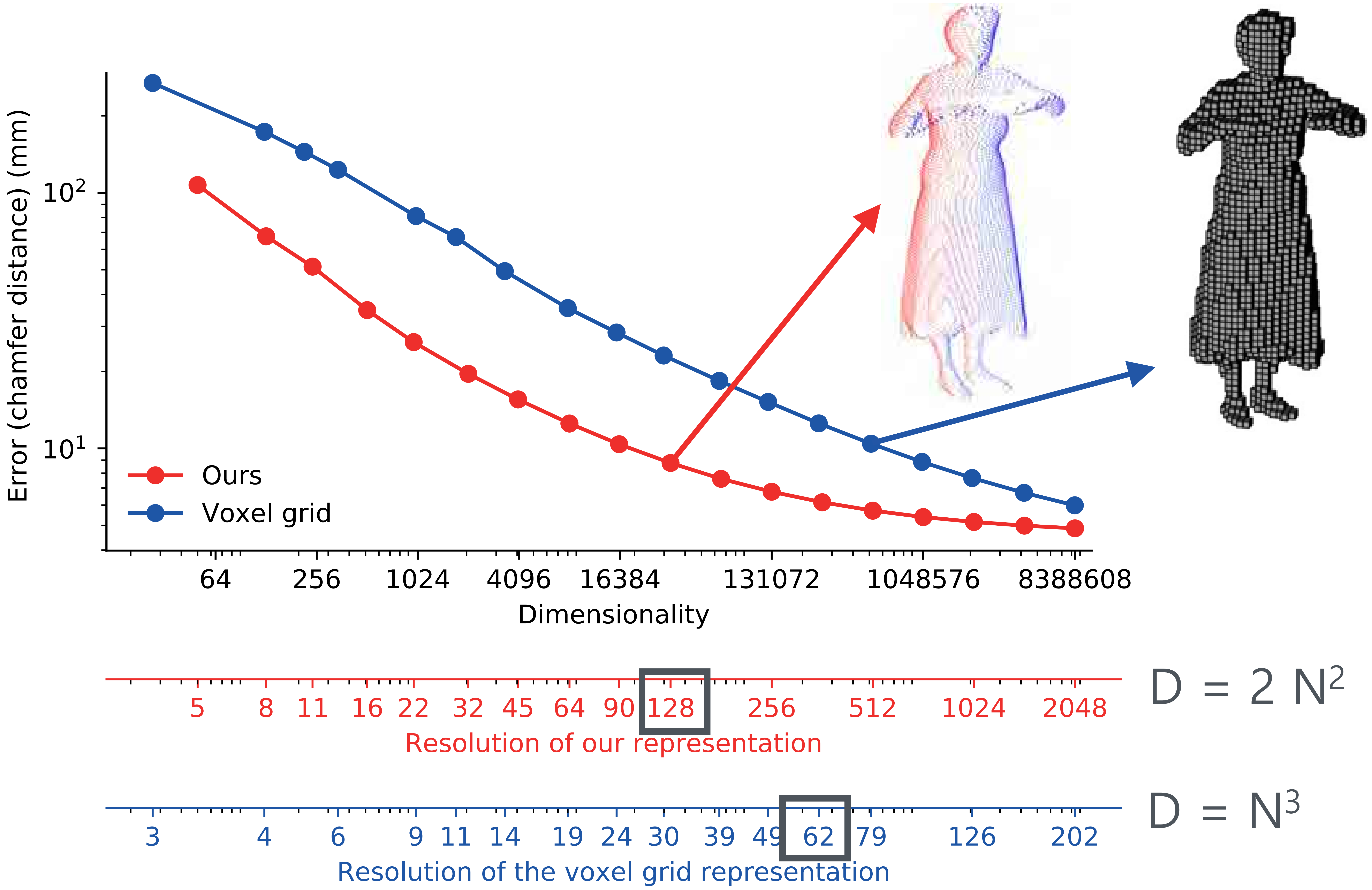


They are not memory efficient or are difficult to handle in a Neural Network.

Our representation is both memory **efficient and easy to handle**.

3D representation

We compare 3D reconstruction error with voxel grids of BodyNet [Varol et al, ECCV'18]

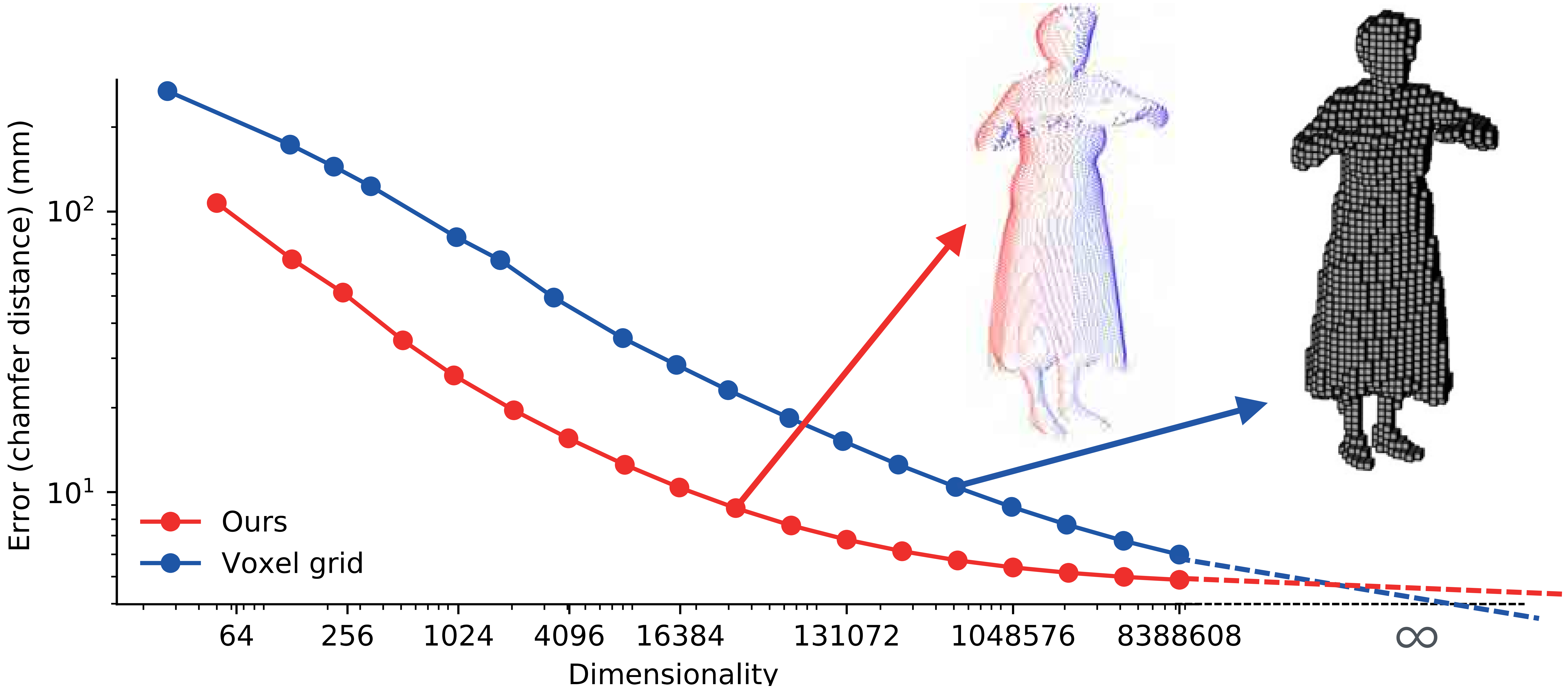


$D = 2\ N^2$

$D = N^3$

3D representation

We compare 3D reconstruction error with voxel grids of BodyNet [Varol et al, ECCV'18]

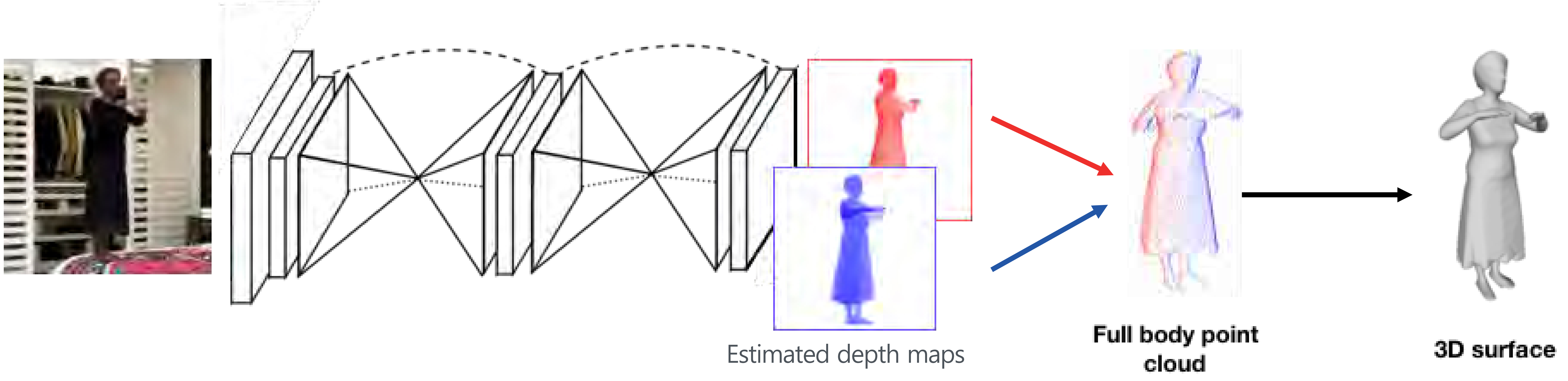


Voxel grids can attain a perfect result with an **infinity of voxels**.

For manageable sizes, our **representation** captures **more details**.

Architecture

We design a **double stacked hourglass** [25] network to estimate both **visible** and **hidden** depth maps from a 256 x 256 input image:



The 2 maps are combined to obtain a 3D surface using Poisson reconstruction [19].

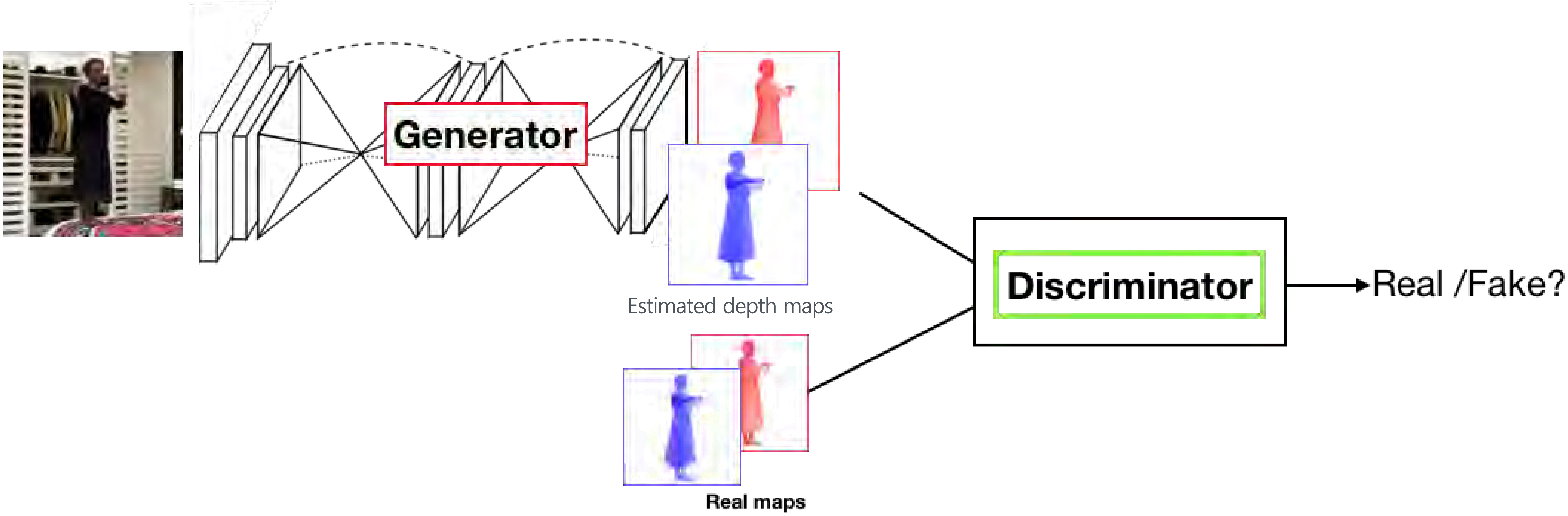
[25] Newell et al., Stacked Hourglass Networks for Human Pose Estimation, ECCV 2016

[19] Kazhdan and Hoppe, Screened Poisson Surface Reconstruction, ACM T. Graph. 2013

Adversarial Training

To improve the accuracy and "**humanness**" of the generated 3D output …

we incorporate a **discriminator in an adversarial manner**.



Dataset

To train and test our method, we generate a **new '3D HUMANS' dataset** of images showing **real persons** in movement with **ground truth 3D shapes**.

Results on our test set

For 3 subjects with varied clothing

Our 3D representation allows a direct **mapping of the image texture** to the surface.

Qualitative results with videos

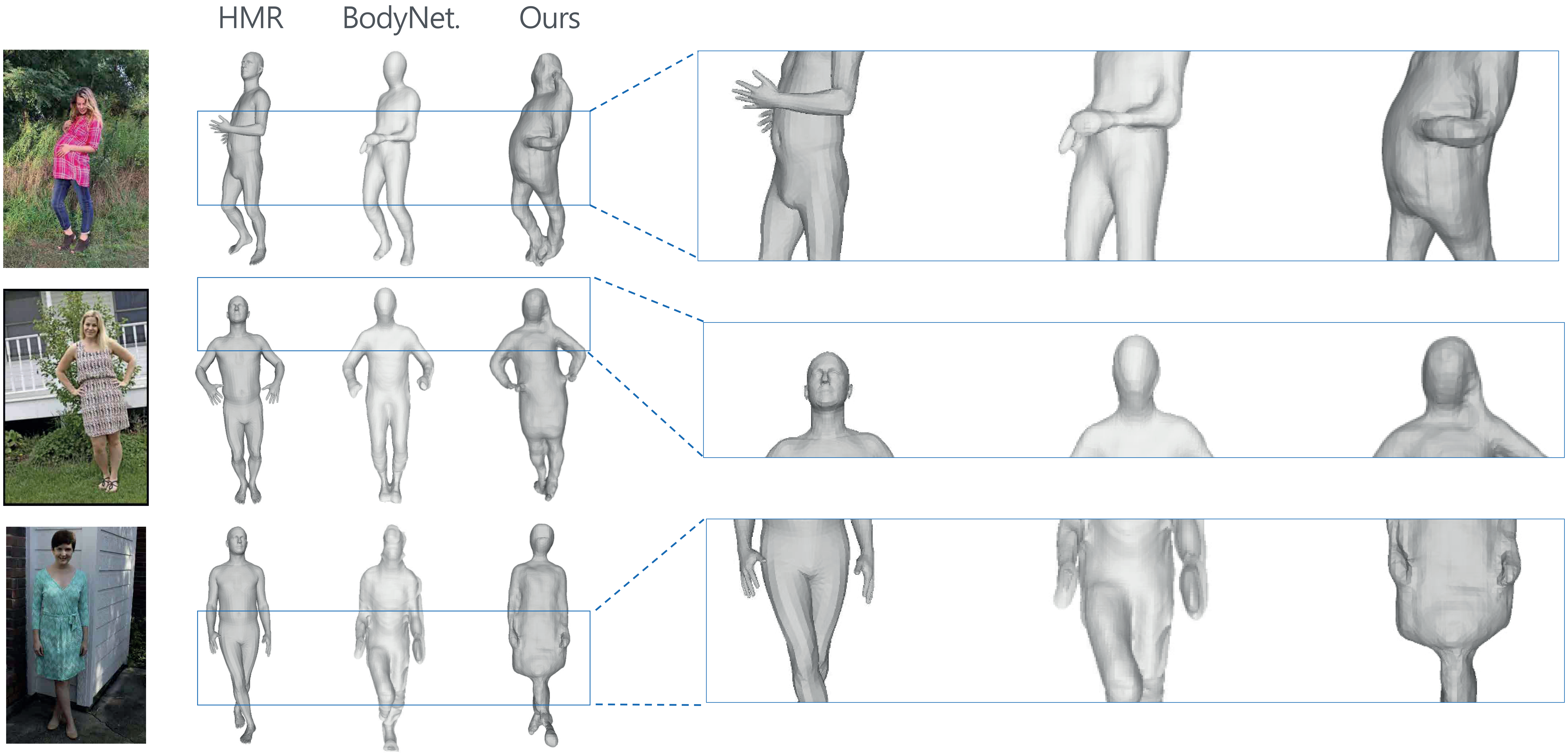
DEVIEW
2019

3D Shape

Textured
3D Shape

Comparaison with state of the art

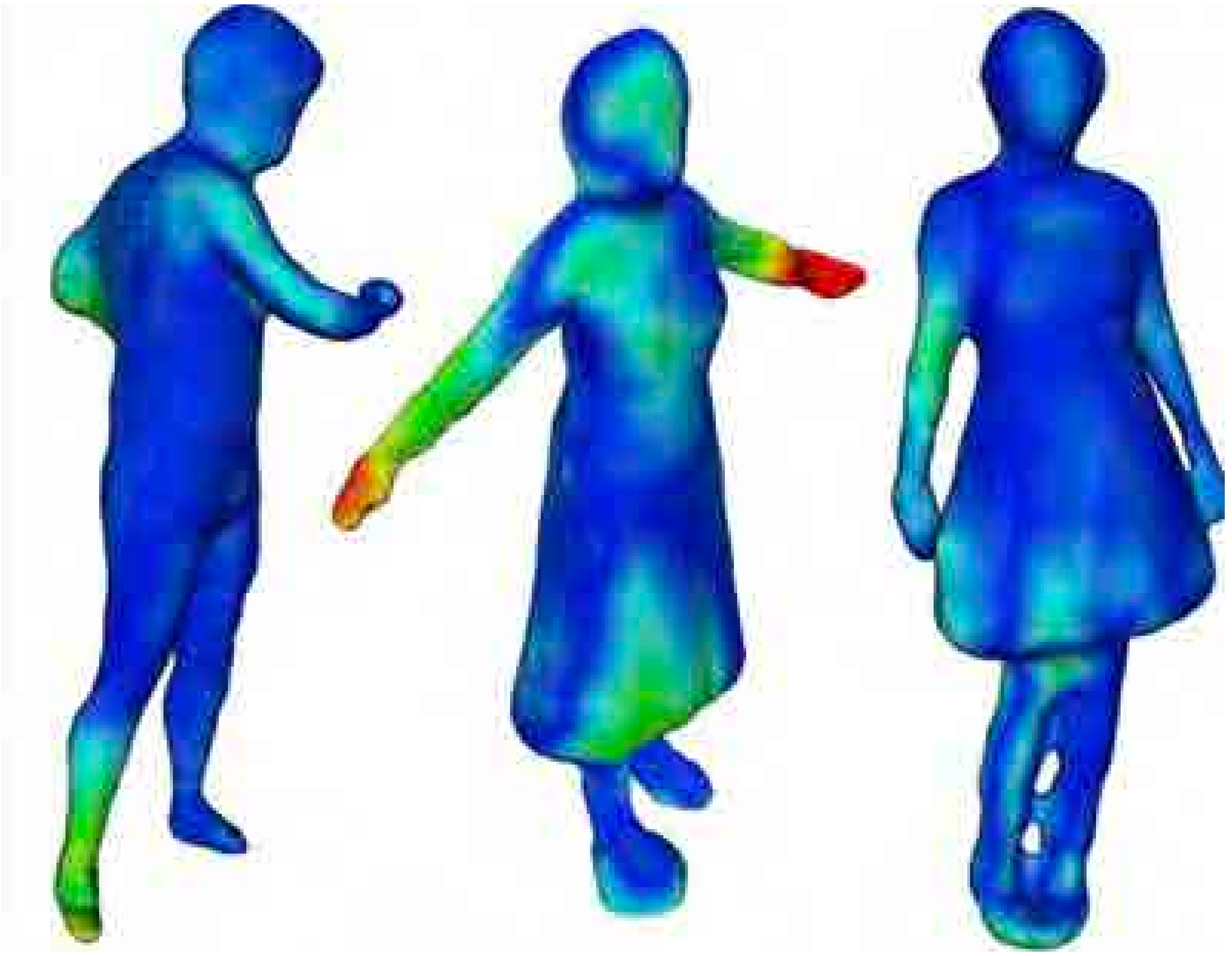
DEVIEW 2019



3D HUMANS dataset

To numerically evaluate our method with realistic scenes, we also rendered our 3D meshes in **realistic 3D environments**:

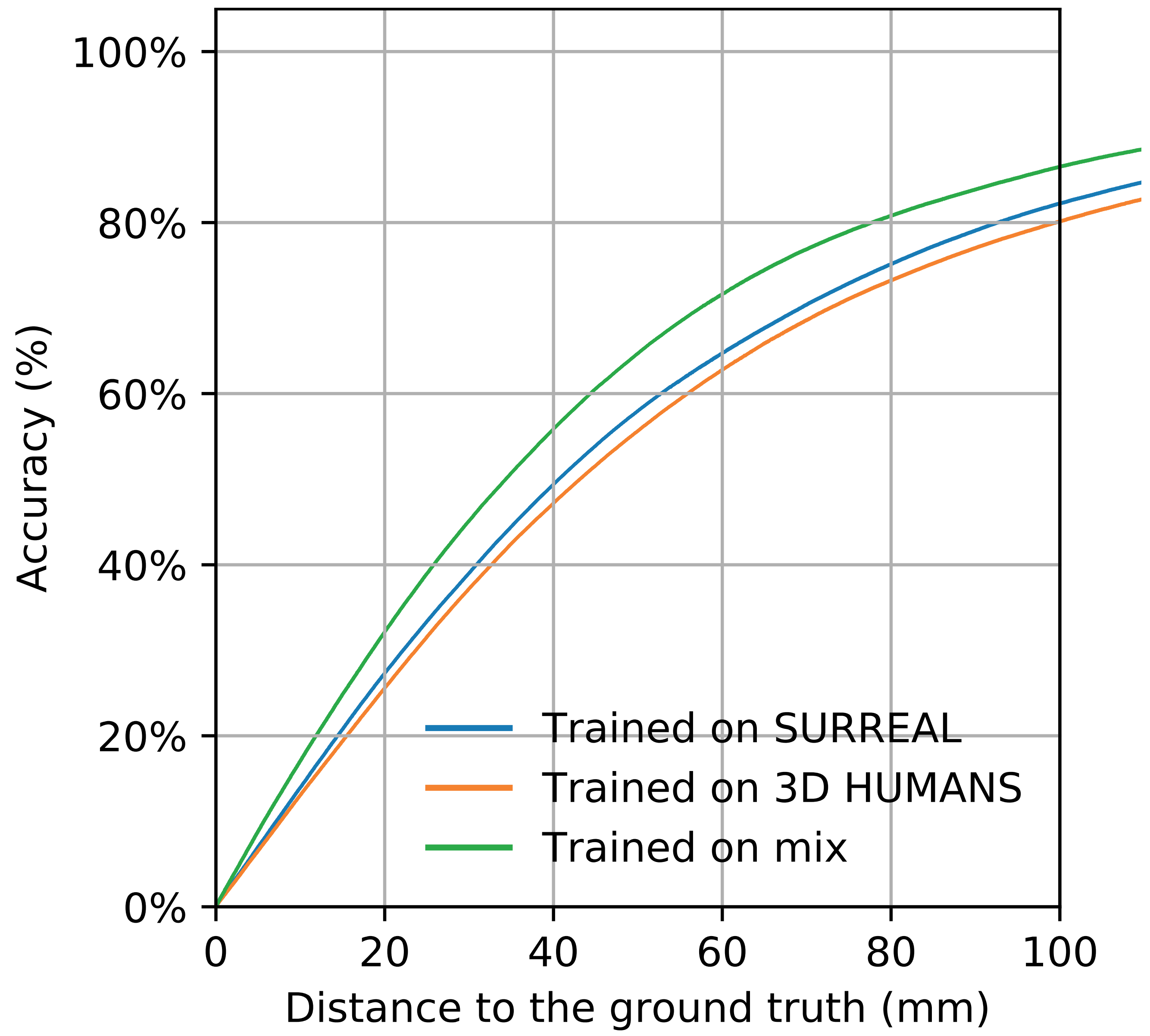
3D HUMANS dataset



Our results show that:

1) this realistic data is **more difficult**

2) adding **synthetic training** data from SURREAL [12] **helps generalisation**.



[12] Varol et al. Learning from Synthetic Humans. CVPR 2017

Results in the wild

3D Shape

Textured
3D Shape

4.Ongoing research and applications

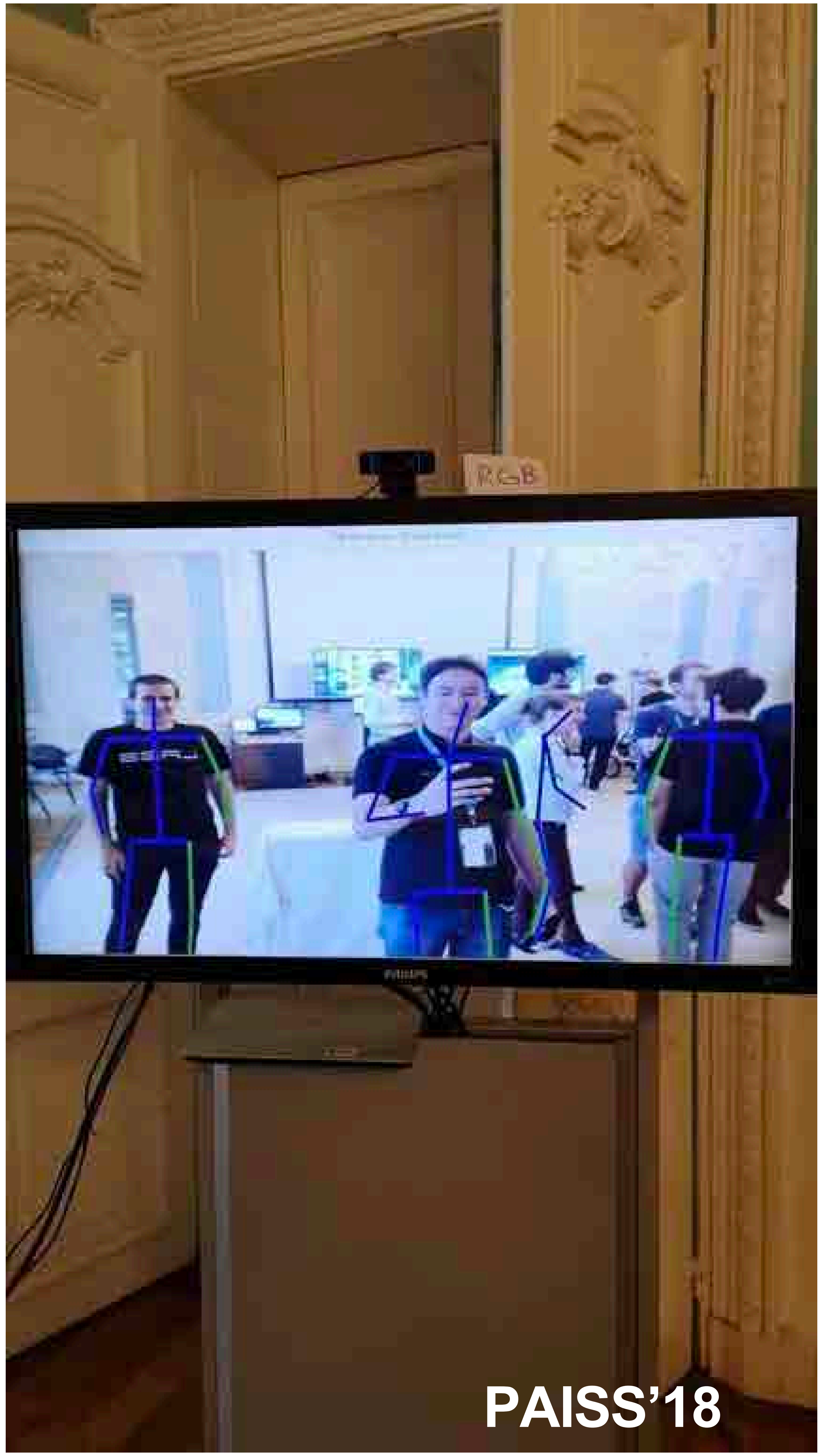
3D surface prediction

Ongoing work:

- More robust & accurate with video
- Scenes with multiple persons
- Occlusions

Our human pose detector:

- Full-body pose
- 2D and 3D pose
- Multi-person
- Real-time
- State-of-art in 3D pose



LCR-Net: Localization-Classification-Regression for. Human Pose. Rogez et al. CVPR 2017

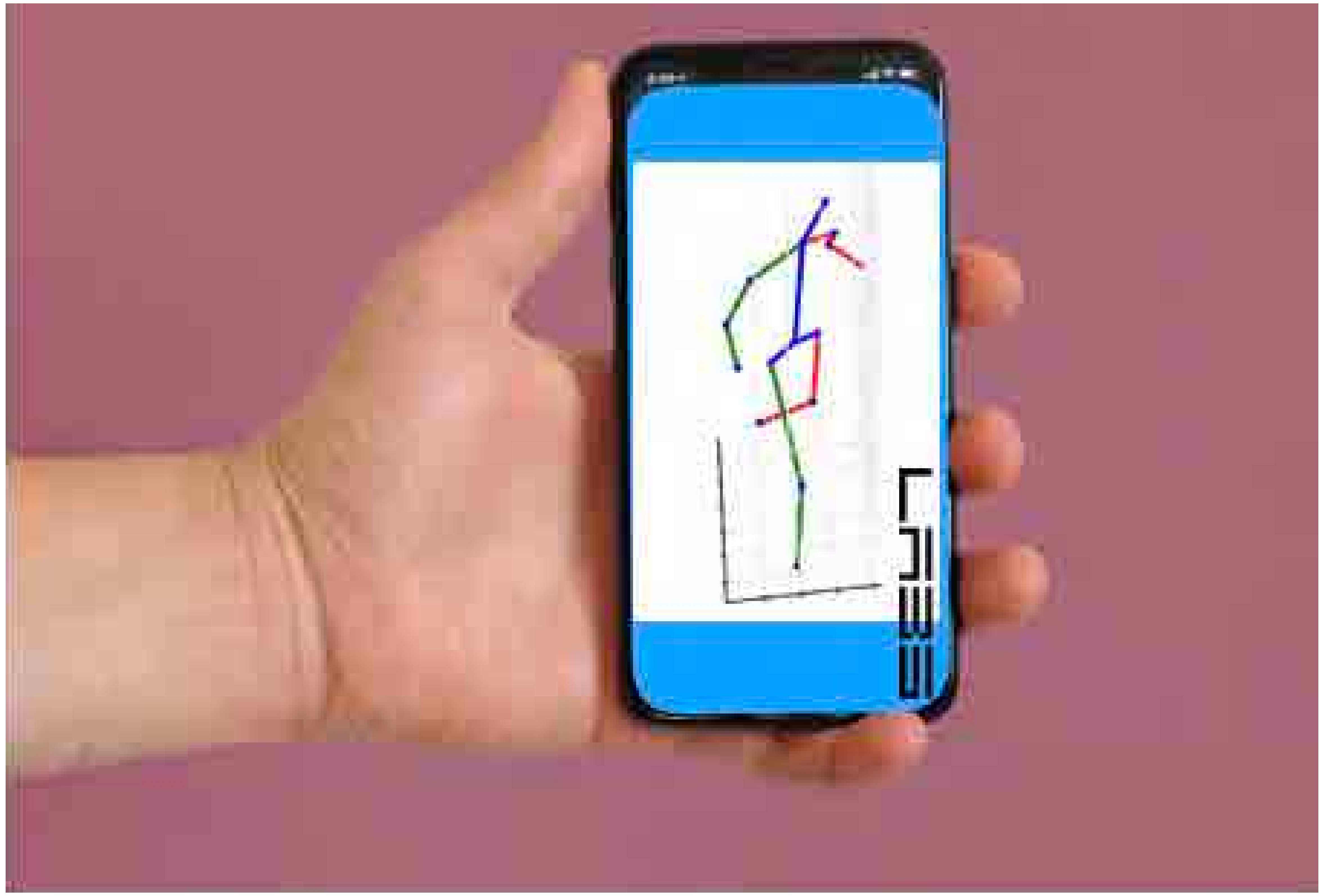
LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. Rogez et al. TPAMI 2019

Robot animation from human pose

Fast human pose on mobile

Ongoing work:

- Robust & faster on Mobile
- **Lighweight** model
- **Distillation** method to train a smaller network to perform as well as a big one



Action Recognition

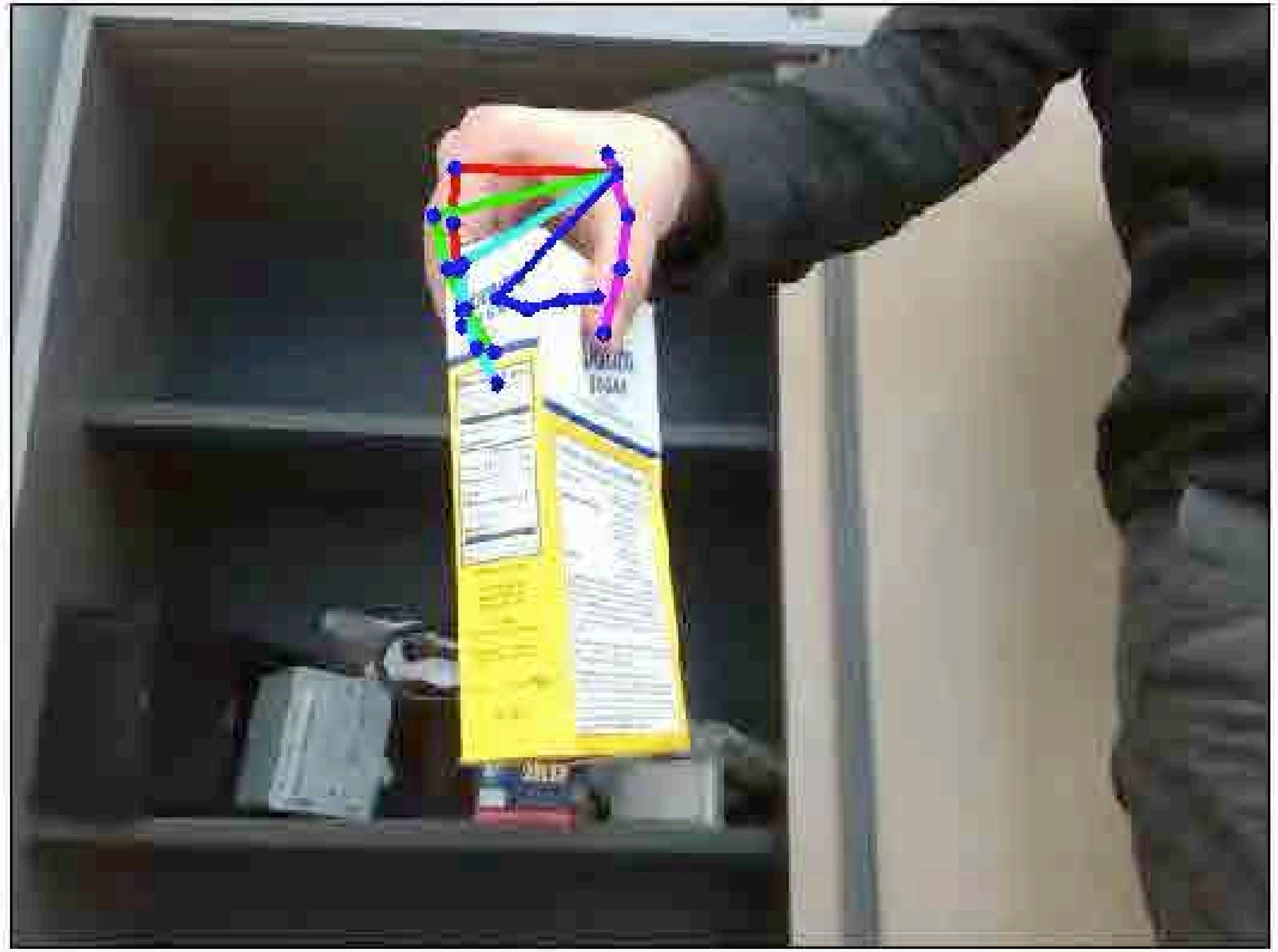
Human can understand Mimes… without context & objects

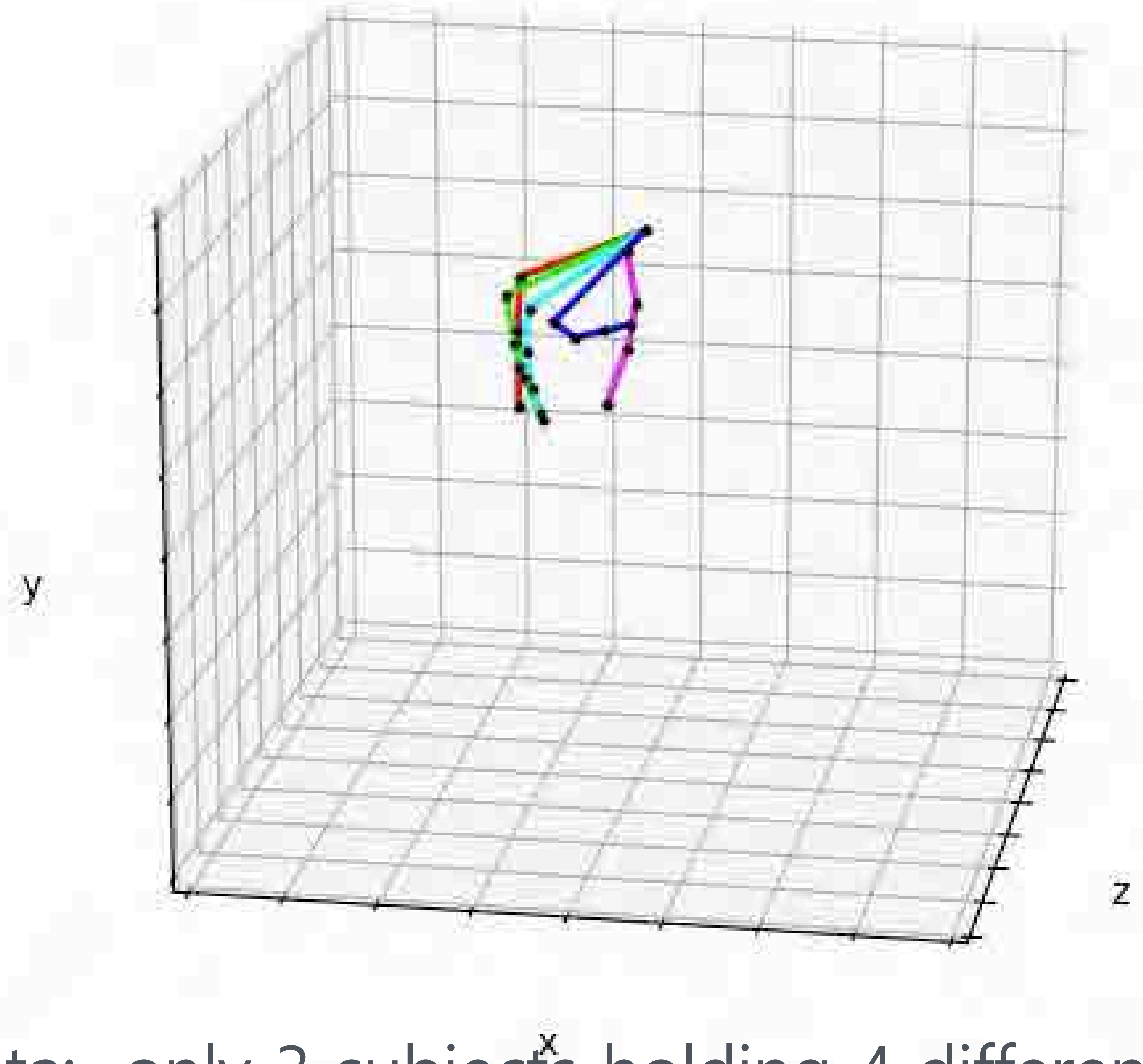
Current methods are biased and not robust

F

Further Understanding Human Actions. Weinzaepfel and Rogez, to be submitted, 2019

3D Hand pose estimation





Results obtained after training on very limited data: only 3 subjects holding 4 differents objects and no synthetic data...

3D Hand pose estimation

ICCV HANDS 2019 challenge

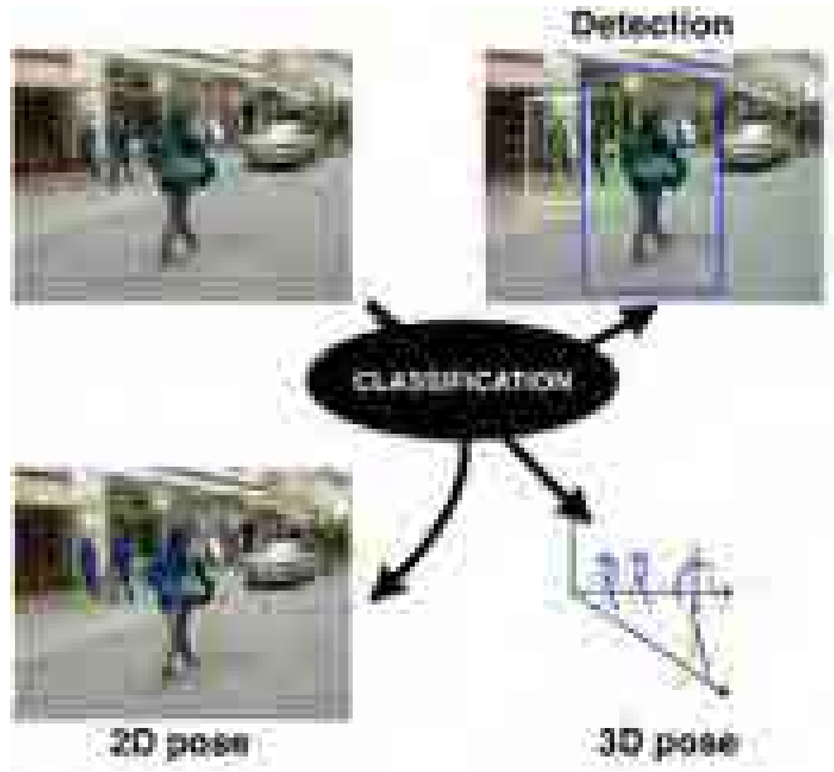
Team Name	EXTRAP. ▲	INTERP. ▲	OBJECT ▲	SHAPE ▲
& meat	24.74 (1)	6.70 (3)	27.36 (2)	13.21 (1)
NLE	29.19 (2)	4.06 (1)	18.39 (1)	15.79 (3)
BT	31.51 (3)	19.15 (5)	30.59 (3)	23.47 (4)
Hasson et al, CVPR'19	38.42 (4)	7.38 (4)	31.82 (4)	15.61 (2)
BT	41.81 (5)	23.52 (6)	72.70 (8)	35.43 (6)
kin	49.64 (6)	46.78 (7)	53.79 (6)	51.32 (8)
kin	57.45 (7)	47.82 (8)	54.81 (7)	50.05 (7)
xteam	80.06 (8)	5.66 (2)	45.34 (5)	29.84 (5)
citrus	80.06 (8)	5.66 (2)	45.34 (5)	29.84 (5)

- 1st in interpolation, i.e. when object & hand pose have been seen before, or when object only is known

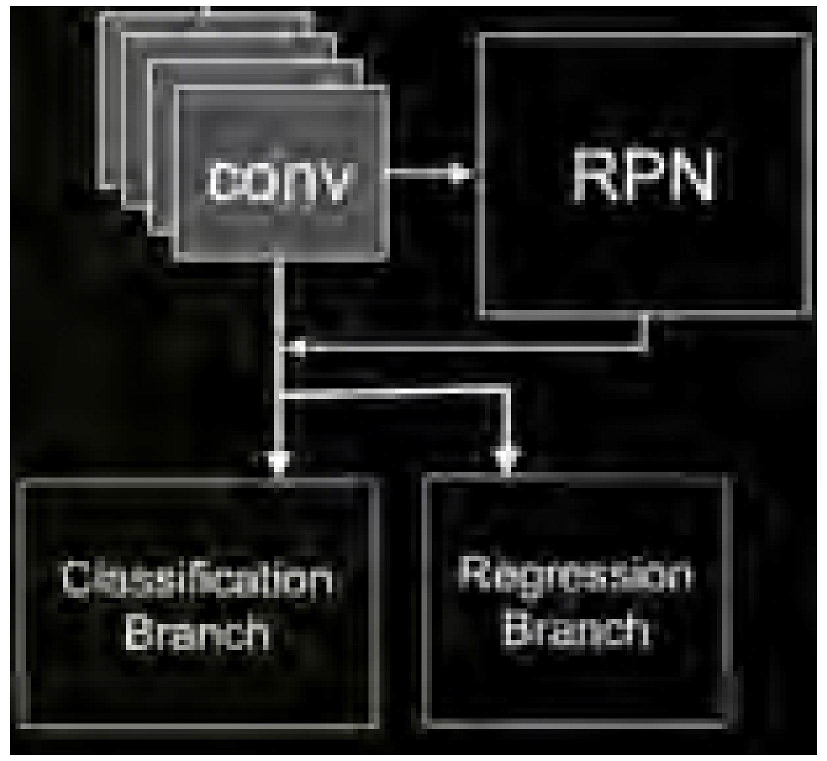
- Overall we outperform CVPR'19 paper by a large margin !

5. Take-home message

Take-home message



Pose detection: localization + 3D/2D pose
- model **multi-modal** distributions
- **holistic** full-body approach



LCR-Net with class-specific regression:
- reduced number of classes (computation), refine the 2D/3D pose
- Very **good interpolating**, less in extrapolation
- Needs in-the-wild data!!

Mould representation for 3D surface:
- **detailed** shape including clothes (non-parametric)
- efficient and **easier to handle** in NN
- Also needs data!!

Collaborators

DEVIEW 2019

Philippe Weinzaepfel
NAVER LABS Europe

Romain Bregier
NAVER LABS Europe

Valentin Gabeur
Inria / Google

Hadrien Combaluzier
NAVER LABS Europe

J.S. Franco
Inria

Cordelia Schmid
Inria / Google

Q & A

Thank You