# 3D Human Sensing from monocular visual data using classification techniques

**Grégory Rogez**

Senior Research Scientist - Computer Vision Team Lead

NAVER LABS Europe

May 2022

# OUTLINE

- Background
- Monocular 3D Human pose estimation
- Classification-based approaches
- Drawbacks and solutions
- and beyond…

# OUTLINE

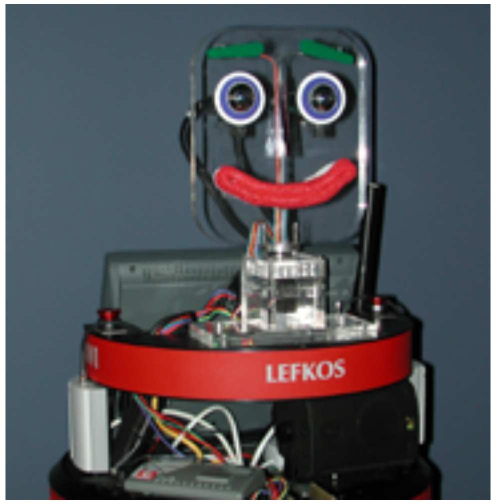
- **Background**
- Monocular 3D Human pose estimation
- Classification-based approaches
- Drawbacks and solutions
- and beyond…

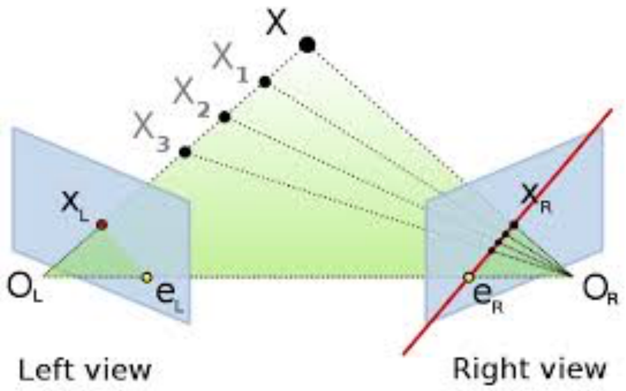
**Eng. Physics**

2002

- GE Medical Systems - Image quality in X-ray

- ICS-FORTH - Stereoscopic vision



# BACKGROUND: OCR & VIDEO-SURVEILLANCE

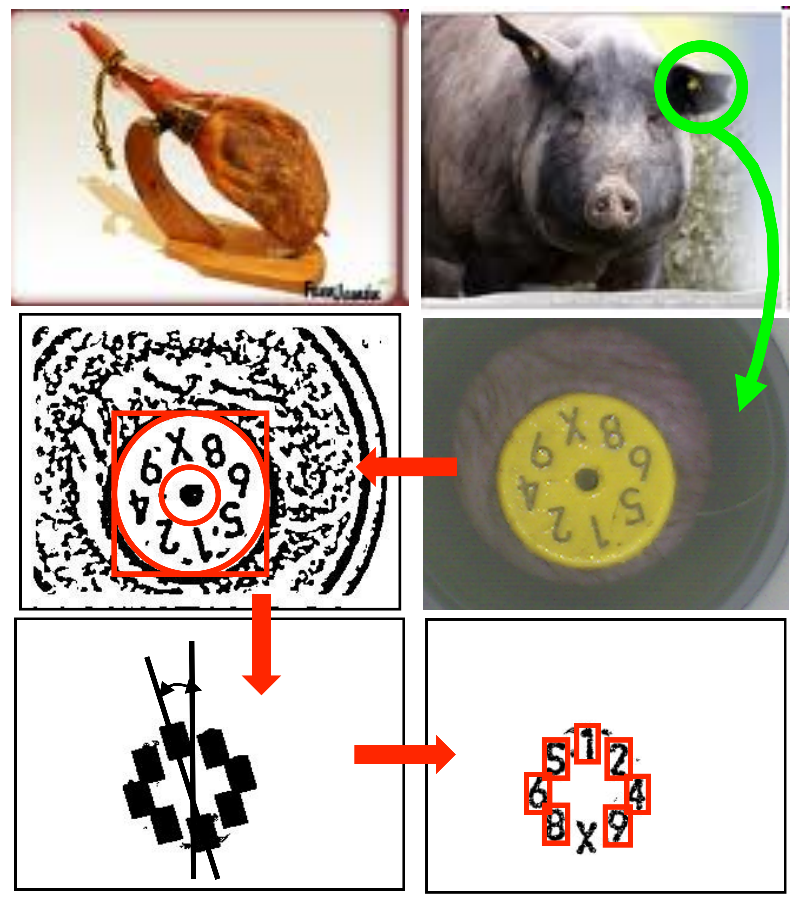
**Eng. Physics**

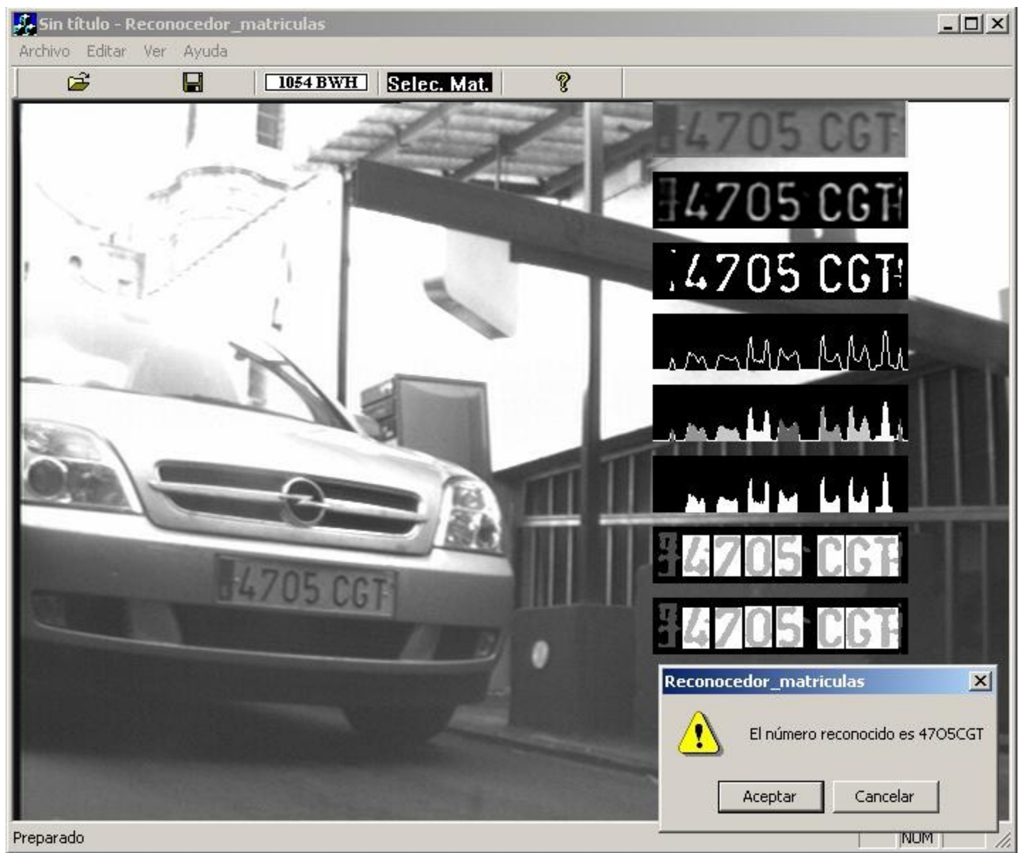
2002

**Engineer (UZ)**

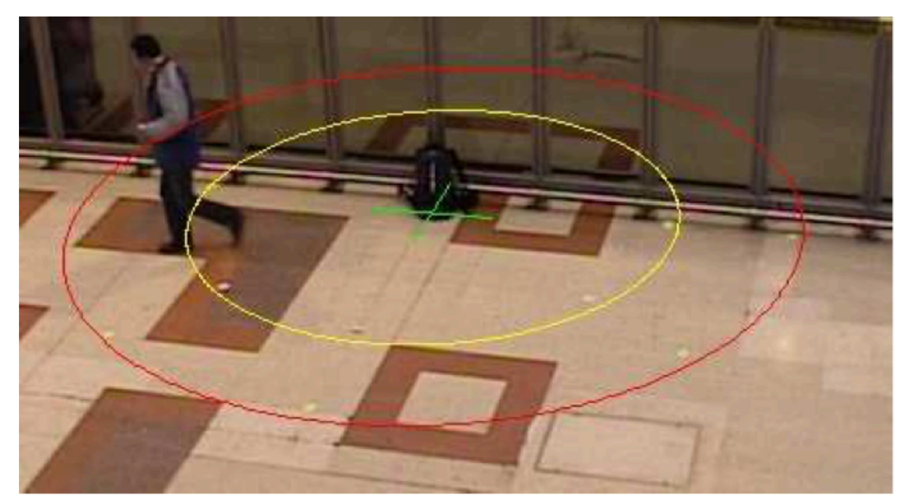
2004

- Optical character recognition






- Video-surveillance



# BACKGROUND: LOOKING AT HUMANS

**Eng. Physics**

2002

**Engineer (UZ)**

2004

**MSc (UZ)**

2006

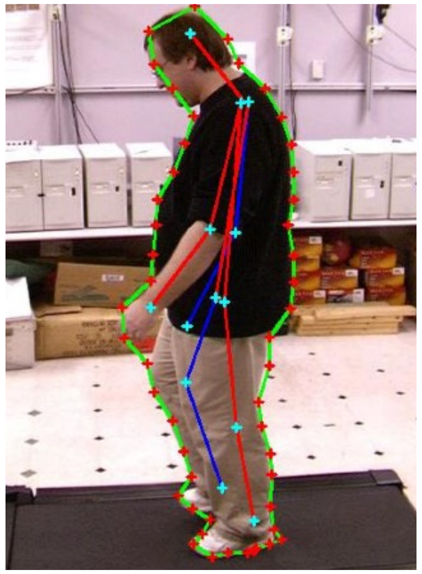
**PhD (UZ)**

2012

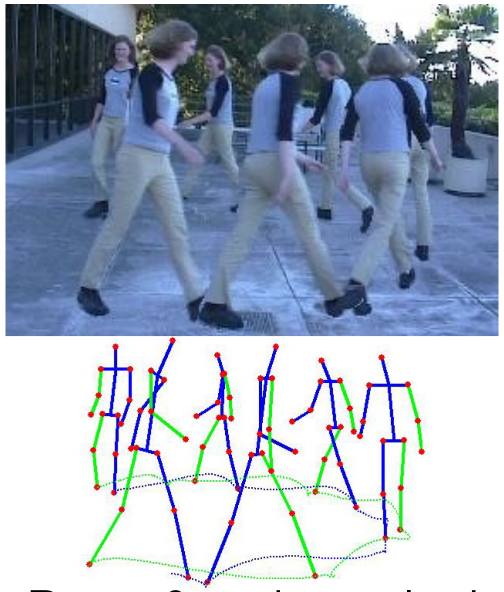
**Postdoc (UCI)**

2015

**Research Scientist (Inria)**

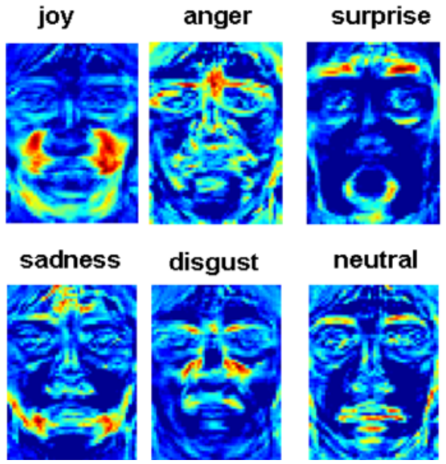


Silhouette - ASM

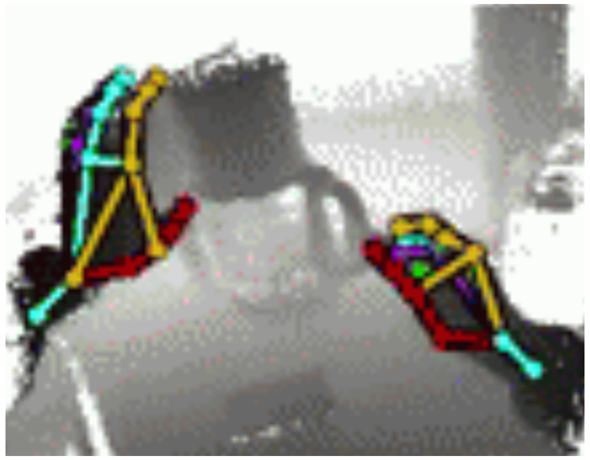


Pose & gait analysis

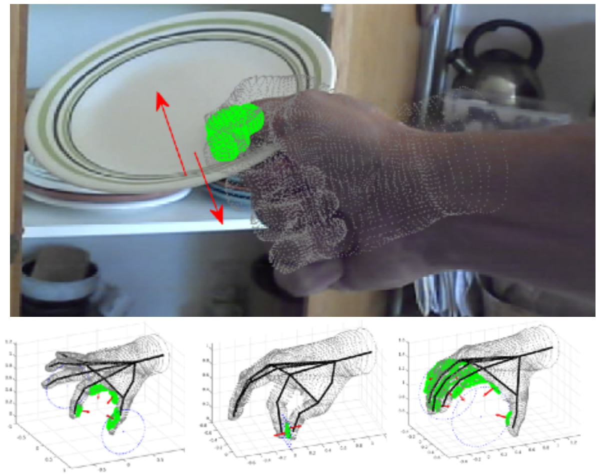
Detection & tracking




Facial expression

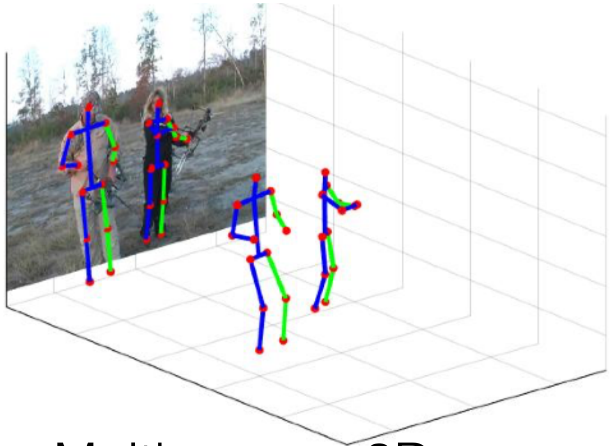


3D hand from depth



Hand-object interactions


Action recognition



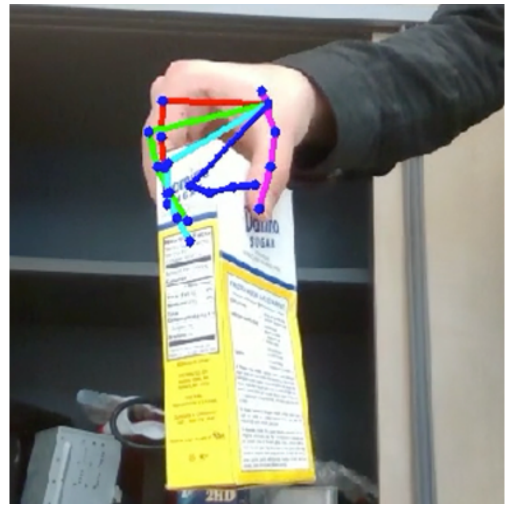
Multi-person 3D pose

3D shape from RGB

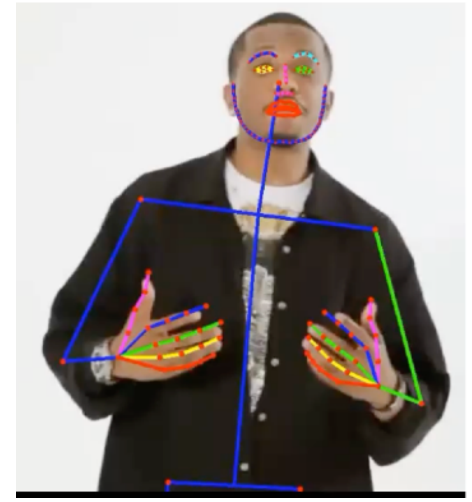
# BACKGROUND: NAVER LABS

2019

**Senior Research Scientist (NLE)**

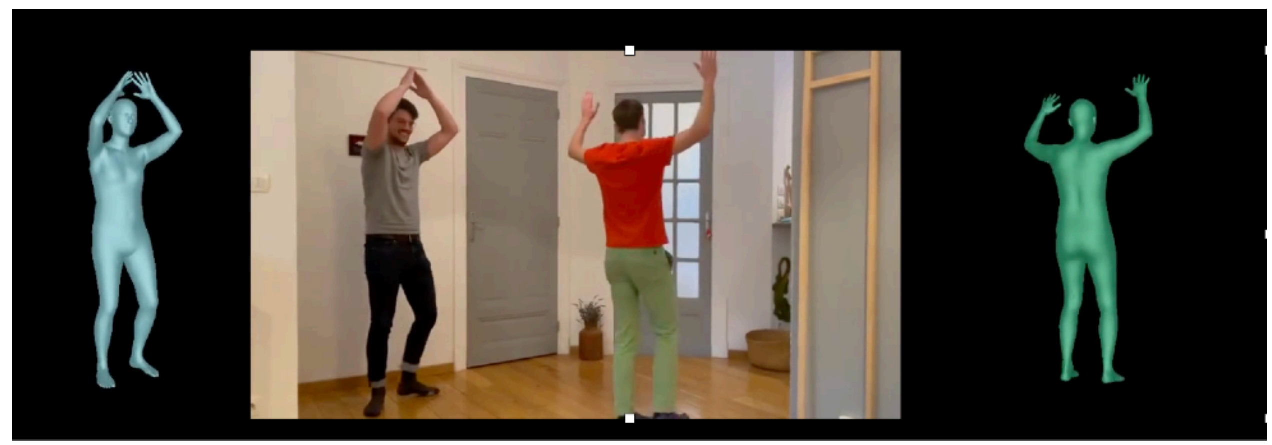


Hand pose estimation



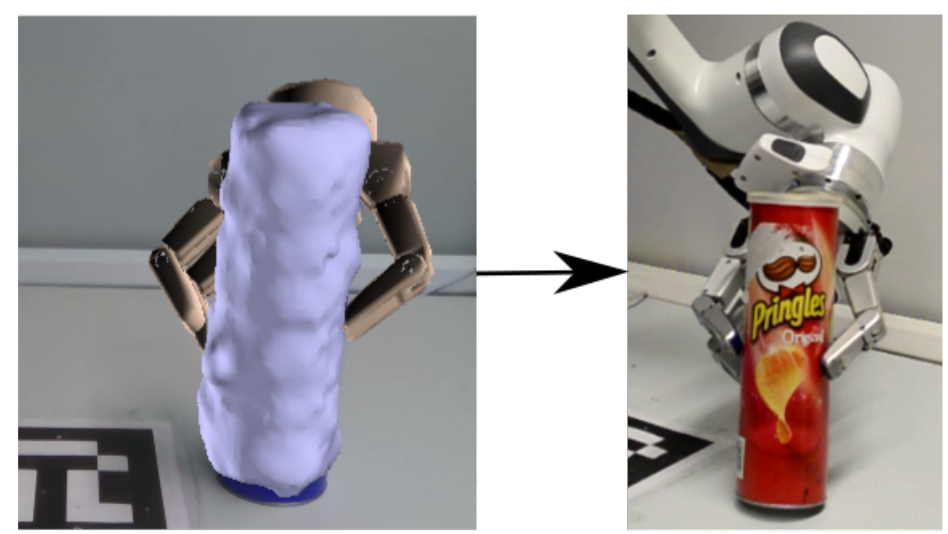
Whole-body pose


Human-robot interaction



3D Body mesh in videos

Grasp/affordance prediction



Robotic grasping

# NAVER LABS EUROPE

We are the European team of NAVER LABS which is the research arm of NAVER, Korea's leading internet company and the part of NAVER responsible for creating future technology.

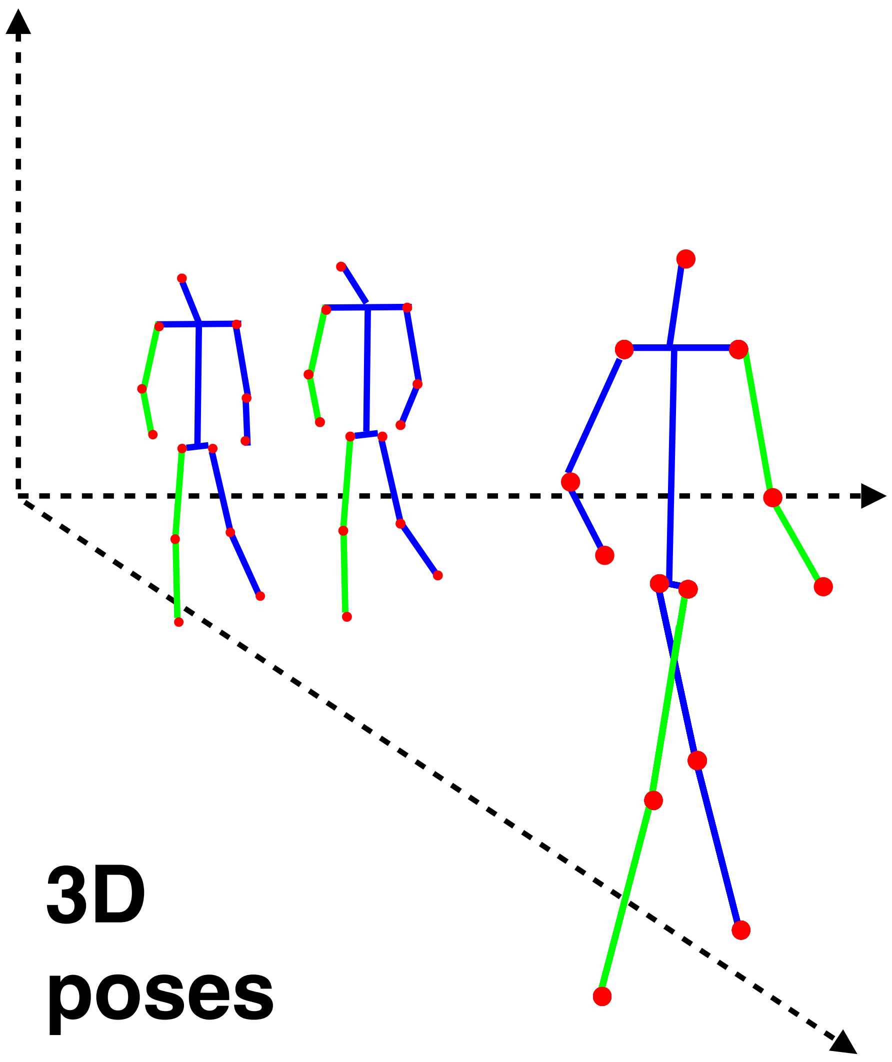
Our scientists conduct fundamental and applied research in the fields of machine learning, computer vision, natural language processing and UX and ethnography. The two main areas of application of research are 'AI for Robotics' and 'AI for our Digital World'.

NAVER LABS Europe is the biggest industrial research lab in AI in France.

# OUTLINE

- Background
- **Monocular 3D Human pose estimation**
- Classification-based approaches
- Drawbacks and solutions
- and beyond…

# MONOCULAR 3D HUMAN POSE ESTIMATION


"Articulated pose estimation is the task that employs computer vision techniques to estimate the configuration of the human body in a given image or a sequence of images". **Sarafianos et al., CVIU 2016**

# WHY IS IT INTERESTING?

- **Human-computer interactions**
- **Video-surveillance**
- **Gaming**
- **Physiotherapy**
- **Movies**
- **Dancing**
- **Proxemics**
- **Sports**
- **Human-robot interactions**

Slide courtesy of
**Sarafianos et al., 2016**

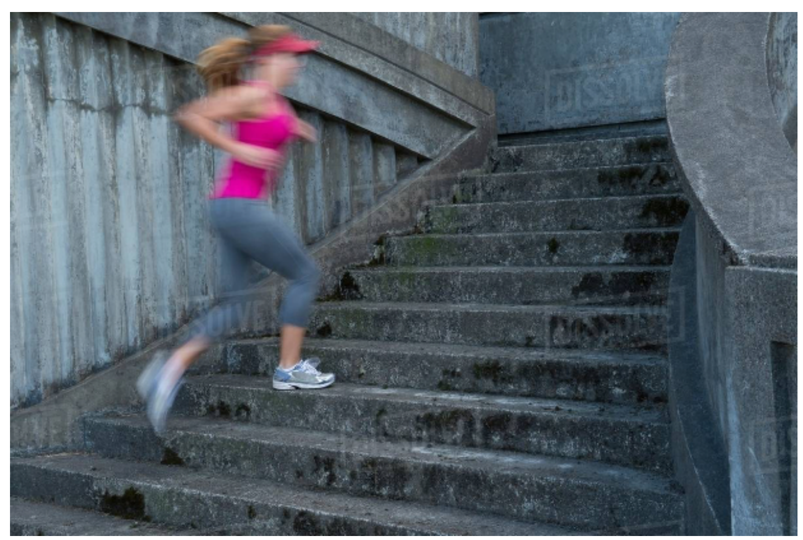
# WHY IS IT DIFFICULT?

variation in illumination

occlusion & clutter

body part foreshortening

variation in appearance



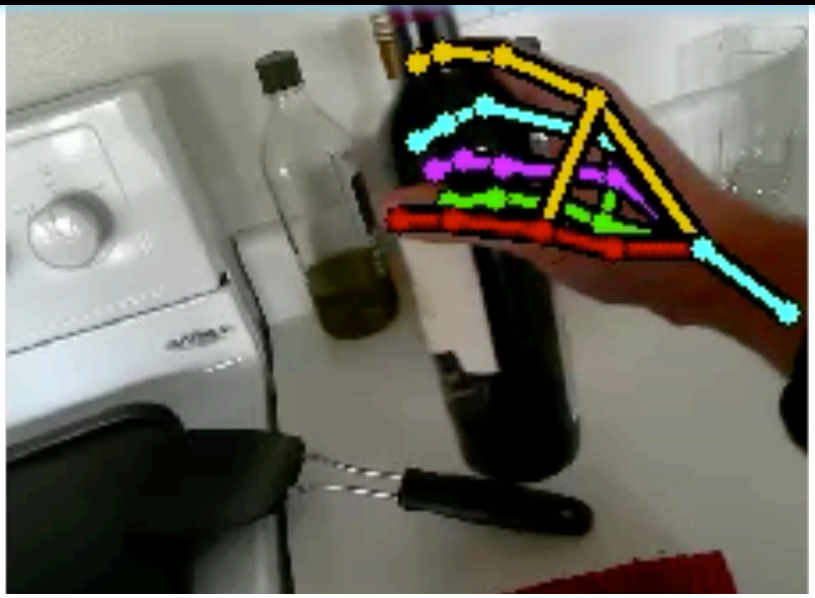
Motion blur

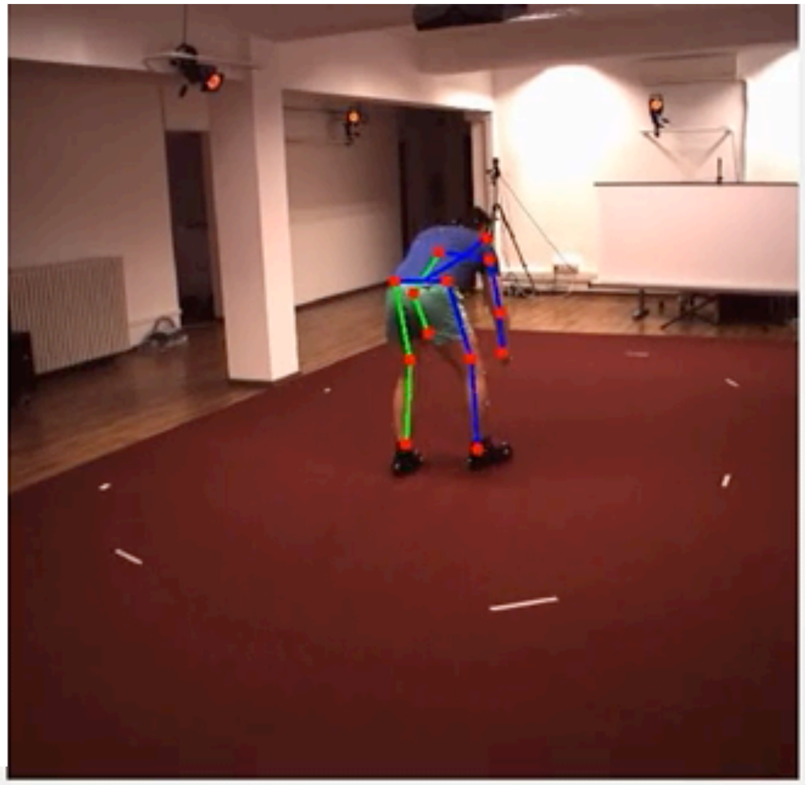
variation in pose, viewpoint

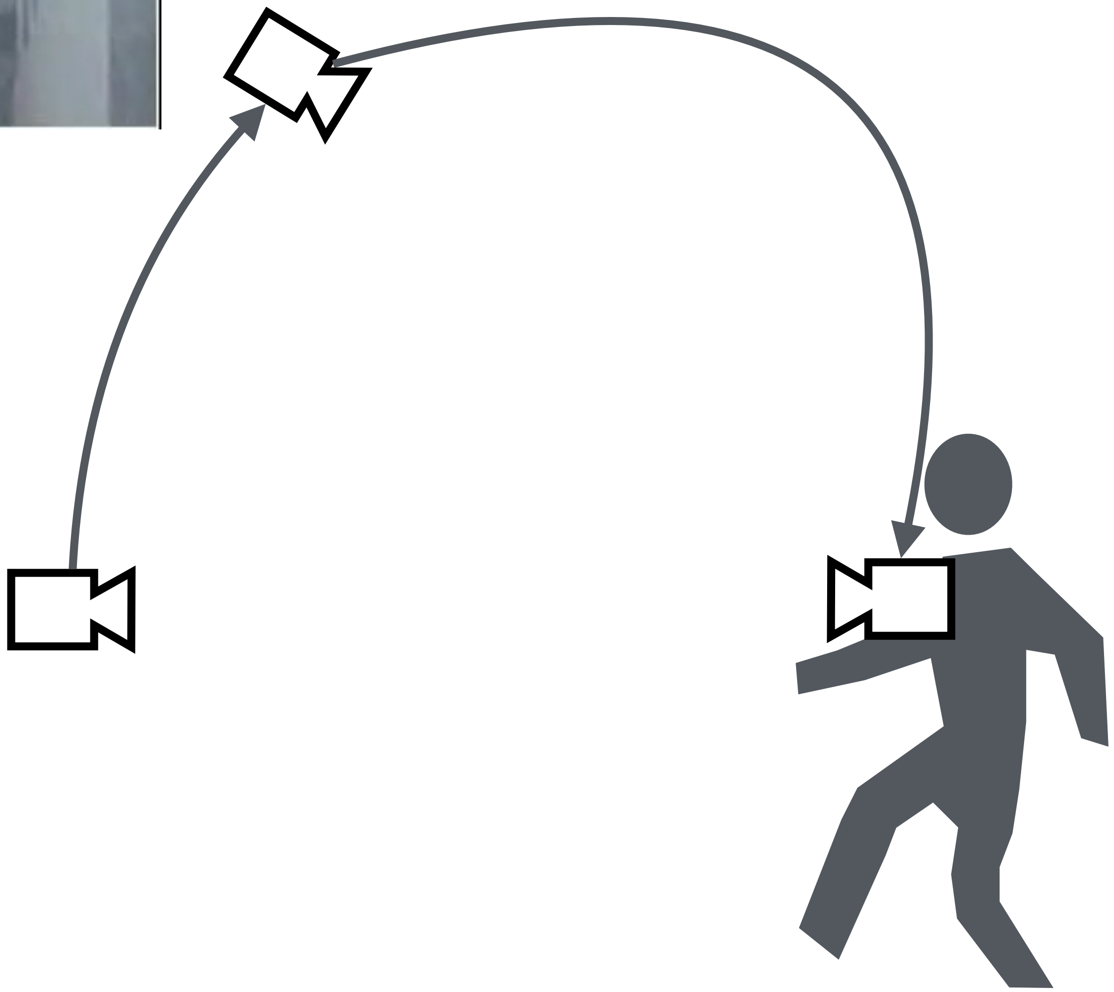
Classic “nuisance factors” for general object recognition

# LONG TAIL DISTRIBUTION

# IMPACT OF VIEWING ANGLE



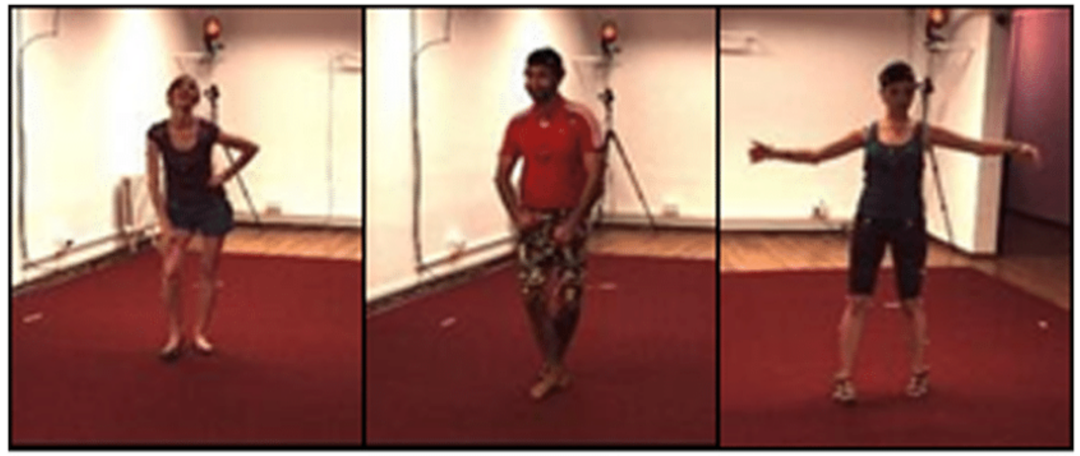




# GROUNDTRUTH 3D POSE DIFFICULT TO OBTAIN

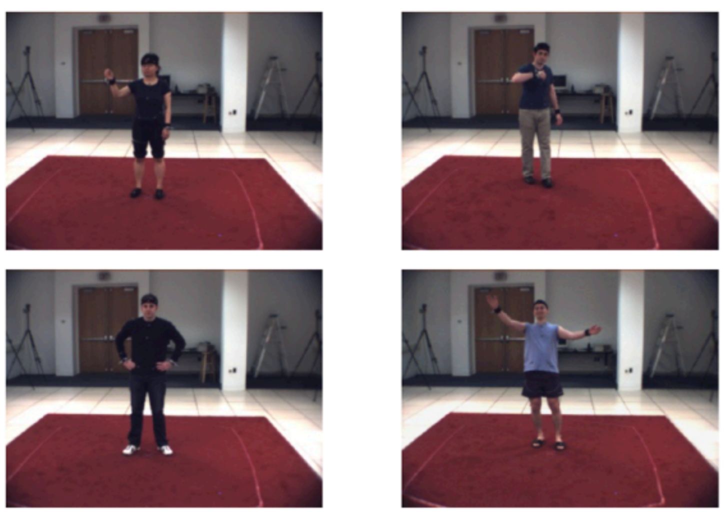
# 3D POSE DATASETS OVER TIME

CMU Graphics Lab
Motion Capture DB:
- 2500 sequences
- No images



Human3.6M:
~200 sequences, 11 subjects, 4 cameras
Images in controlled env.

AGORA:
~17k images/scenes
4k 3D scans
semi-synthetic images



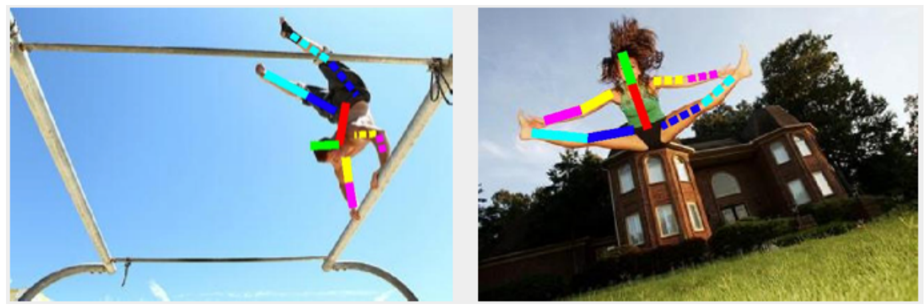
HumanEVA:
7 sequences, 4 subjects
Images in controlled env.

3DPW:
~60 sequences
Pseudo 3D groud-truth
Images in the wild

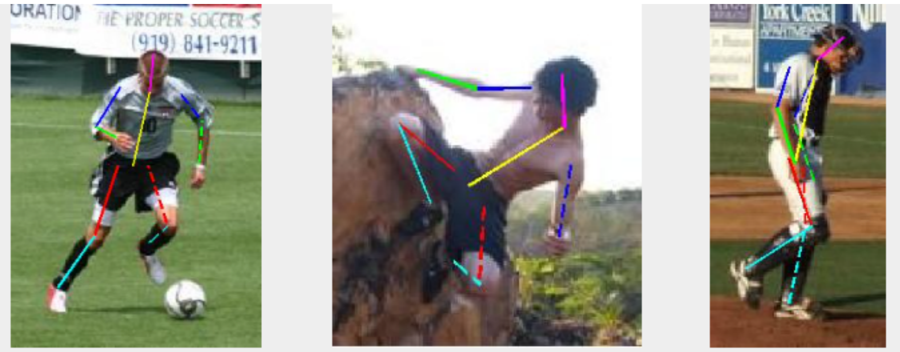
2000 2005 2010 2015 2020

# 2D POSE DATASETS OVER TIME

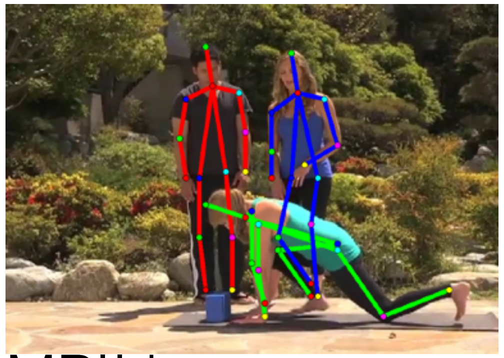
MS COCO:
40k image images
56k humans



LSPE dataset:
10k images




LSP dataset:
2k images

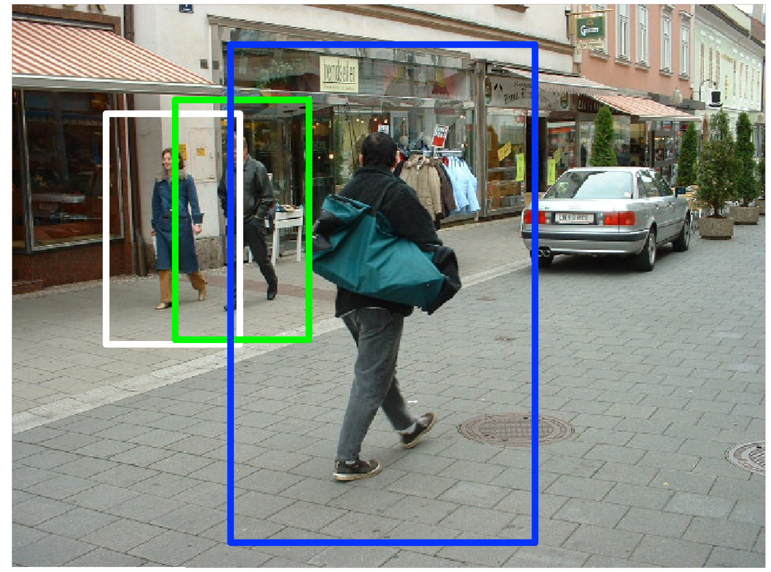


MPII human pose dataset:
25k images
40k humans

2000 2005 2010 2015 2020

# SEVERAL PATHS TO 3D HUMAN POSE

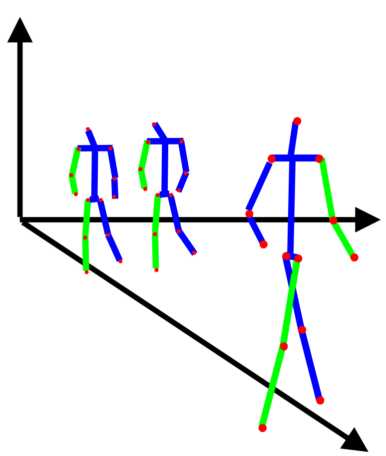
**Detection**



5

**2D pose**

4



**3D pose**

1 [Dalal & Triggs, CVPR'05]

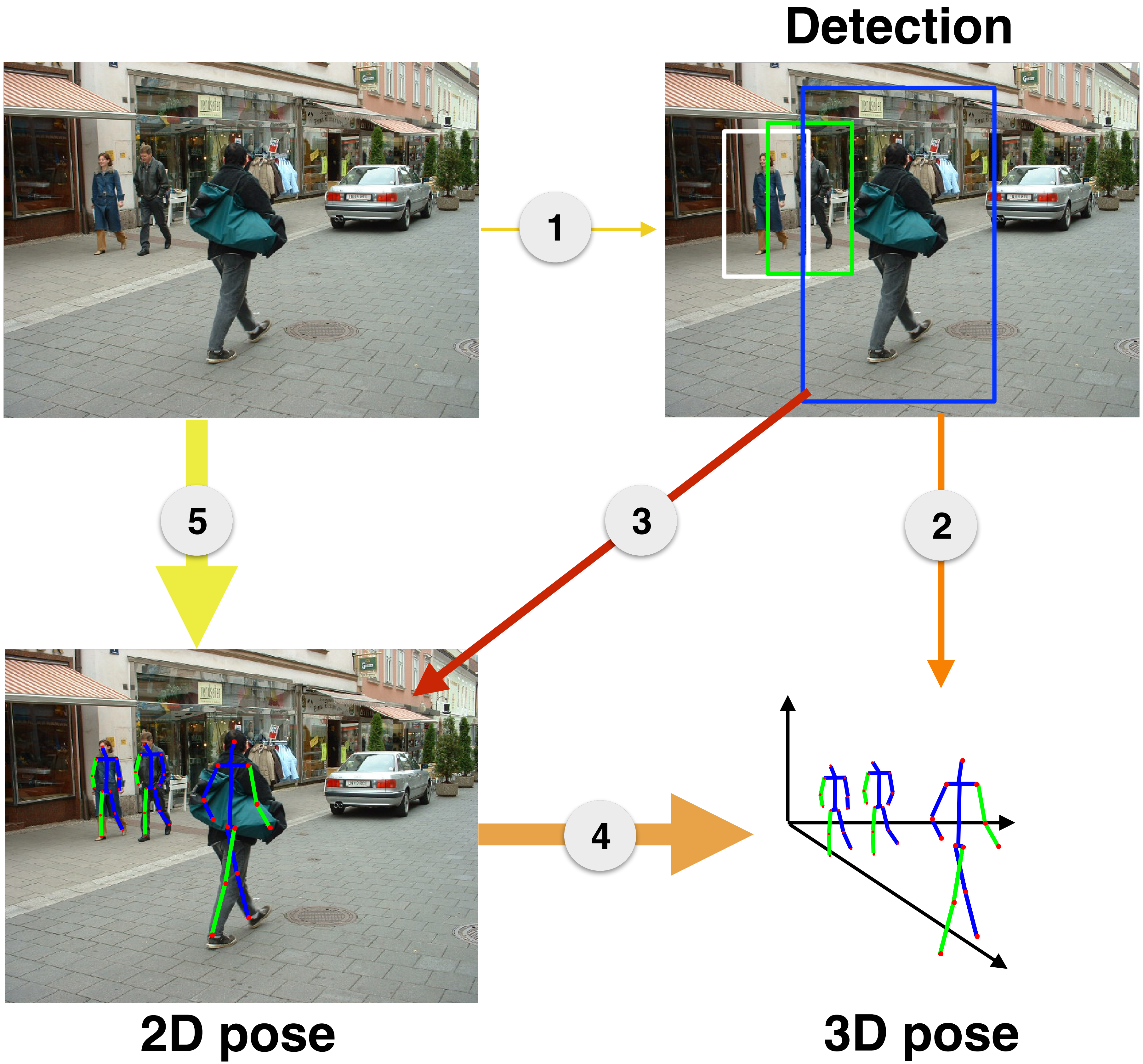
2 [Li et al, ICCV'15, Tekin et al, Zhou et al, CVPR'16]

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]

5 [Pishchulin et al, CVPR'16, Iqbal&Gall, ECCVw'16]

# SEVERAL PATHS TO 3D HUMAN POSE




1 [Dalal & Triggs, CVPR’05]

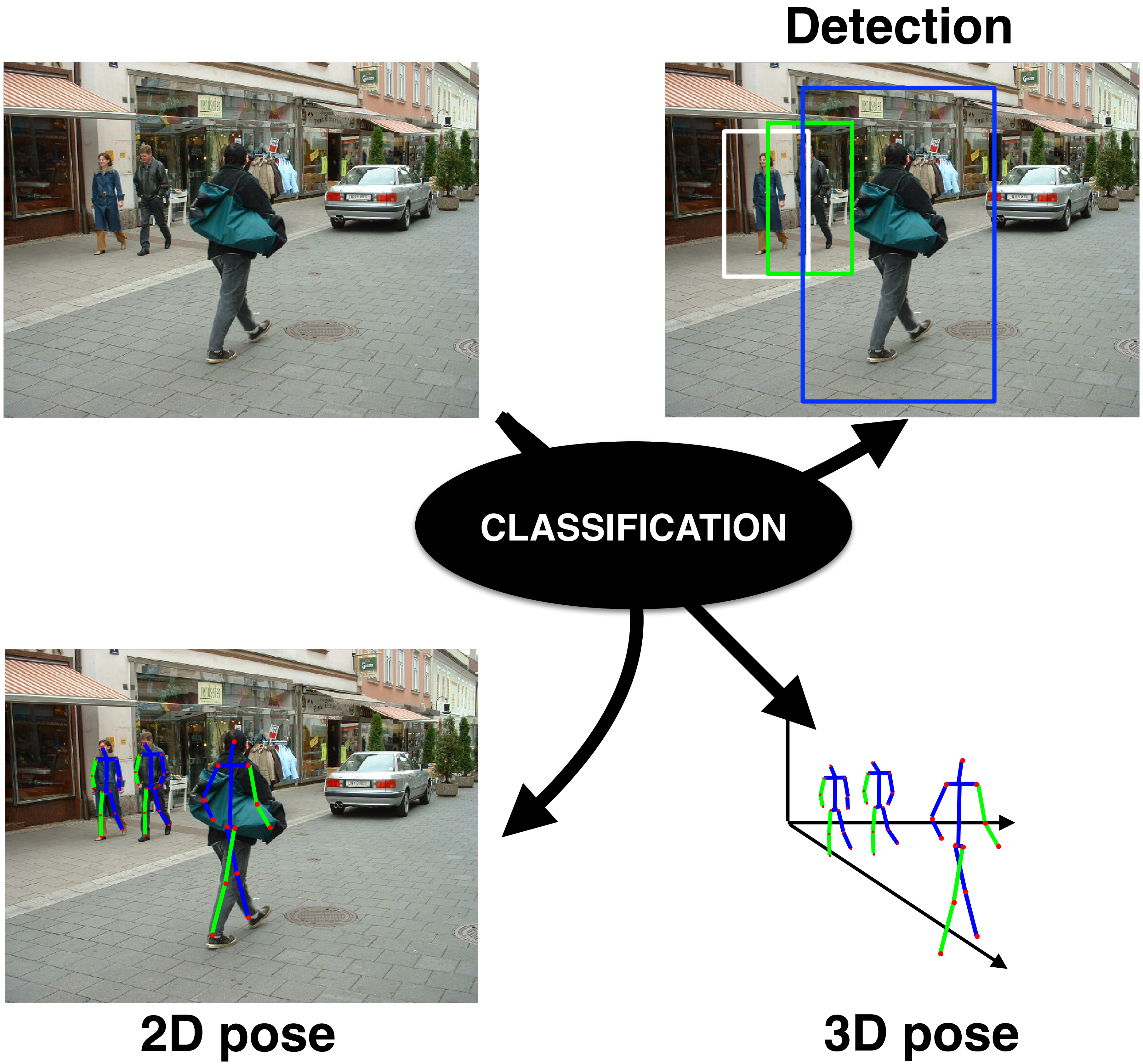
2 [Li et al, ICCV’15, Tekin et al, Zhou et al, CVPR’16]

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]

5 [Pishchulin et al, CVPR’16, Iqbal&Gall, ECCVw'16]

# SEVERAL PATHS TO 3D HUMAN POSE



# OUTLINE

- Background
- Monocular 3D Human pose estimation
- **Classification-based approaches**
- Drawbacks and solutions
- and beyond…

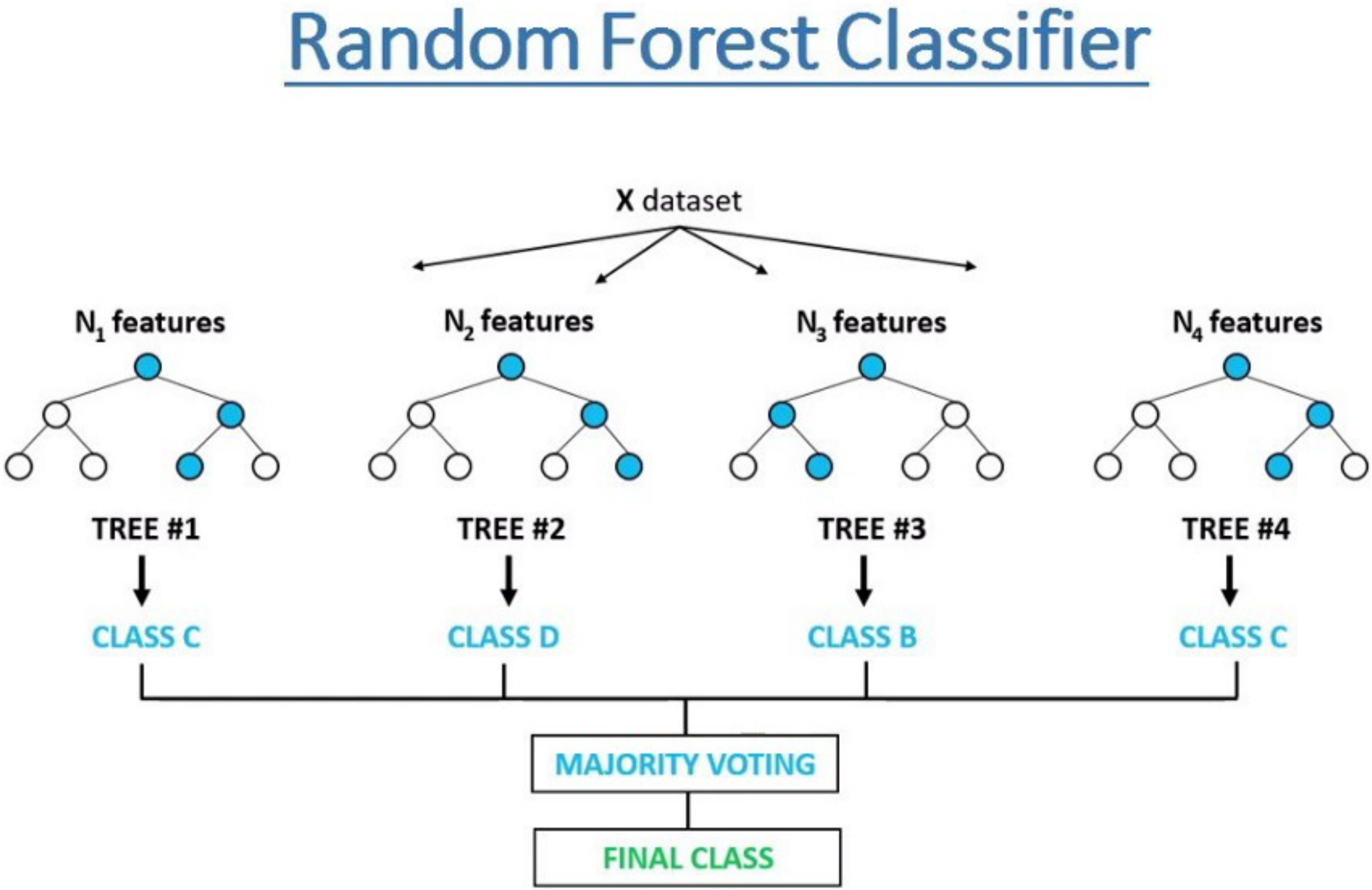
# 3D HUMAN POSE ESTIMATION AS A CLASSIFICATION PB

Back in 2007:

Greg, MSR just hired Jamie Shotton.
They will work on human pose estimation using Random Forest.
We should do it first!!

**Philip H.S. Torr**
Oxford University

I know Random Forest classifiers but
human pose space is continuous…



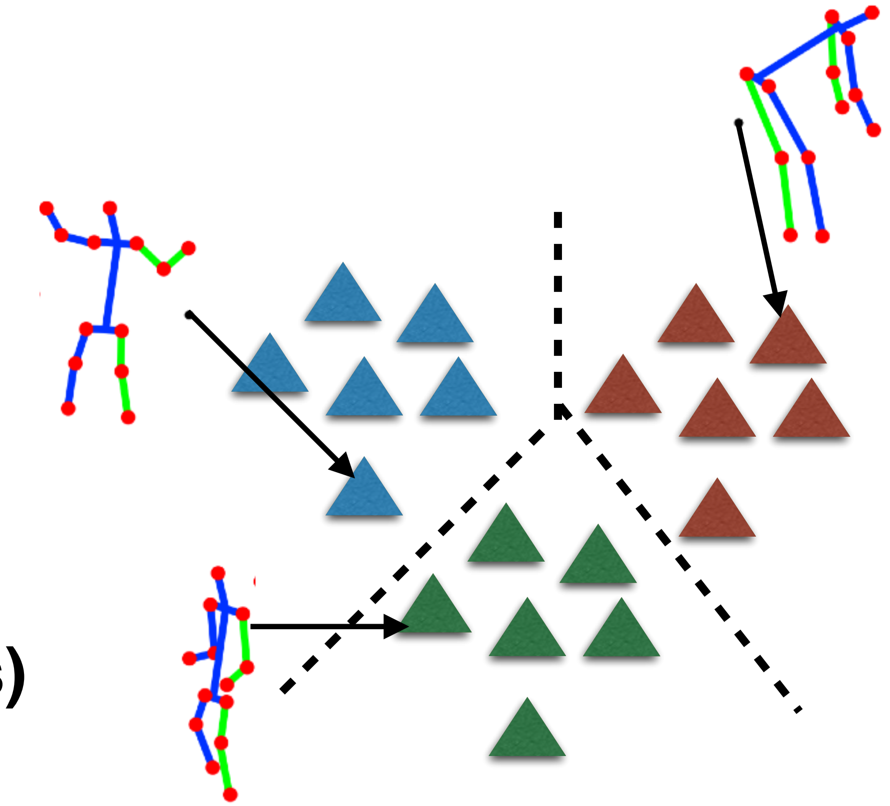
Random Forest Classifier
X dataset
N1 features
N2 features
N3 features
N4 features
TREE #1
TREE #2
TREE #3
TREE #4
CLASS C
CLASS D
CLASS B
CLASS C
MAJORITY VOTING
FINAL CLASS

# 3D HUMAN POSE ESTIMATION AS A CLASSIFICATION PB

1. Partition the space of body poses into K classes

2. Train a K-way classifier (here a RF).

3. Perform "pose detection":
 - Consider K+1 classes (additional background class)
 - Joint localization and pose estimation



4. Return center of top scoring classes or a weighted average

# 3D HUMAN POSE ESTIMATION AS A CLASSIFICATION PB

**Randomized Trees for Human Pose Detection**

Grégory Rogez[1], Jonathan Rihan[2], Srikumar Ramalingam[2], Carlos Orrite[1] and Philip H.S. Torr[2]

[1]Computer Vision Lab - I3A
University of Zaragoza, SPAIN
{grogez, corrite}@unizar.es
http://www.cv.i3a.unizar.es

[2]Department of Computing
Oxford Brookes University, UK
{jon.rihan,srikumar.ramalingam,philiptorr}@brookes.ac.uk
http://cms.brookes.ac.uk/research/visiongroup/

**Abstract**

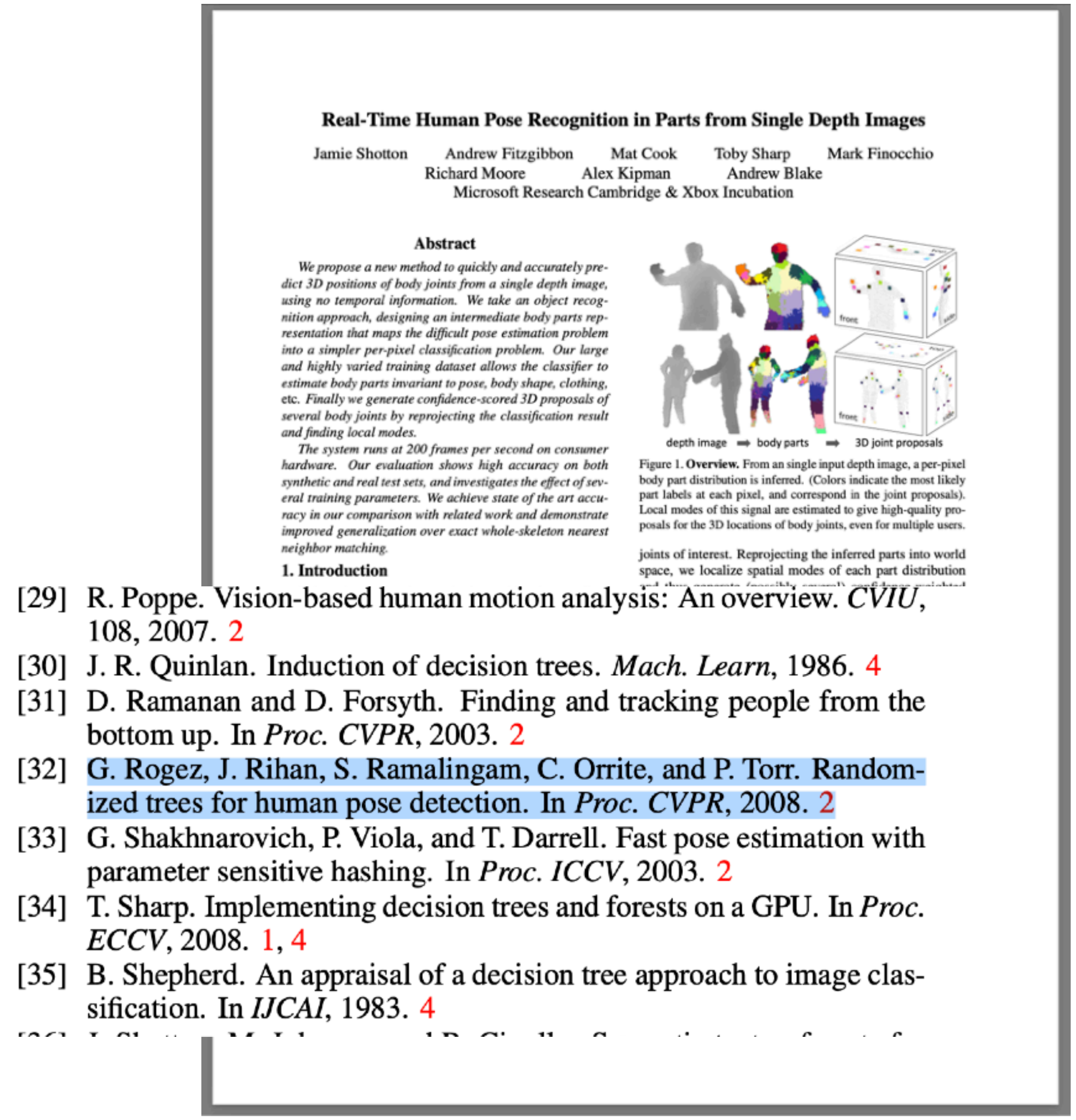
*This paper addresses human pose recognition from video sequences by formulating it as a classification problem. Unlike much previous work we do not make any assumptions on the availability of clean segmentation. The first step of this work consists in a novel method of aligning the training images using 3D Mocap data. Next we define classes by discretizing a 2D manifold whose two dimensions are camera viewpoint and actions. Our main contribution is a pose detection algorithm based on random forests. A bottom-up approach is followed to build a decision tree by recursively clustering and merging the classes at each level. For each node of the decision tree we build a list of potentially discriminative features using the alignment of training images;*

In this work, we propose an efficient method to jointly localize and recognize the pose of humans, using an exemplar based approach and fast search technique. Such pose detector would be very useful for initializing model-based-approaches[17], tracking algorithms [24] or segmentation algorithms [7].

**1.1. Related Previous Work**

Exemplar based approaches have been very successful in pose recognition [16]. However, in a scenario involving a wide range of viewpoints and poses, a large number of exemplars would be required. As a result the computational time would be very high to recognize individual poses. One approach, based on efficient nearest neighbours search using histogram of gradient features, addressed the problem of

[3] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, and A. Ng. Discriminative learning of markov random fields for segmentation of 3D scan data. In *Proc. CVPR*, 2005. 2
[4] Autodesk MotionBuilder. 3
[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. PAMI*, 24, 2002. 4
[6] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *Proc. ICCV*, 2009. 2
[7] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. CVPR*, 1998. 1, 2
[8] L. Breiman. Random forests. *Mach. Learning*, 45(1):5–32, 2001. 4
[9] CMU Mocap Database. http://mocap.cs.cmu.edu/. 3



**Real-Time Human Pose Recognition in Parts from Single Depth Images**

Jamie Shotton Andrew Fitzgibbon Mat Cook Toby Sharp Mark Finocchio
Richard Moore Alex Kipman Andrew Blake
Microsoft Research Cambridge & Xbox Incubation

**Abstract**

*We propose a new method to quickly and accurately predict 3D positions of body joints from a single depth image, using no temporal information. We take an object recognition approach, designing an intermediate body parts representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. Our large and highly varied training dataset allows the classifier to estimate body parts invariant to pose, body shape, clothing, etc. Finally we generate confidence-scored 3D proposals of several body joints by reprojecting the classification result and finding local modes.*

*The system runs at 200 frames per second on consumer hardware. Our evaluation shows high accuracy on both synthetic and real test sets, and investigates the effect of several training parameters. We achieve state of the art accuracy in our comparison with related work and demonstrate improved generalization over exact whole-skeleton nearest neighbor matching.*

**1. Introduction**

depth image ➡ body parts ➡ 3D joint proposals

Figure 1. **Overview.** From an single input depth image, a per-pixel body part distribution is inferred. (Colors indicate the most likely part labels at each pixel, and correspond in the joint proposals). Local modes of this signal are estimated to give high-quality proposals for the 3D locations of body joints, even for multiple users.

joints of interest. Reprojecting the inferred parts into world space, we localize spatial modes of each part distribution

[29] R. Poppe. Vision-based human motion analysis: An overview. *CVIU*, 108, 2007. 2
[30] J. R. Quinlan. Induction of decision trees. *Mach. Learn*, 1986. 4
[31] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *Proc. CVPR*, 2003. 2
[32] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. Torr. Randomized trees for human pose detection. In *Proc. CVPR*, 2008. 2
[33] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *Proc. ICCV*, 2003. 2
[34] T. Sharp. Implementing decision trees and forests on a GPU. In *Proc. ECCV*, 2008. 1, 4
[35] B. Shepherd. An appraisal of a decision tree approach to image classification. In *IJCAI*, 1983. 4

**CVPR 2008** **~ 250 citations**

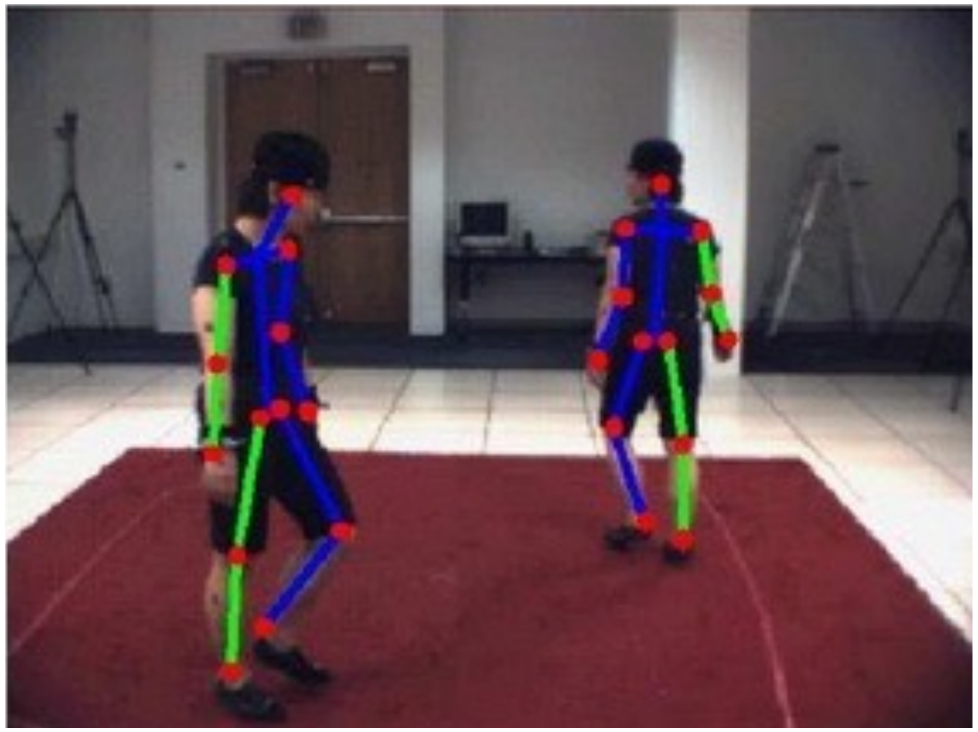
**CVPR 2011** **4000+ citations**

What is MSR doing?

# POSE ESTIMATION BY CLASSIFICATION: 3 CASES

## Case 1: Full-body walking poses

[Rogez, Rihan, Ramalingam, Orrite and Torr, CVPR'08]

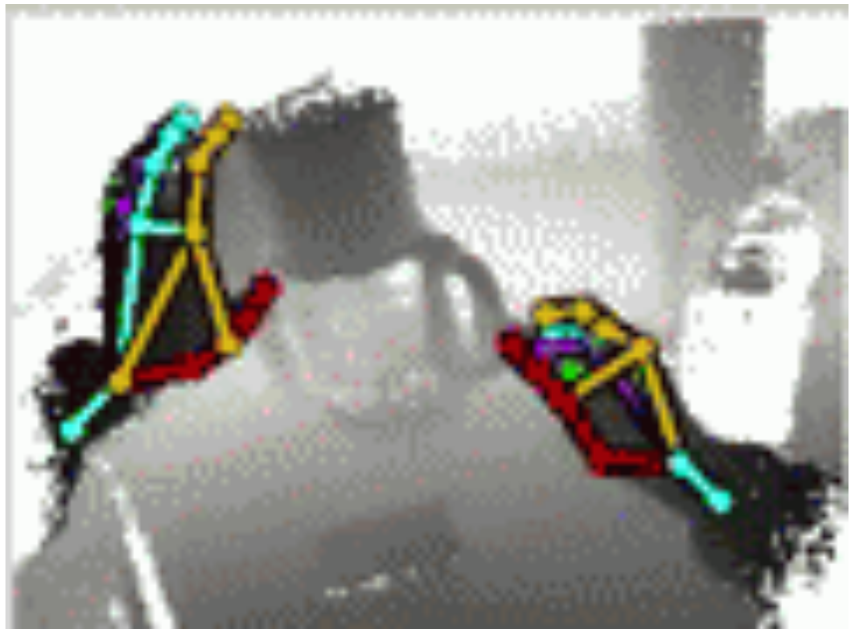


## Case 1b: Full-body walking poses

[Rogez, Rihan, Orrite and Torr, IJCV'12]

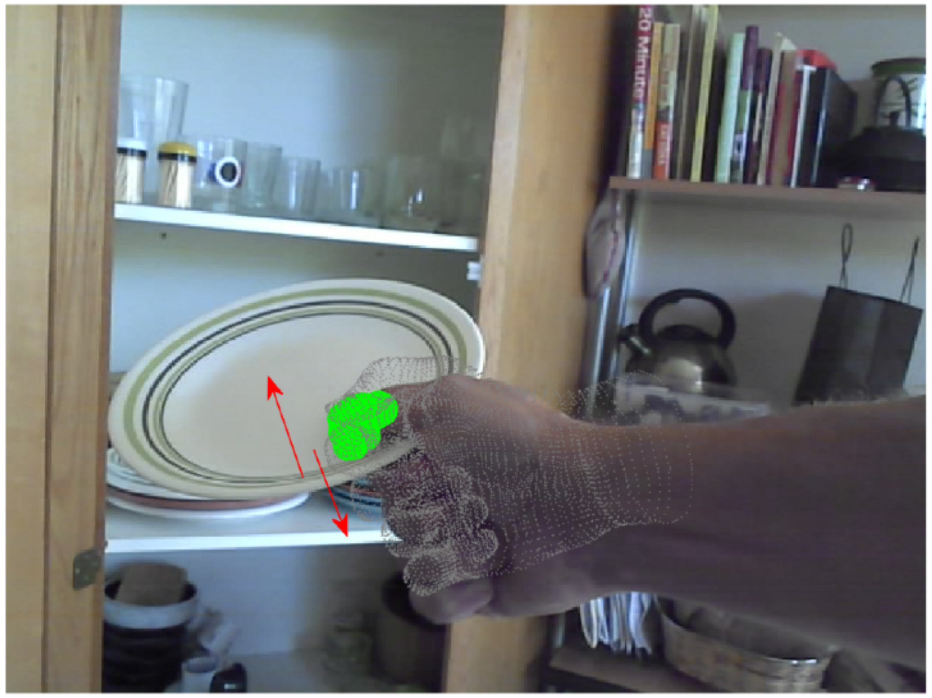
## Case 2: Upper-limb egocentric view

[Rogez, Supancic and Ramanan, CVPR'15]



## Case 3: Grasping hand

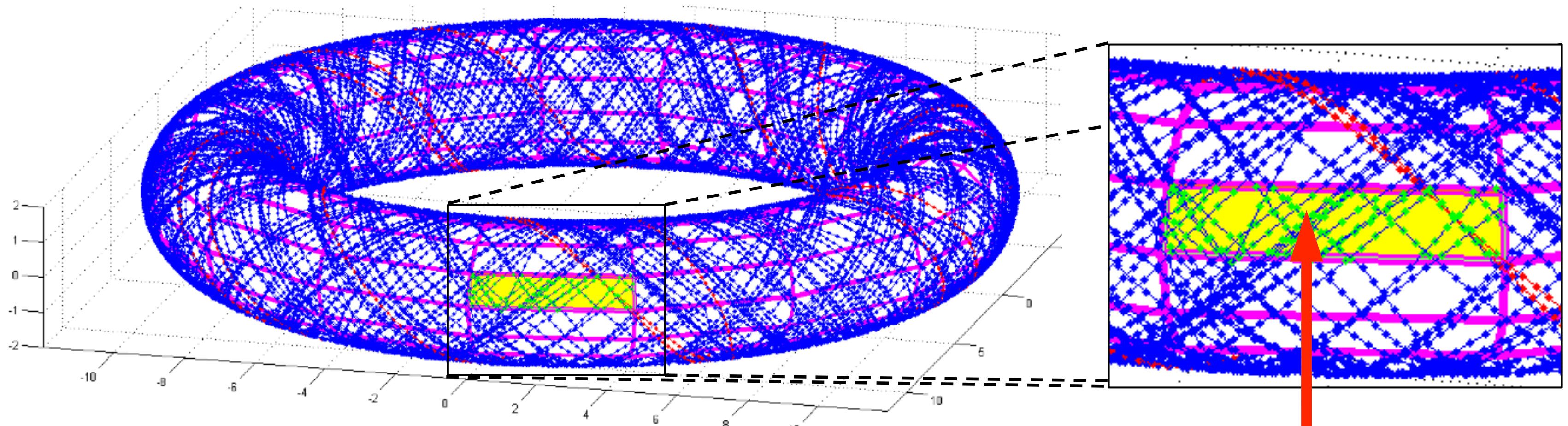
[Rogez, Supancic and Ramanan, ICCV'15]

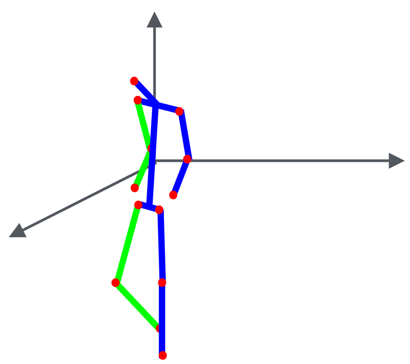


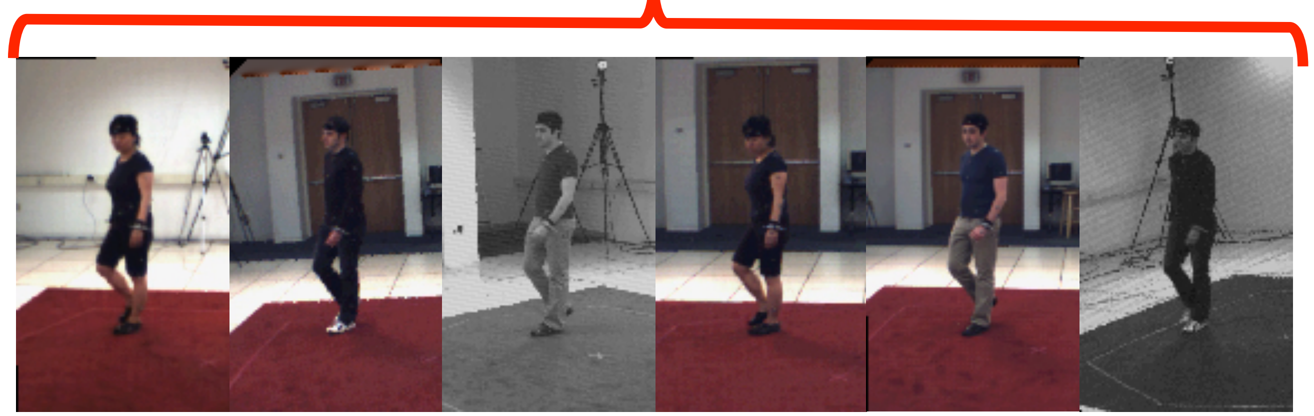
# CASE 1: FULL-BODY WALKING POSES

**Class definition:** Torus manifold to model viewpoint and pose of cyclic motion + grid

[Rogez et al *Pattern Recognition* 2008]

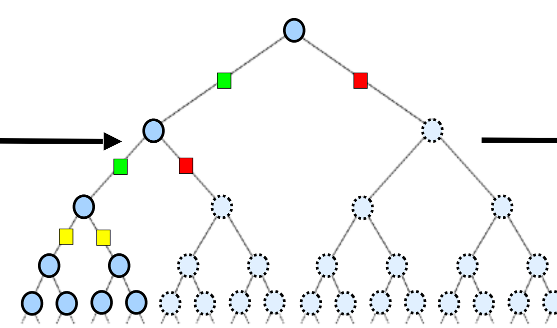


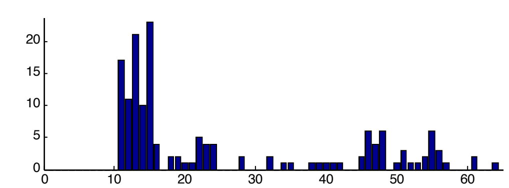


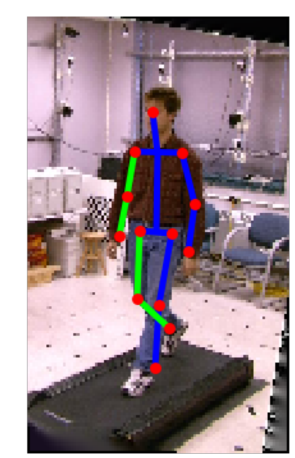


**2D pose** **3D pose**

**Holistic full-body** pose estimation

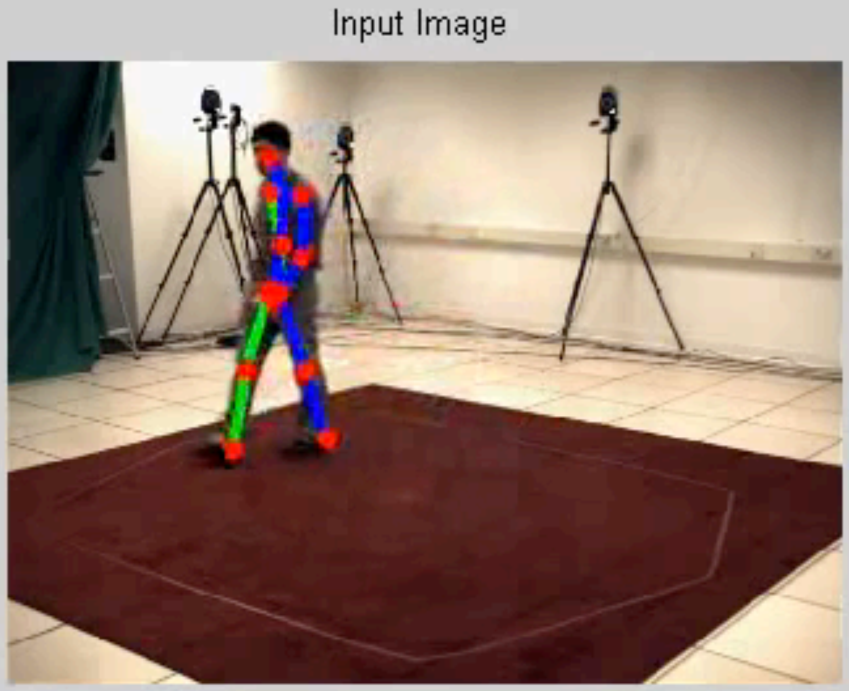


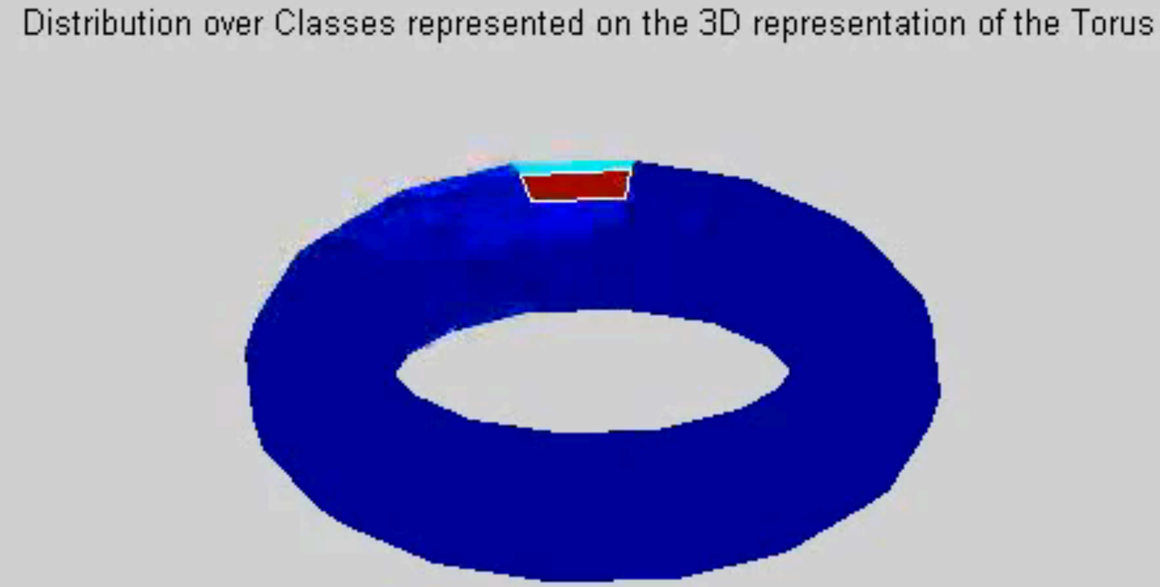




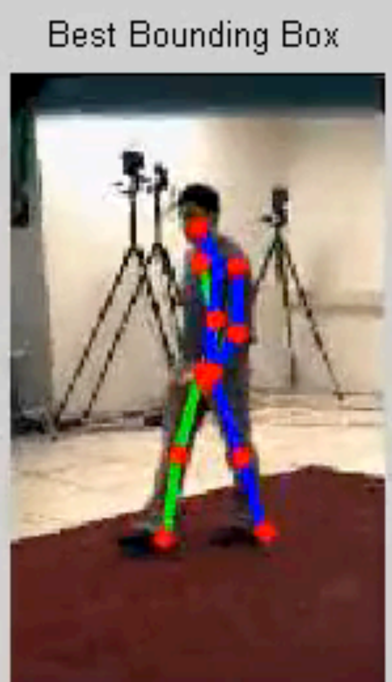
[Rogez, Rihan, Ramalingam, Orrite & Torr, Randomized trees for human pose detection, CVPR'08]

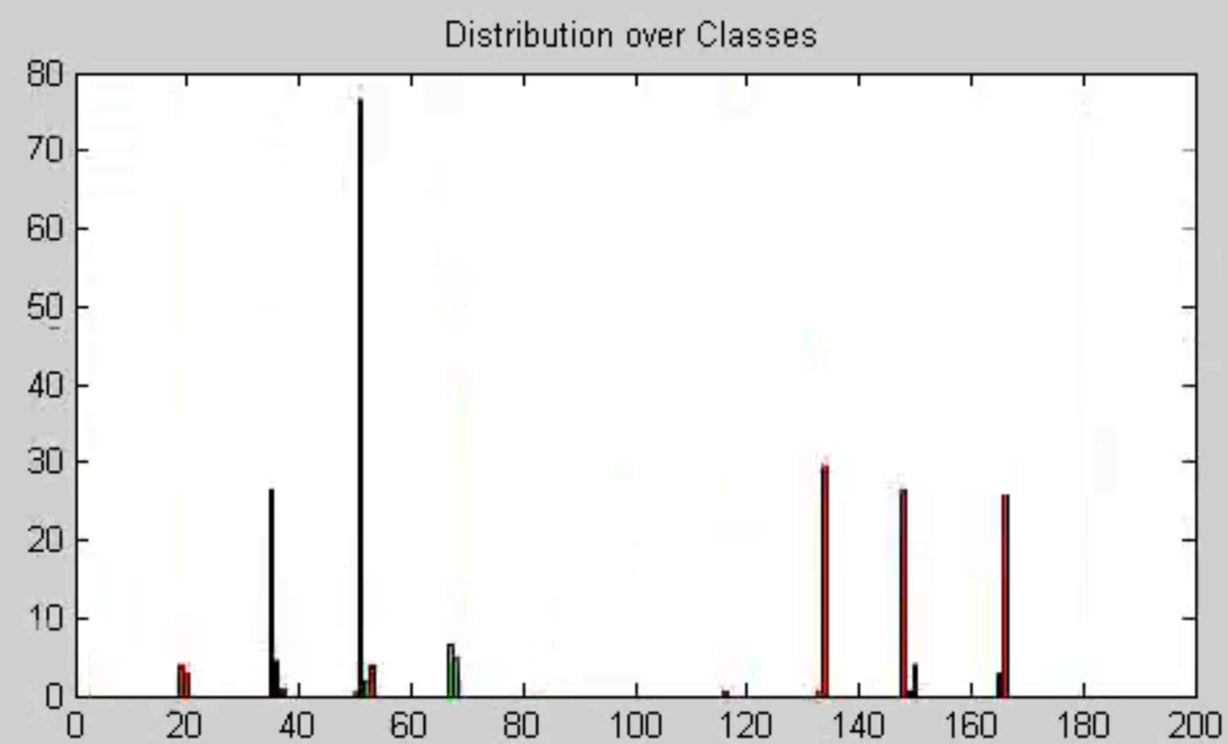
# CASE 1: FULL-BODY WALKING POSES




**Good for tracking + model multi-modal distributions**



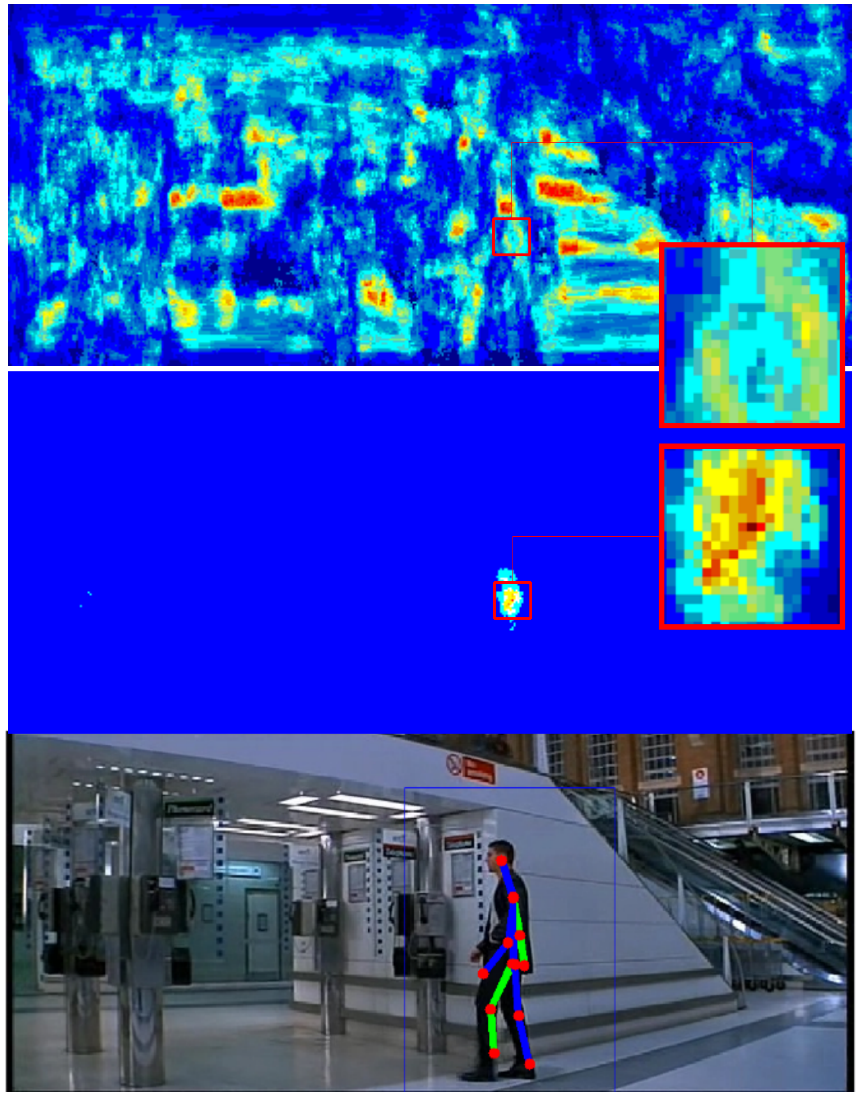

[Rogez, Rihan, Ramalingam, Orrite & Torr, Randomized trees for human pose detection, CVPR'08]

# FULL-BODY WALKING POSES IN THE WILD

To make the pose detector work in-the-wild:

Data augmentation

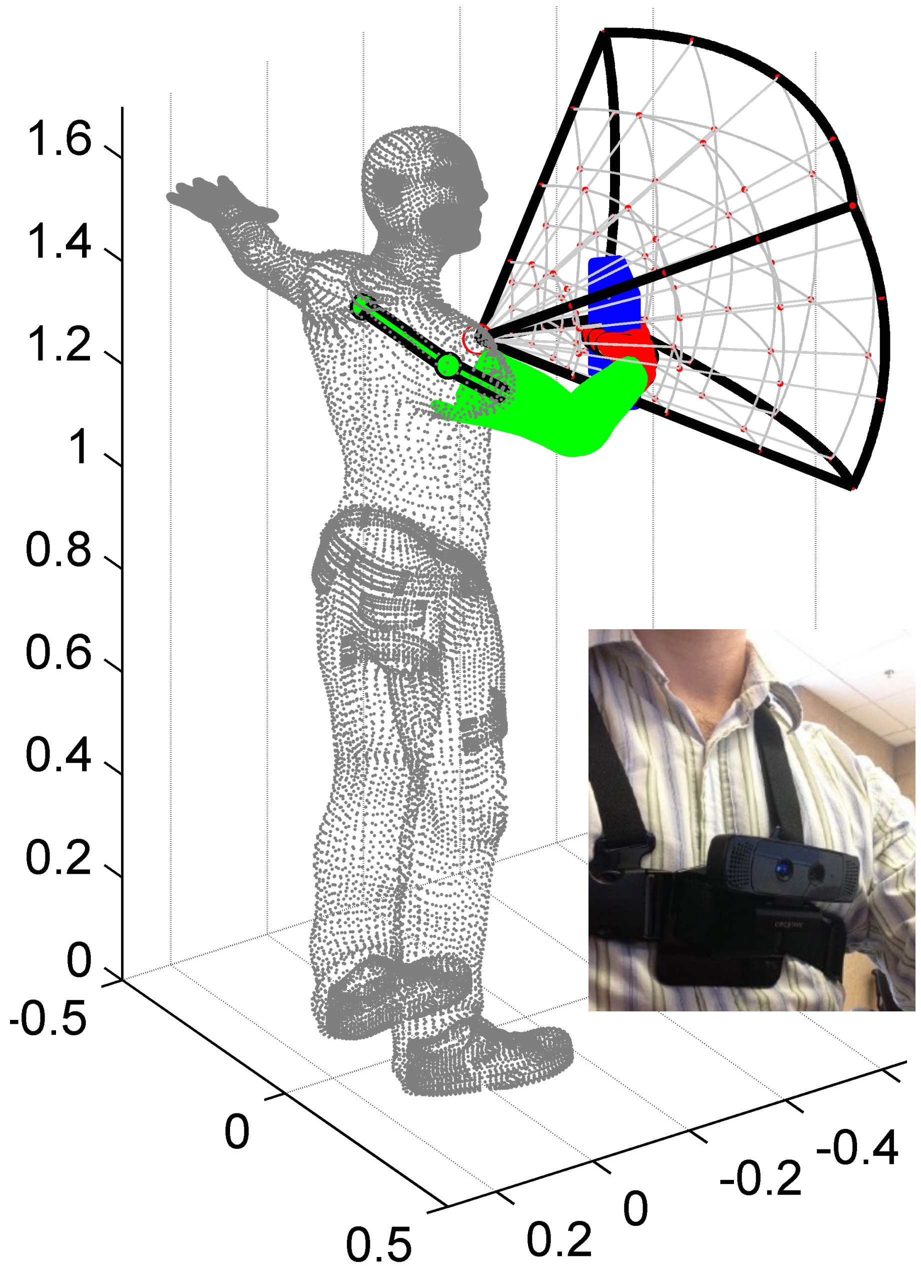
Hard negative mining

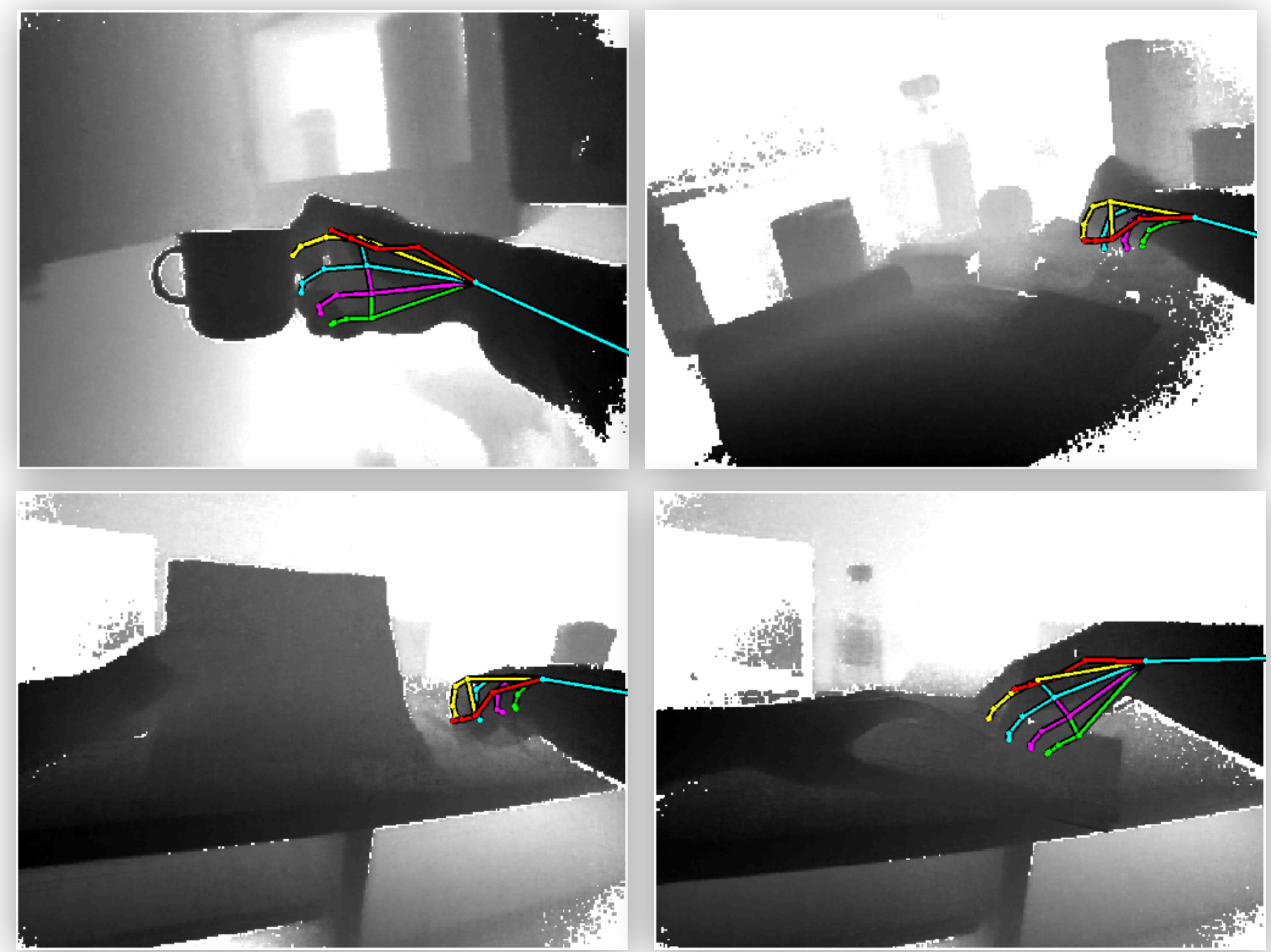


Real-time detection + pose estimation

[Rogez et al., Fast Human Pose Detection Using Randomized Hierarchical Cascades of Rejectors, IJCV'12

# CASE 2: UPPER-LIMB EGOCENTRIC VIEW



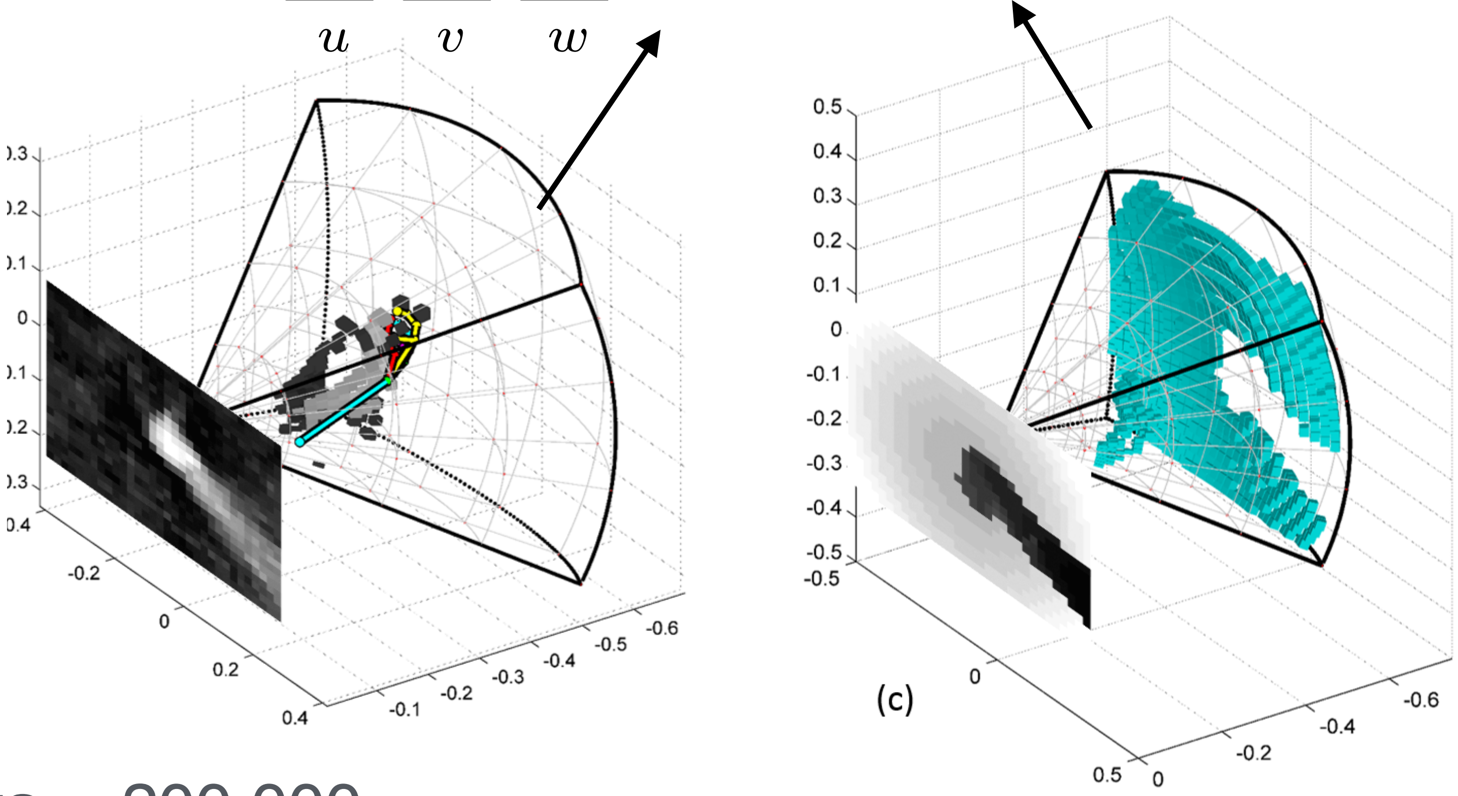
**Class Definition:** K-means on Arm+Hand 3D pose inside "egocentric workspace"

The class directly returns 2D-3D hand pose **+ location & scale** in the image

[Rogez, Supancic & Ramanan, First-person pose recognition using egocentric workspaces. CVPR'15]

# CASE 2: UPPER-LIMB EGOCENTRIC VIEW

$$\text{score}[k] = \sum_u \sum_v \sum_w \beta_k[u,v,w] \cdot b[u,v,w].$$

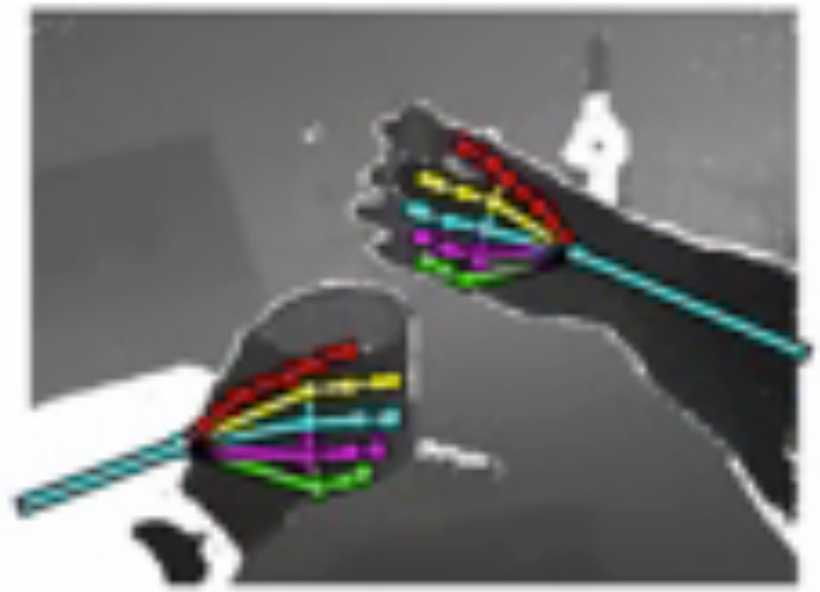


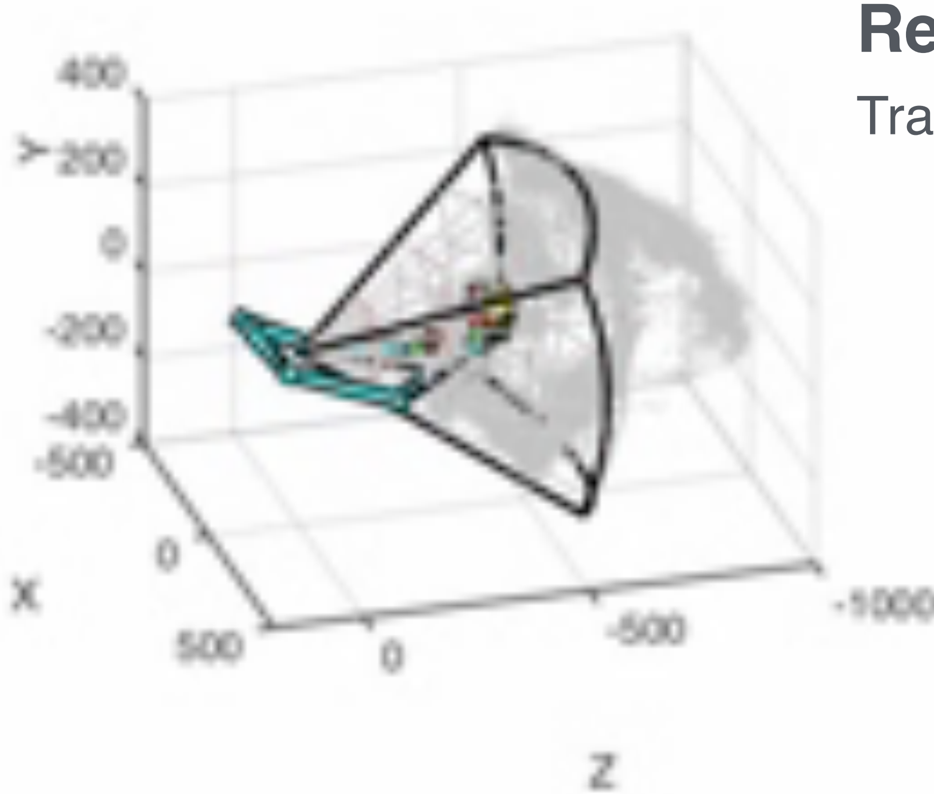

**Classifier:** SVM on binary depth features

**Data:** ~200,000 synthetic "egocentric workspaces"


[Rogez, Supancic & Ramanan, First-person pose recognition using egocentric workspaces. CVPR'15]

# CASE 2: UPPER-LIMB EGOCENTRIC VIEW

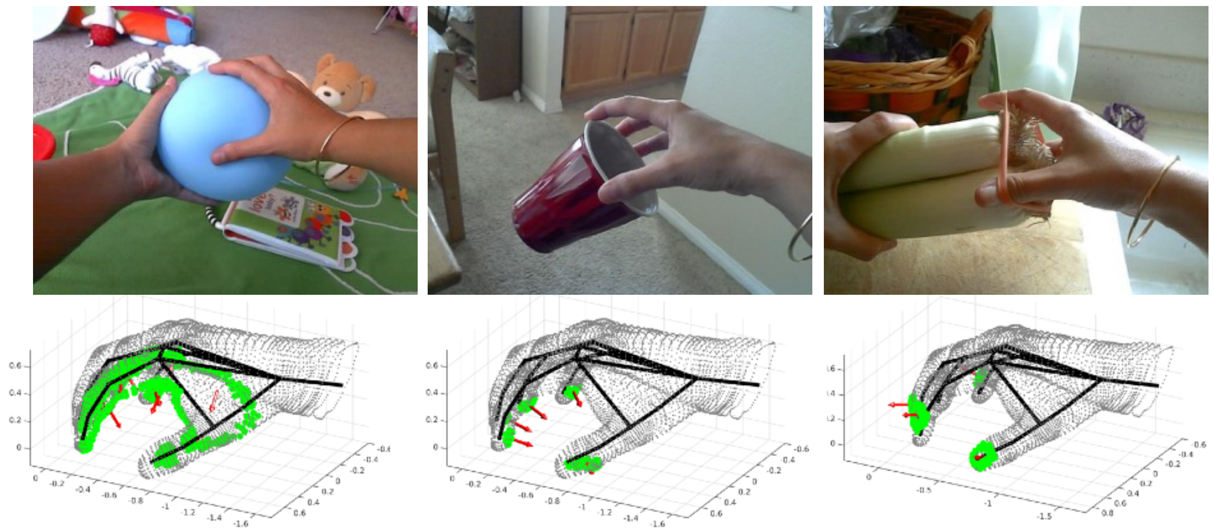


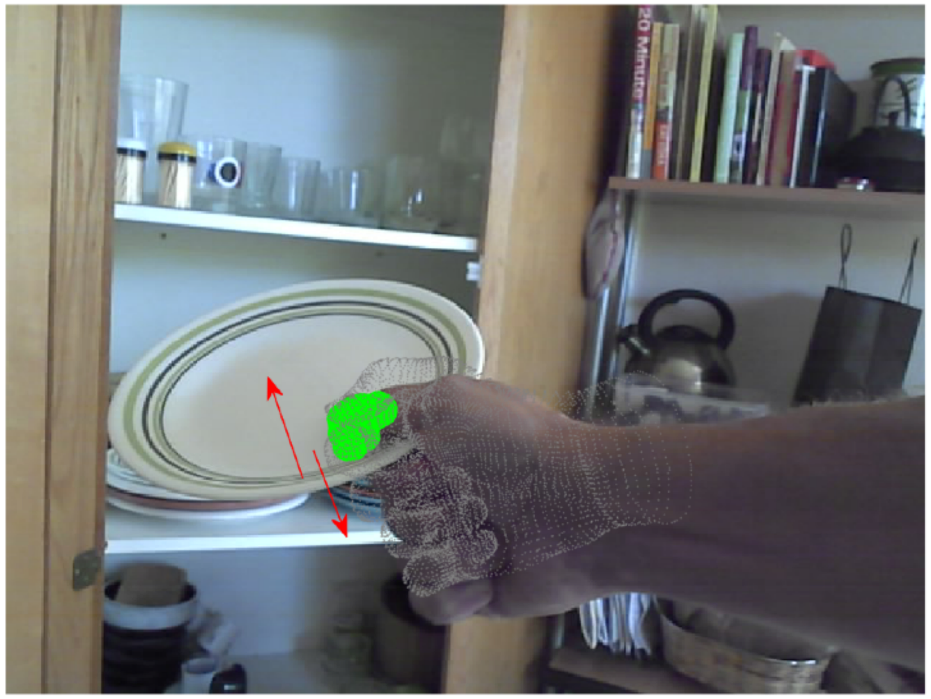



**Real-time demo** running on Matlab with K=750

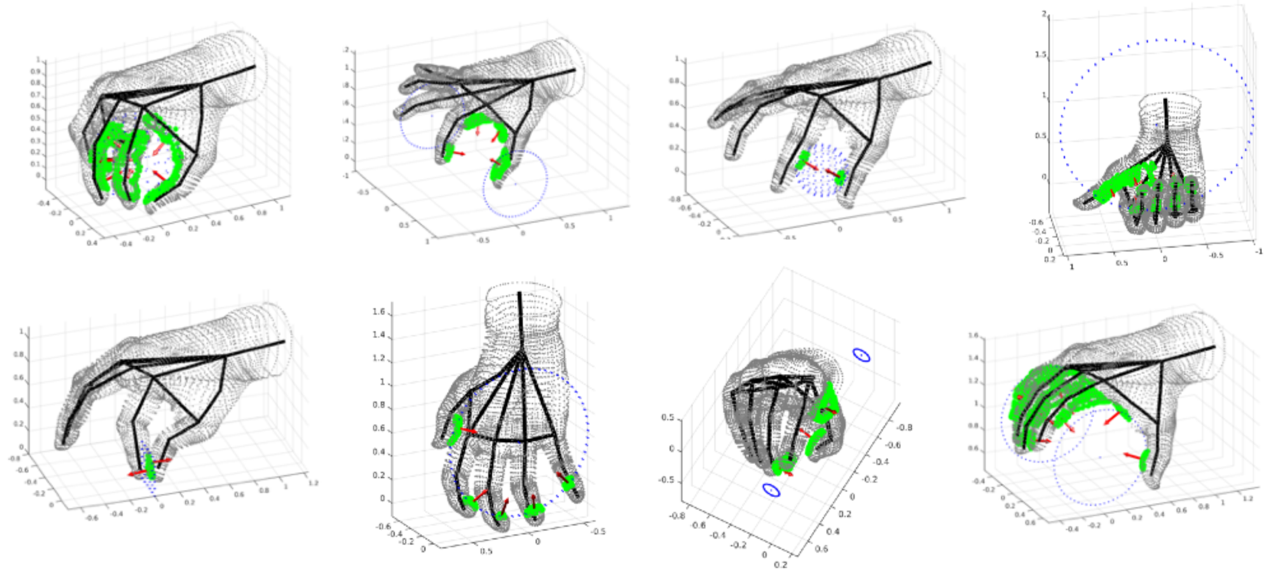
Trade-off precision / computation

# CASE 3: MORE THAN JUST POSE





**A same kinematic pose** can be used for dramatically **different functional manipulations**

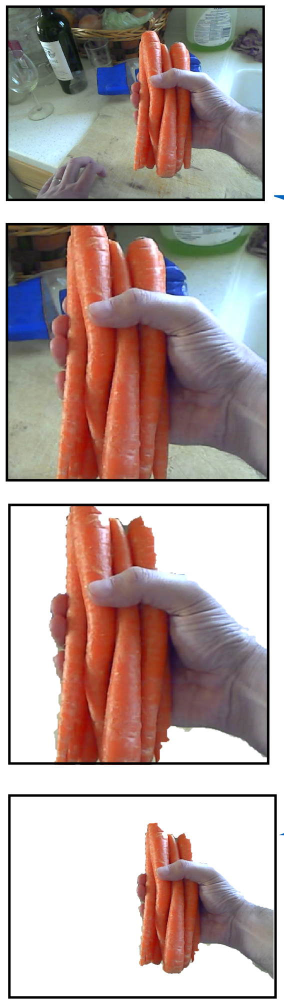


**Class Definition:** fine-grained grasps (pose+contact+forces)

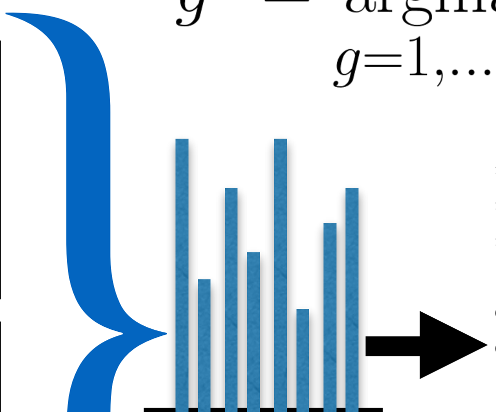
71-class taxonomy [Liu et al, Humanoid'14]

[Rogez, Supancic & Ramanan, Understanding Hands in Action. ICCV'15]

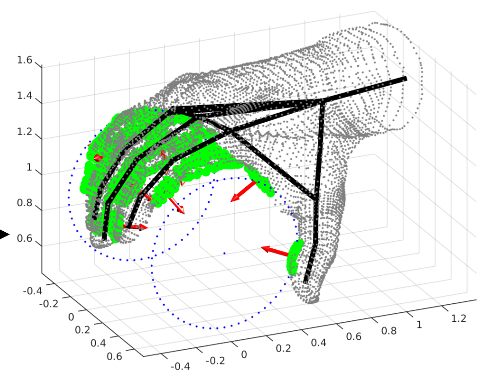
# CASE 3: MORE THAN JUST POSE



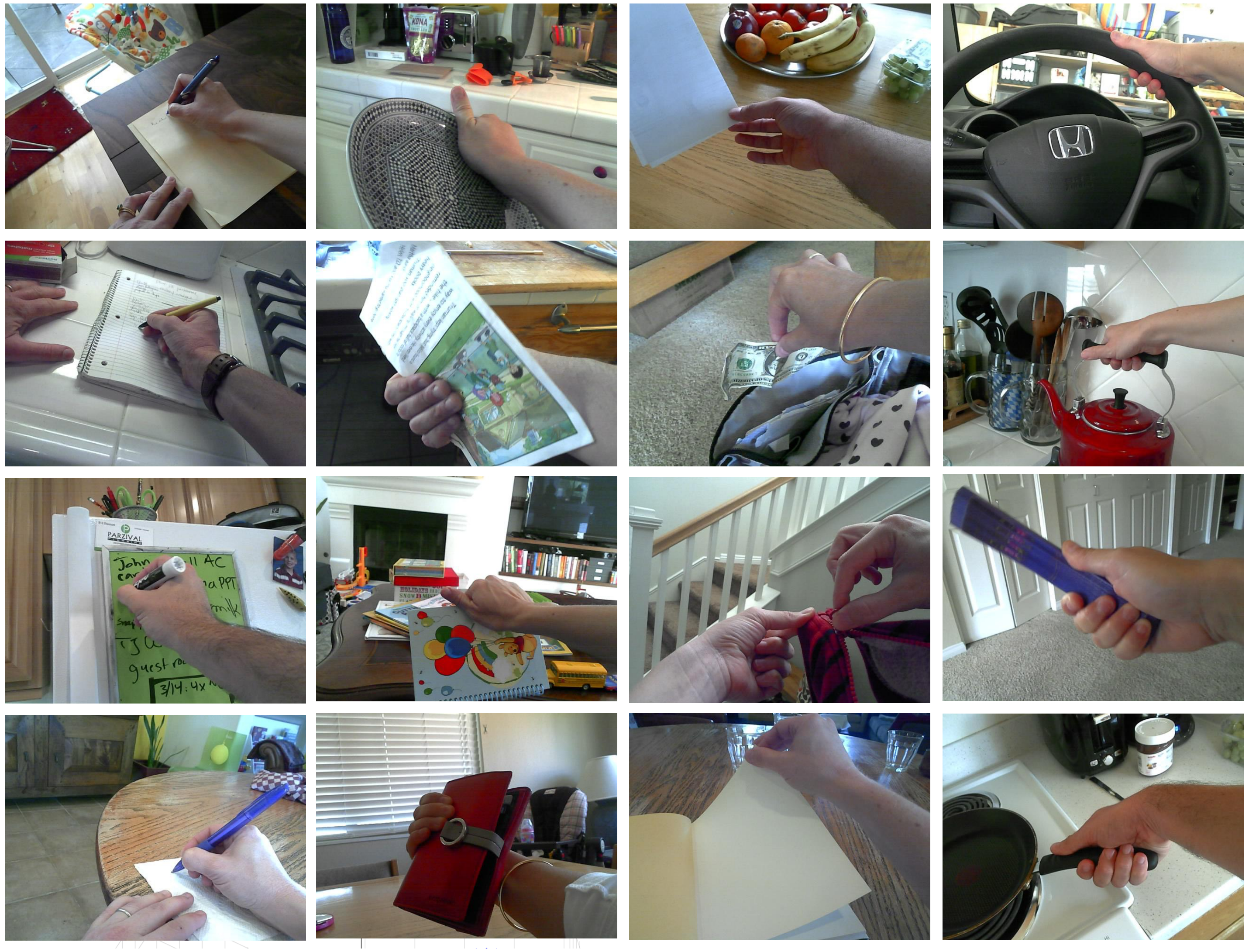
$$g^* = \operatorname*{argmax}_{g=1,\ldots,71} \left(w_g^T \phi(I) + b_g\right)$$



**CNN features**



**grasp estimation**

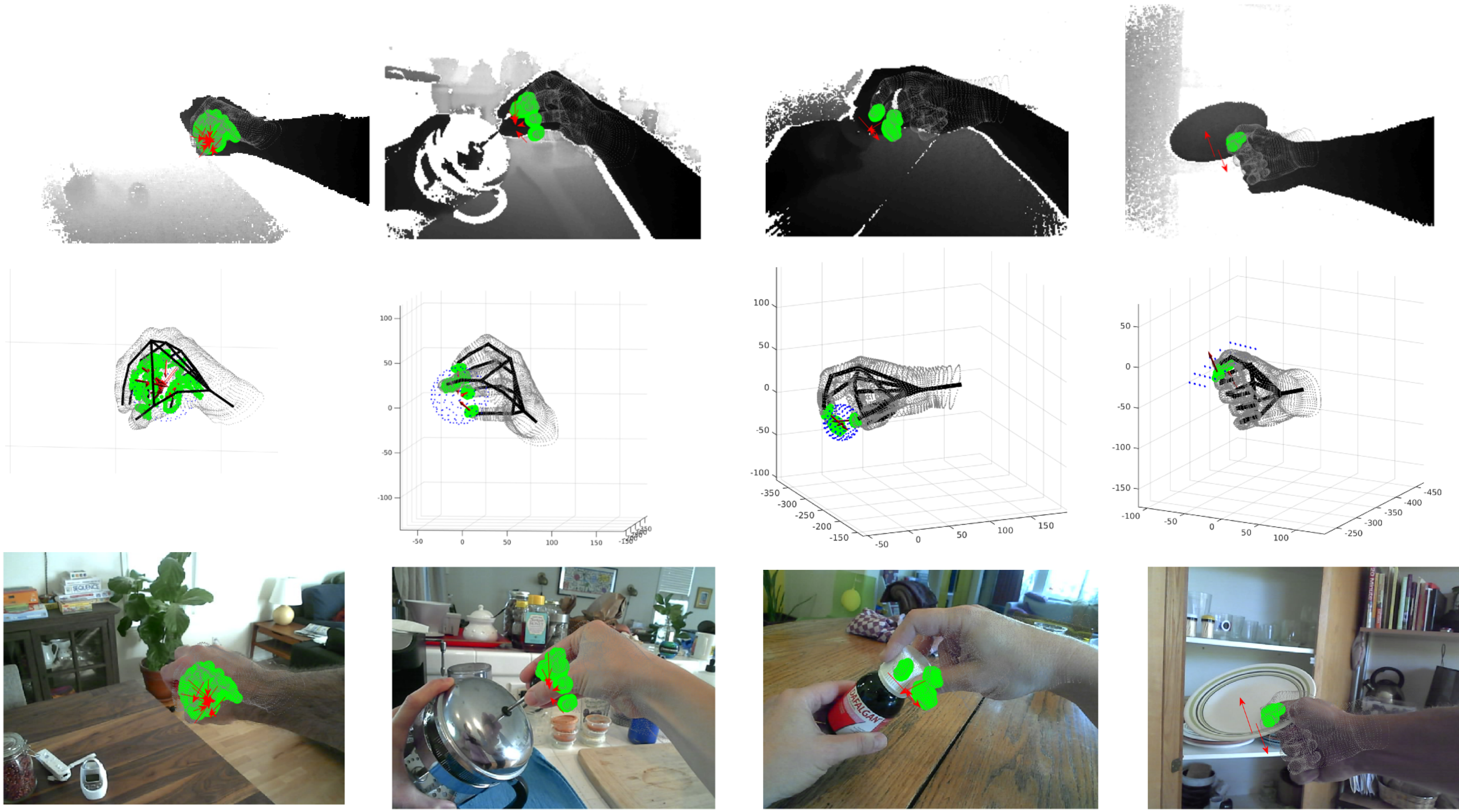


**Classifier:** SVM on deep features

**Data:** ~12k RGBD images (25 obj./grasp, 8 subj)

[Rogez, Supancic & Ramanan, Understanding Hands in Action. ICCV'15]

# CASE 3: GRASPING HAND



[Rogez, Supancic & Ramanan, Understanding Hands in Action. ICCV’15]

# CLASSIFICATION: LESSONS LEARNT

**Pose detection:**
localization + 3D/2D pose

Model **multi-modal**
distributions

Holistic **full-body**
approach

Additional
**attributes**

# OUTLINE

- Background
- Monocular 3D Human pose estimation
- Classification-based approaches
- **Drawbacks and solutions**
- and beyond…

# CLASSIFICATION: LESSONS LEARNT

Requires **large scale training data (images+3D pose)**

**SYNTHESIS**

Won't work with **unseen poses**

Only **coarse pose** estimation

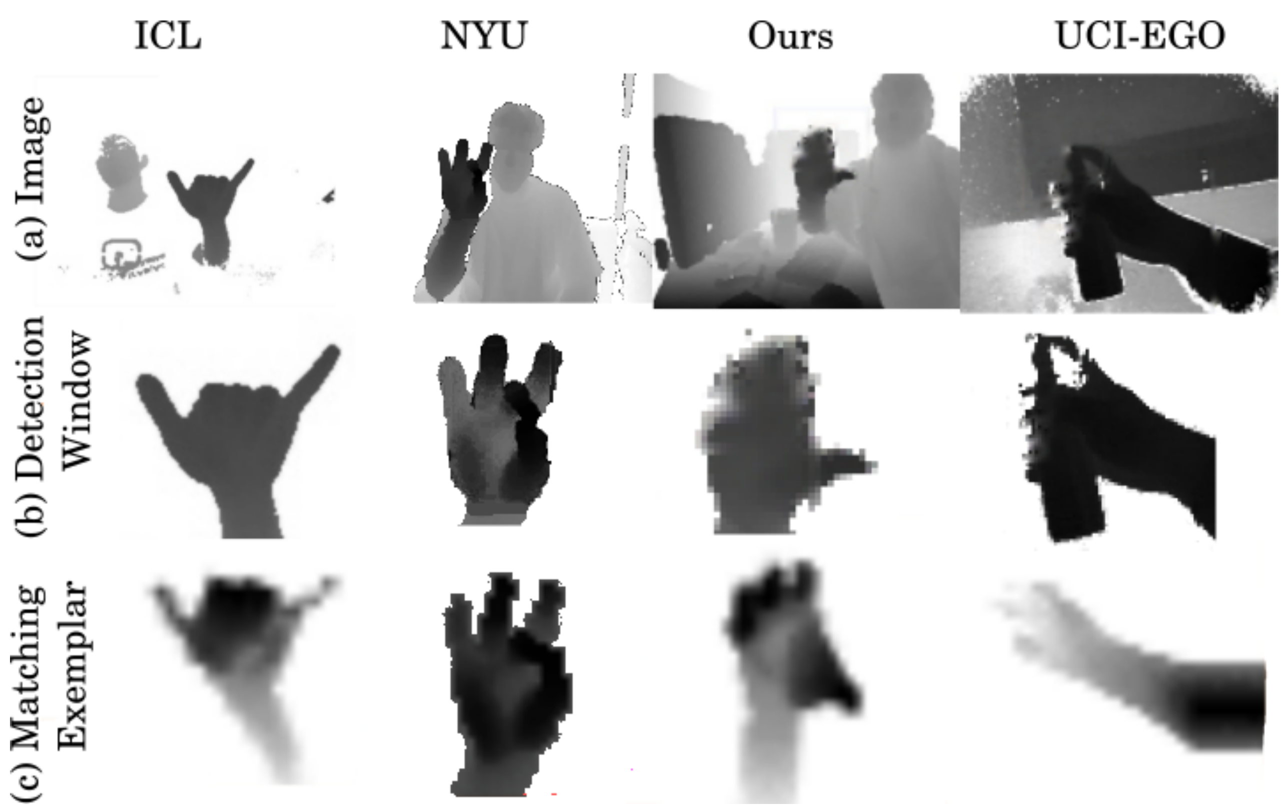
**CNN**

Computational **cost**

# REALISTIC SYNTHETIC HUMANS FOR DEPTH

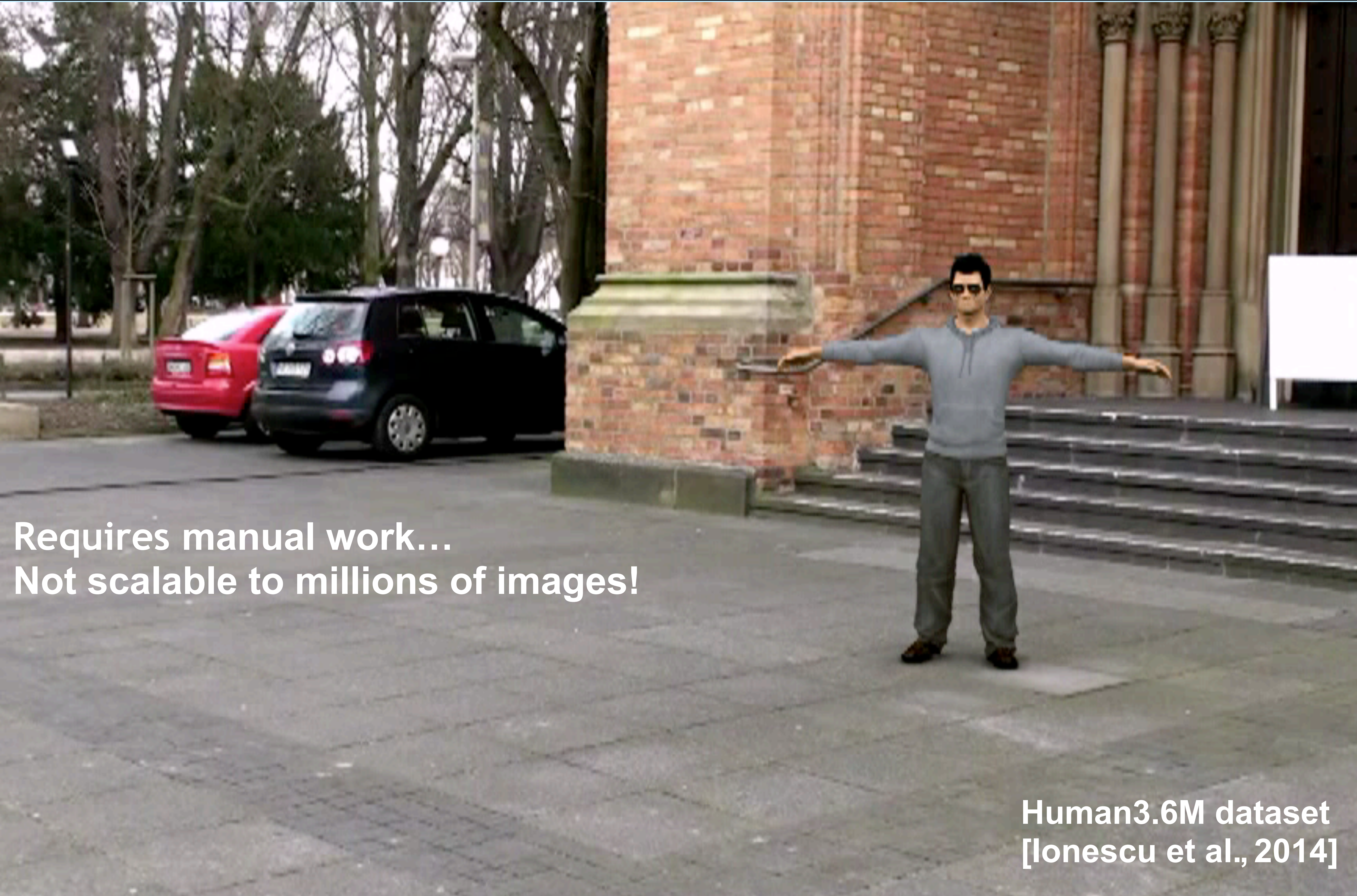
**Data:** ~200,000 synthetic “egocentric workspaces”

[Rogez, Supancic & Ramanan, First-person pose recognition using egocentric workspaces. CVPR’15]




[Supancic, Rogez, Yang, Shotton & Ramanan, Depth-Based Hand Pose Estimation: Data, Methods, and Challenges. ICCV’15 and IJCV’18]

# WHAT ABOUT RGB?



**Requires manual work…**
**Not scalable to millions of images!**

**Human3.6M dataset**
**[Ionescu et al., 2014]**

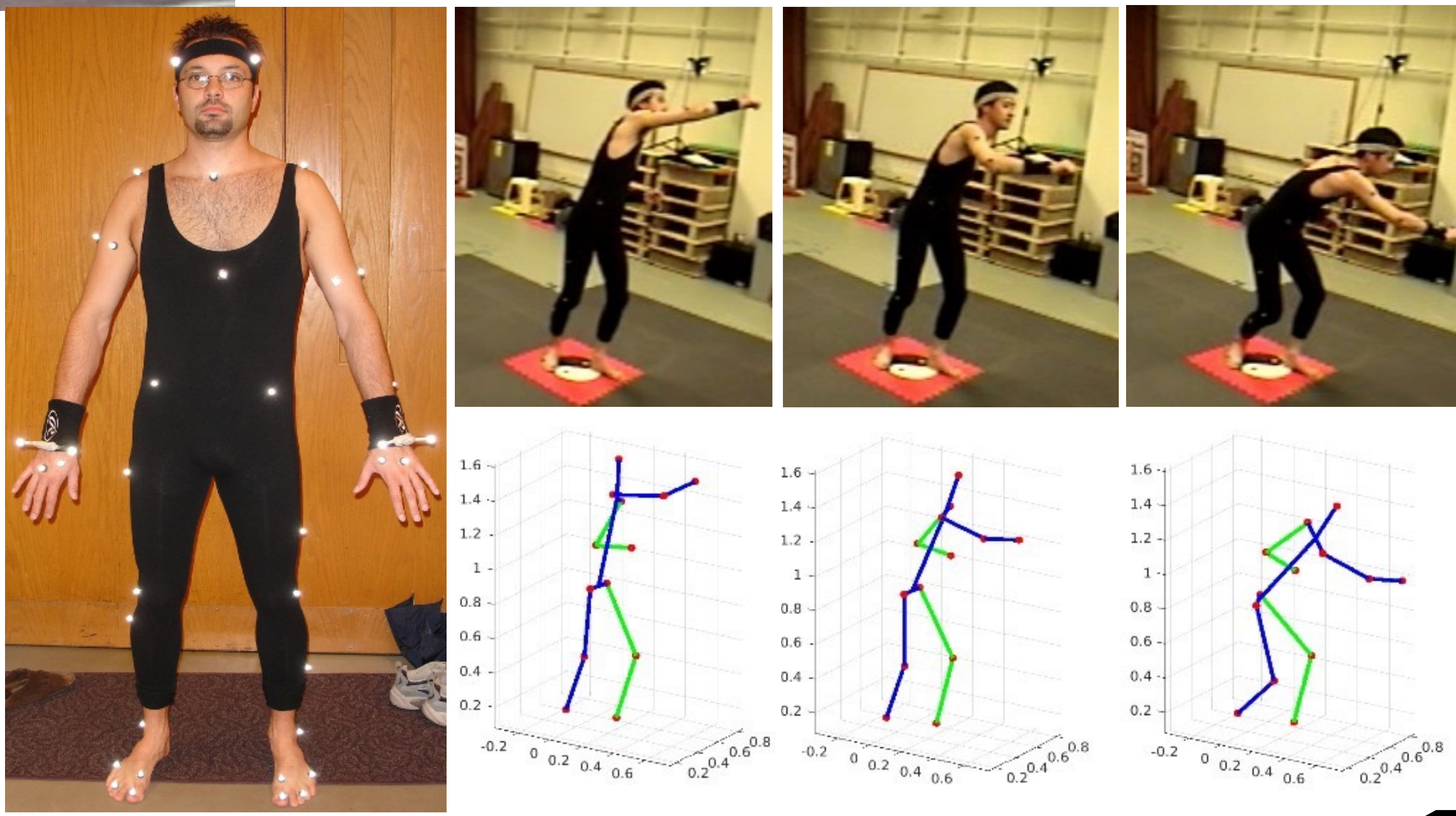
# WHAT DATA DO WE HAVE?

**2D source:** real images with 2D pose annotations

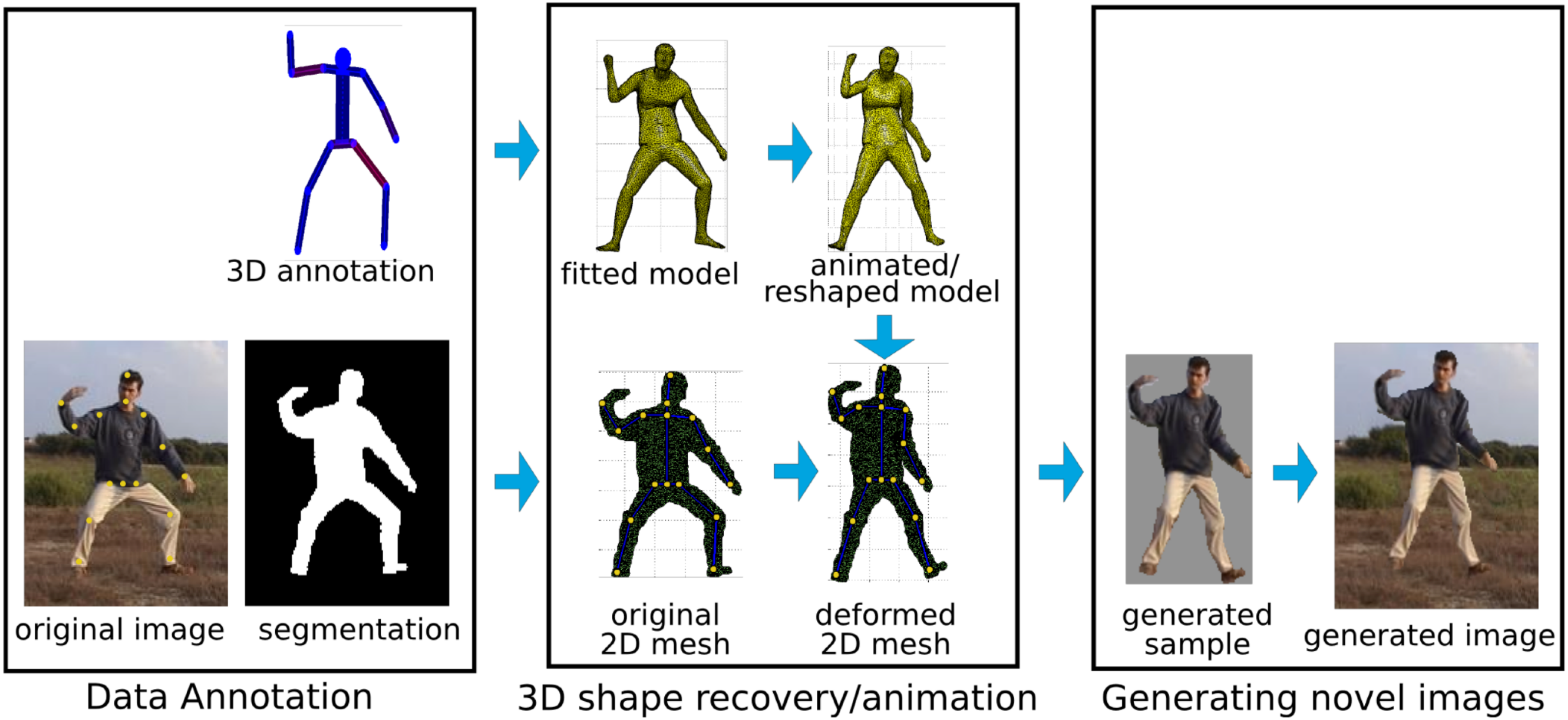
- Leeds Sport Dataset (LSP): 2,000 images [Johnson & Everingham 2010]
- Leeds Sport Dataset Extended (LSPE): 10,000 images [Johnson & Everingham 2011]
- MPII Human Pose Dataset: 25,000 images [Andriluka et al., 2014]
- COCO 39,000 images [Lin et al., 2014]

**3D source:** Motion Capture (MoCap) data

- CMU Graphics Lab Motion Capture MoCap Dataset 2500 sequences
- Pose Prior: [Akhter & Black 2015]
- Human3.6M dataset: 3.6M poses [Ionescu et al., 2014]

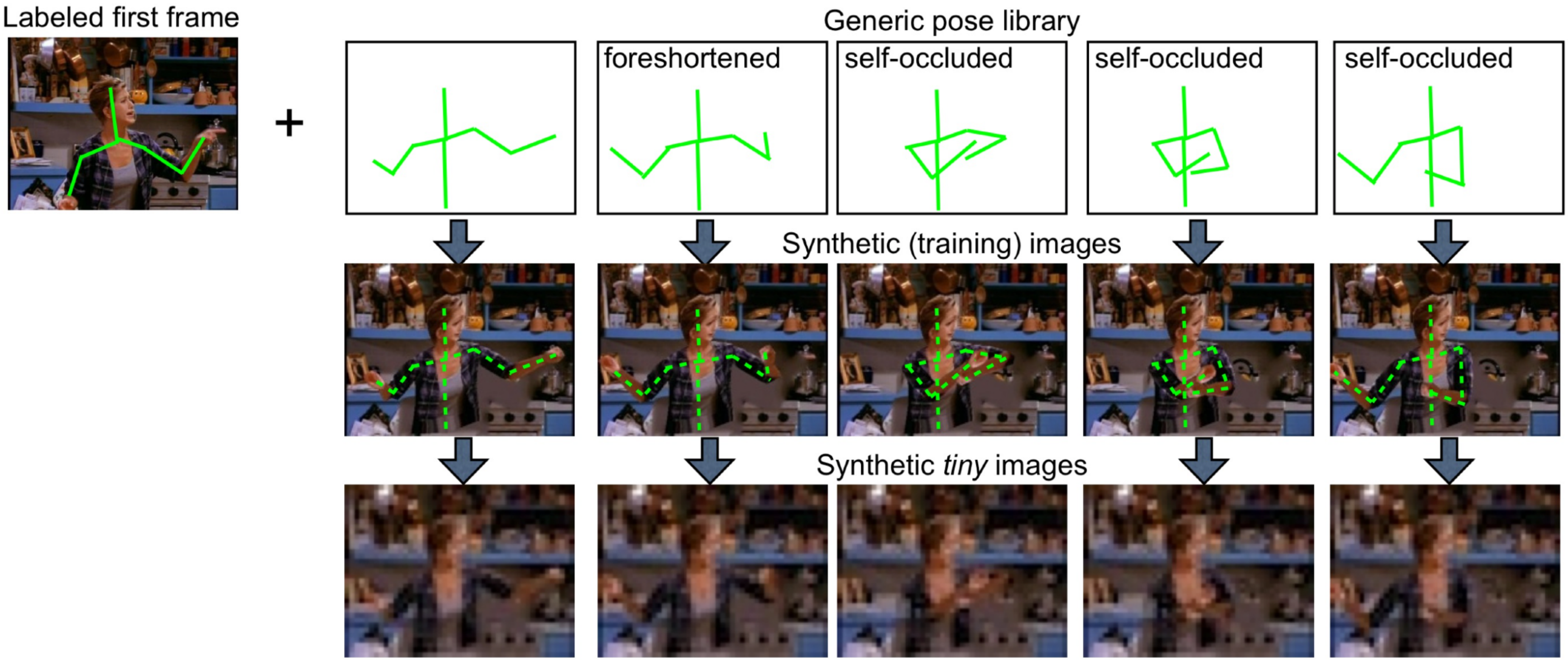


# POSE-BASED DATA AUGMENTATION




Reshaping the future [Pishchulin et al., CVPR 2012]

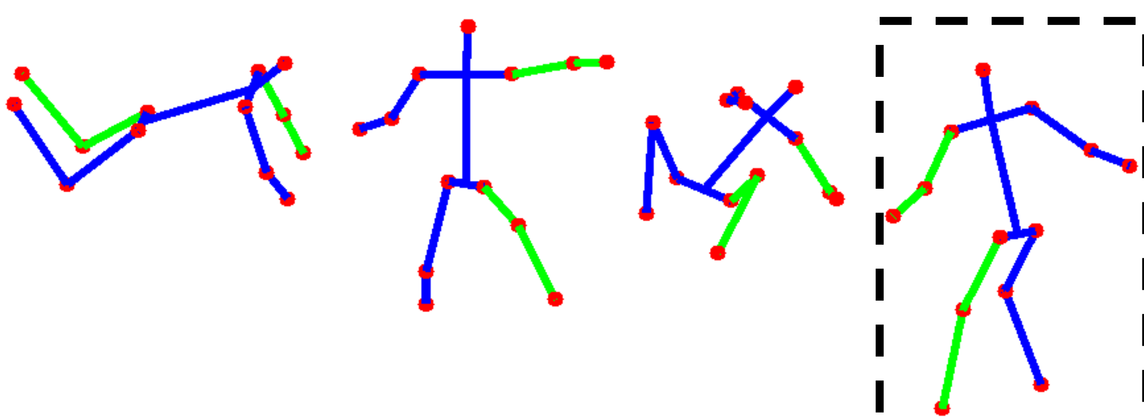
# POSE-BASED DATA AUGMENTATION




Tiny videos [Park and Ramanan, CVPRW 2015]

# OUR IMAGE-BASED SYNTHESIS ENGINE

**3D source:** Motion Capture (MoCap) data

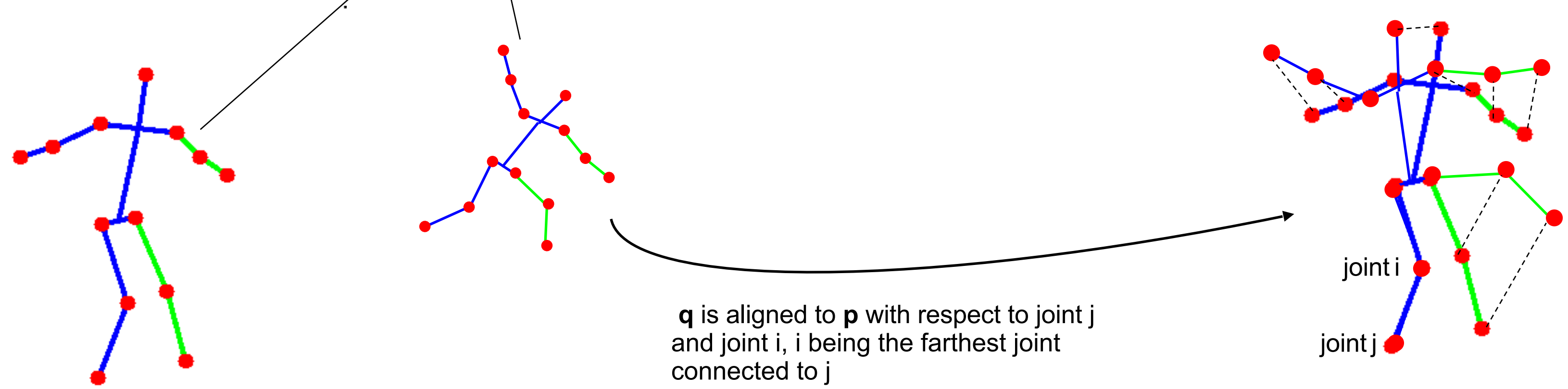


Camera views (x200 )

[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

- We define a distance between 2D poses **p** and **q** conditioned on one particular joint j:

$$D_j(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{n} (w_k^j(\mathbf{p}) + w_k^j(\mathbf{q}))\, d_{\mathrm{E}}(p_k, q'_k)$$




**q** is aligned to **p** with respect to joint j and joint i, i being the farthest joint connected to j

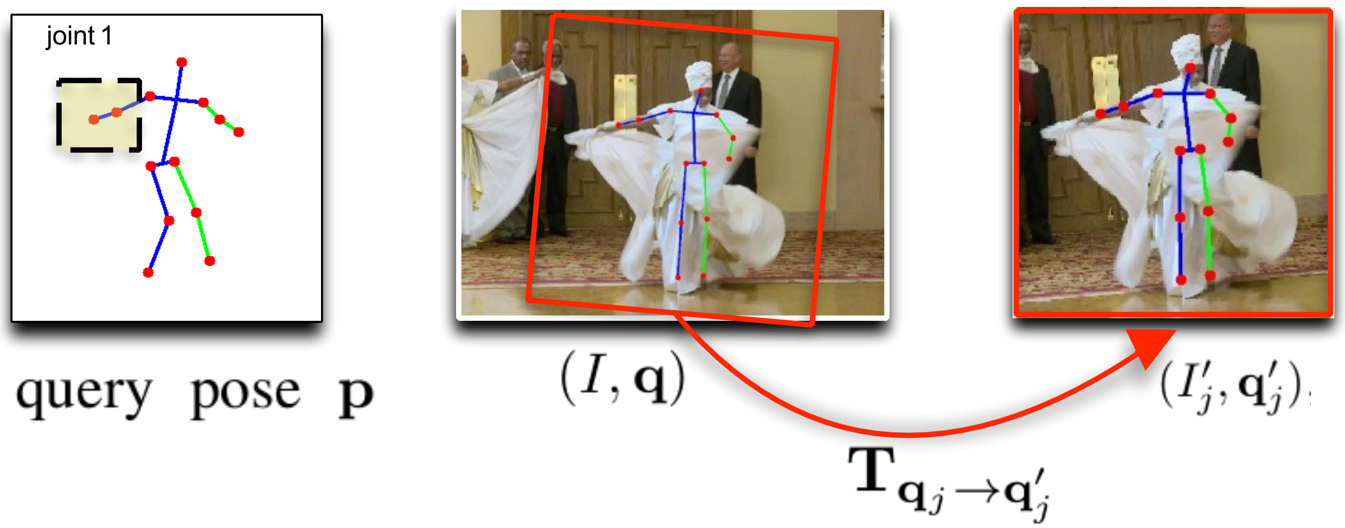
$\mathbf{T}_{\mathbf{q}_j \rightarrow \mathbf{q}'_j}$ respects: $q'_j = p_j$ and $q'_i = p_i$

We compute the Euclidean distance $d_{\mathrm{E}}(p_k, q'_k)$ between each pair of joints in **q'** and **p**

[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

# MOCAP-GUIDED IMAGE MOSAICING

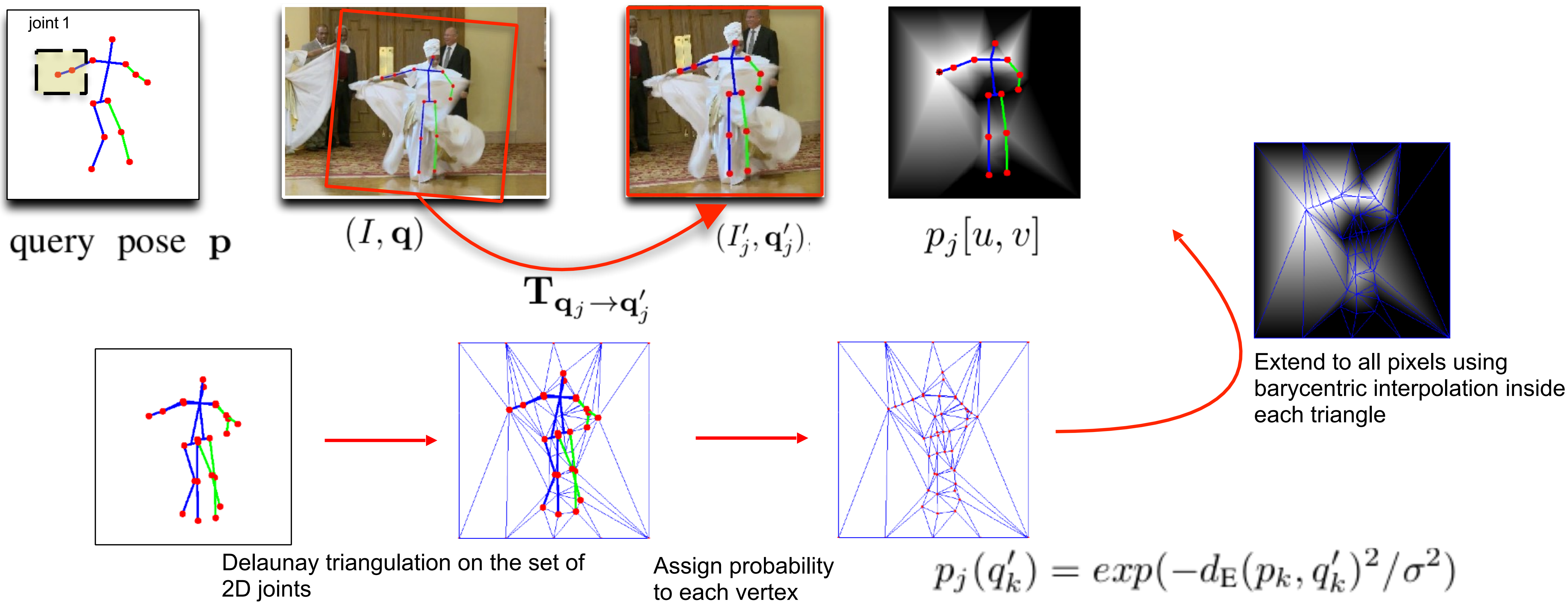
- For each joint of the query pose, we search our dataset of annotated 2D poses to find the image with similar local kinematic configuration.




$$\mathbf{q}_j = \operatorname{argmin}_{\mathbf{q} \in \mathbb{Q}} D_j(\mathbf{p}, \mathbf{q})$$

[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

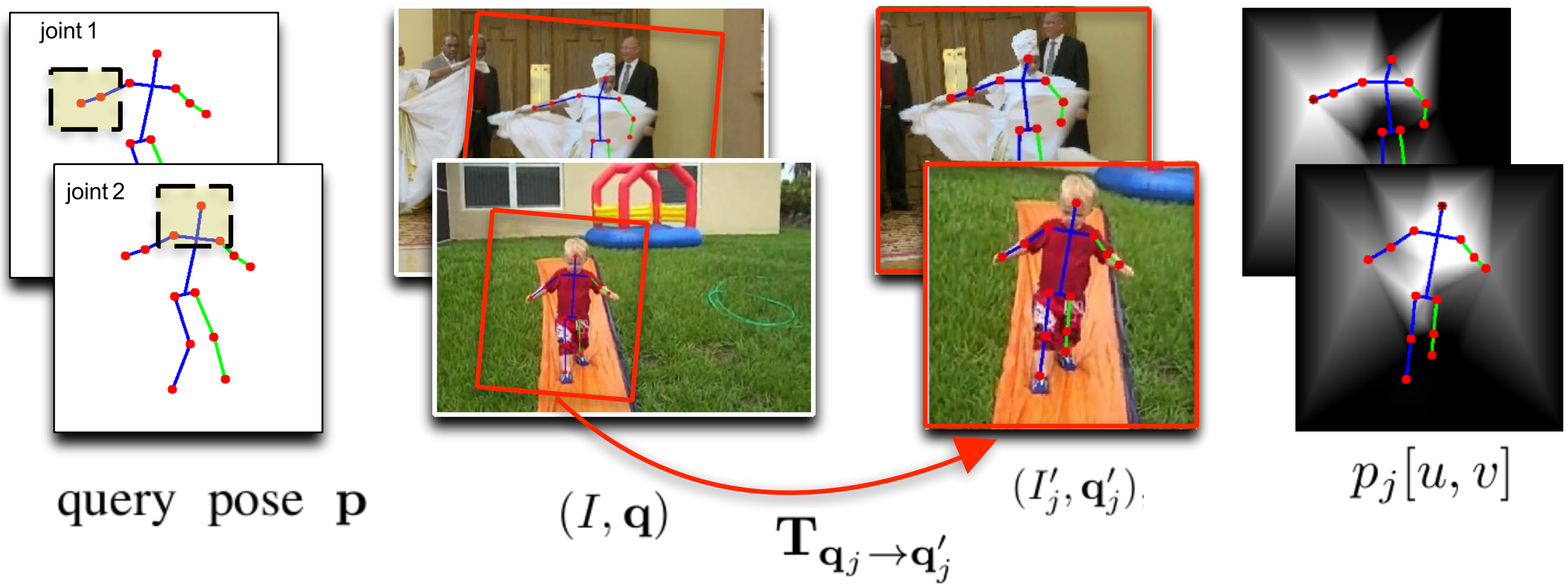
- For each joint of the query pose, we compute a probability map $p_j[u,v]$




[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

# MOCAP-GUIDED IMAGE MOSAICING

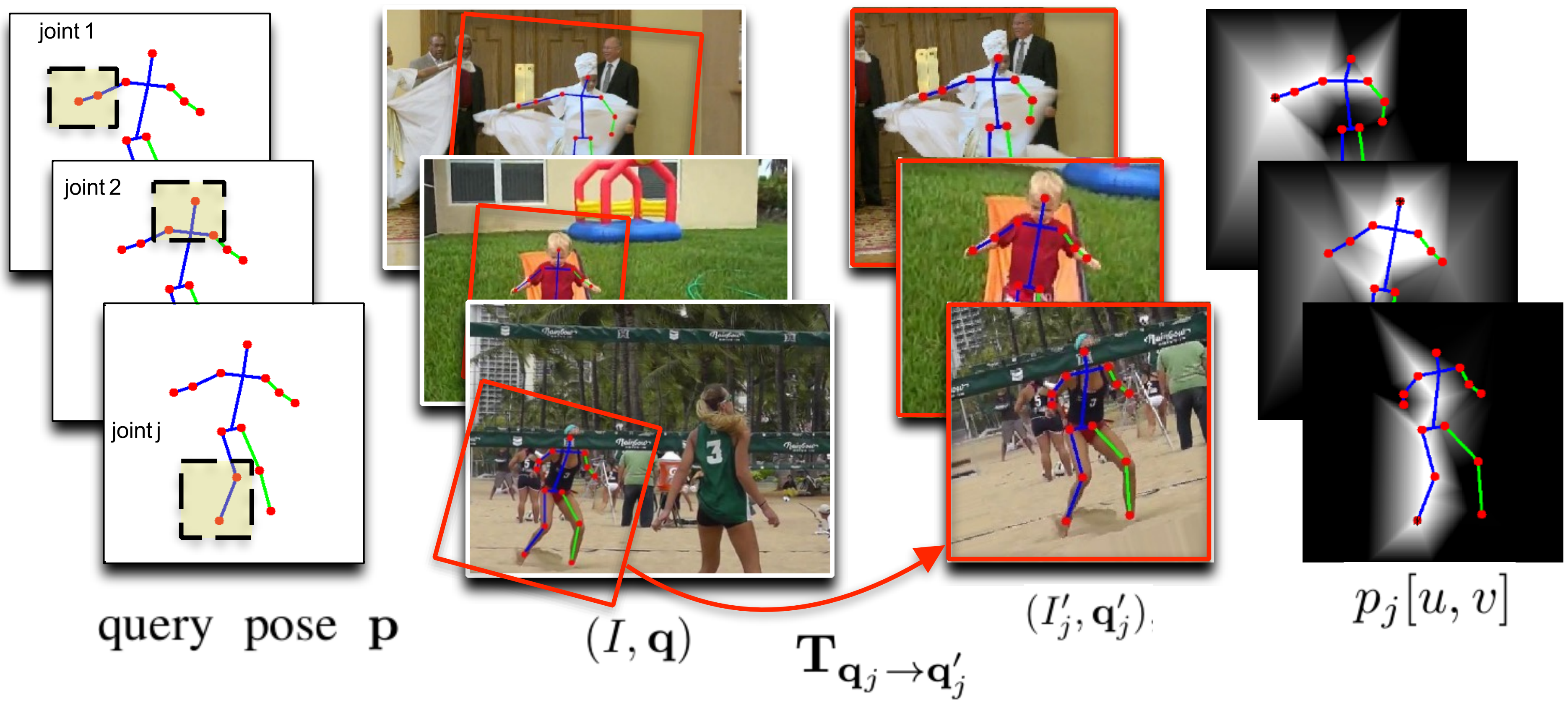
- Repeat the process for all joints




[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

# POSE AWARE IMAGE BLENDING

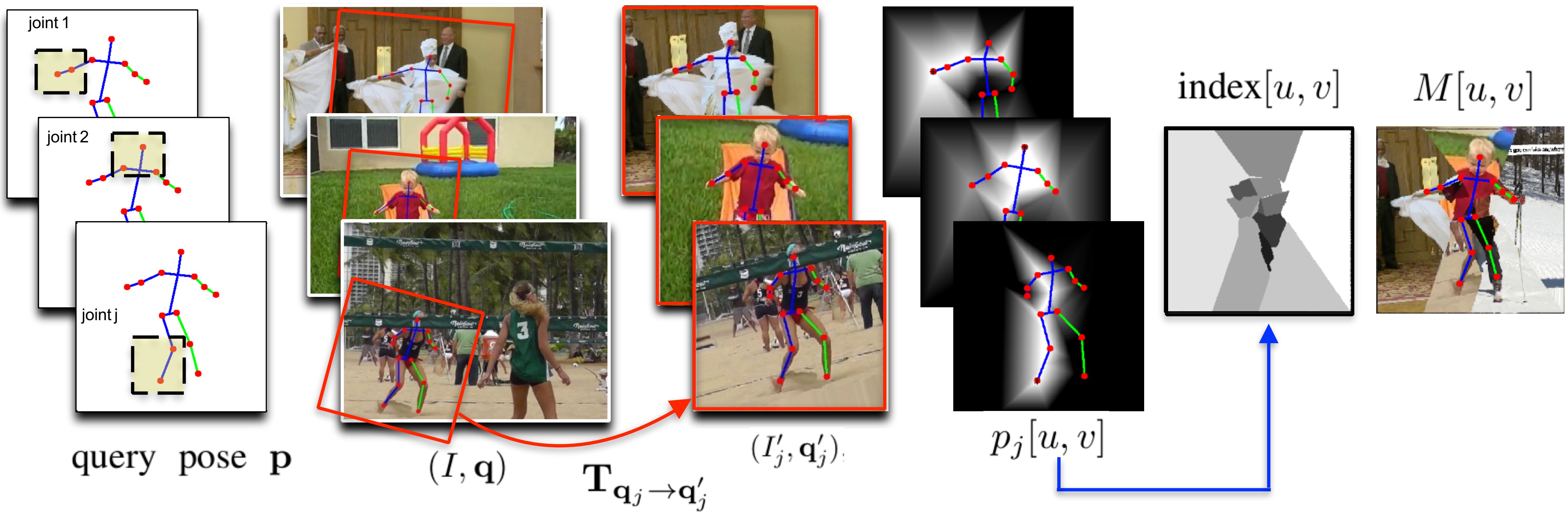
- Repeat the process for all joints




Obtaining a list of n matches $\{(I'_j, \mathbf{q}'_j), j = 1...n\}$

[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

- Taking the argmax over probability maps generate a mosaic with artifacts

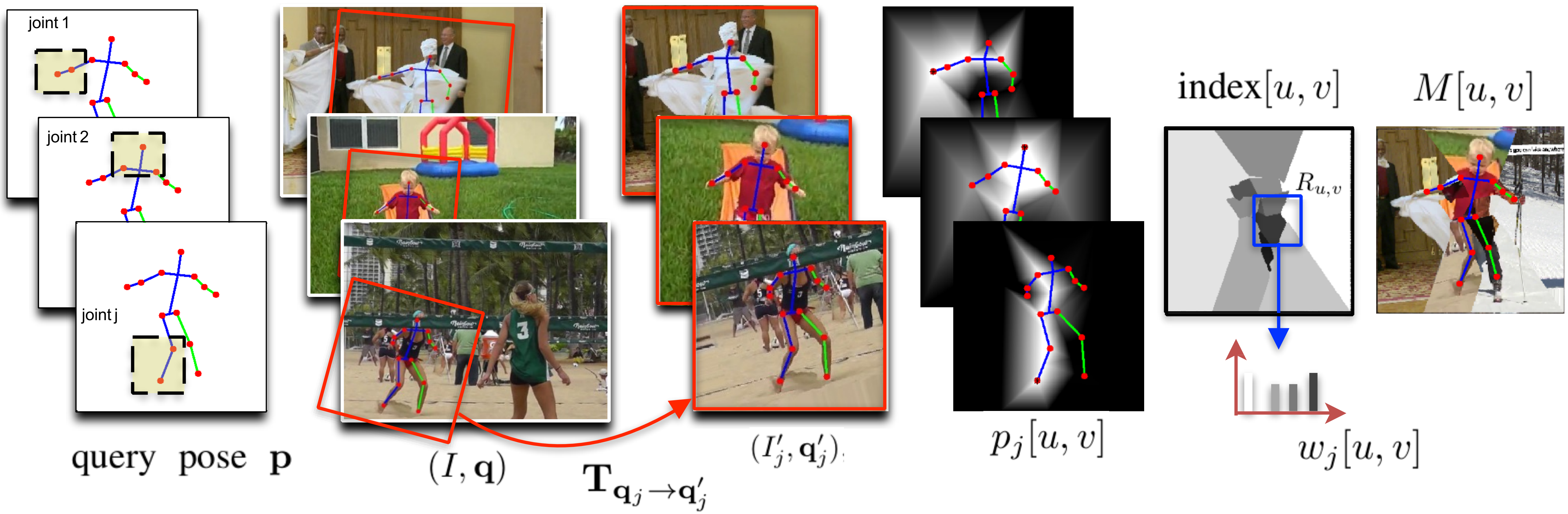



$$\text{index}[u, v] = \text{argmax}_{j \in \{1 \ldots n\}}\, p_j[u, v]$$

[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

# POSE AWARE IMAGE BLENDING

- Instead, we use a pose-aware image blending algorithm with a squared region $R_{u,v}$ whose size varies with the distance to the pose

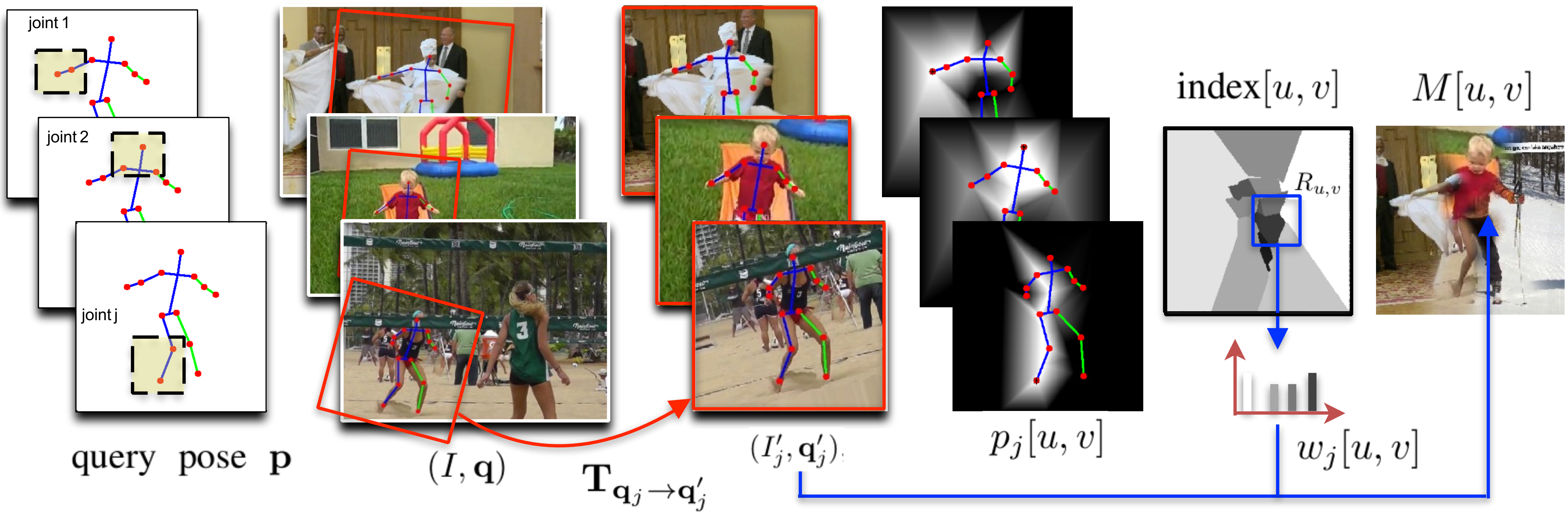



Build an histogram of indices inside

[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

# POSE AWARE IMAGE BLENDING

- Instead, we use a pose-aware image blending algorithm with a squared region $R_{u,v}$ whose size varies with the distance to the pose



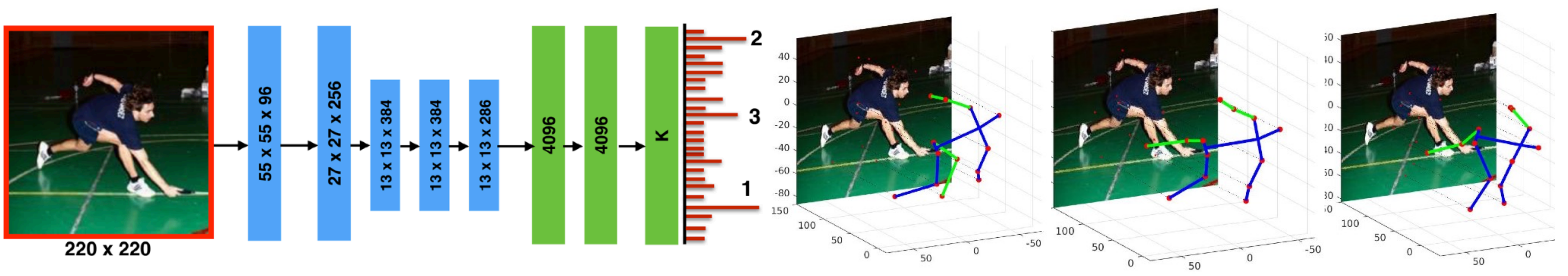

Compute pixel values as a weighted sum over all images:

$$M[u, v] = \sum_j w_j[u, v] I'_j[u, v]$$

[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

# CNN FOR FULL-BODY 3D POSE

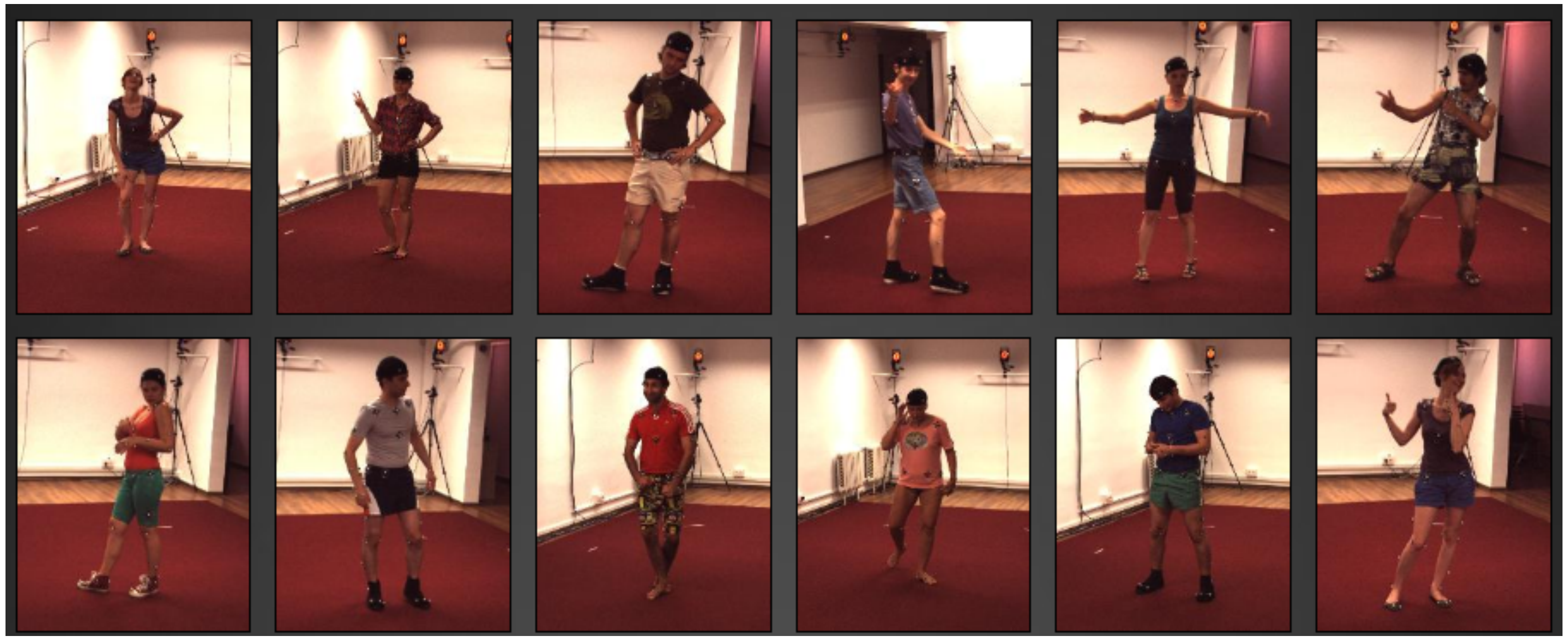
- 3D pose space partitioned into K clusters (K=5000)
- AlexNet adapted to output a probability distribution over pose classes.




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

[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS’16]

# EVALUATION ON HUMAN3.6M



**H3.6M details:**
- **3.6 millions** images (4 cameras)
- **1 environment** (MoCap room)
- **11 actors**
- **17 activities** (discussion, smoking, taking photo, talking on the phone…)
- **3D MoCap data** (human poses)
- **2D joint location**

| Type of images | 2D source size | 3D source size | 3D pose error (mm) |
|---|---|---|---|
| Real | 190,000 | 190,000 | 97.7 |
| Synth | 17,000 | 190,000 | 97.2 |
| Synth + Real | 190,000 | 190,000 | **88.1** |

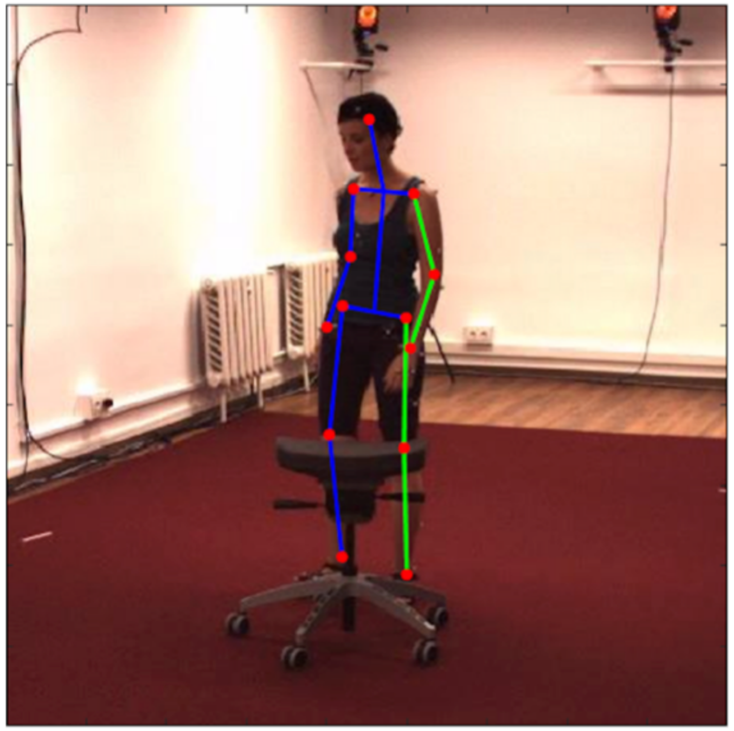
Our data is **useful:** classifier performs slightly better when trained on Synth data

Our data is **different:** classifier performs better when trained on Real and Synth data together

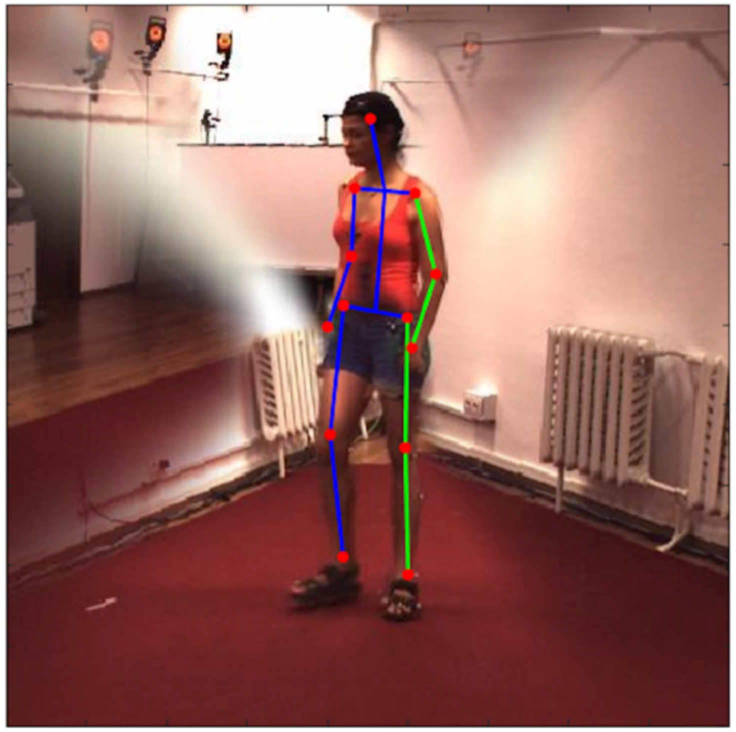
[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

# COMPARAISON WITH "CLASSICAL" SYNTHESIS

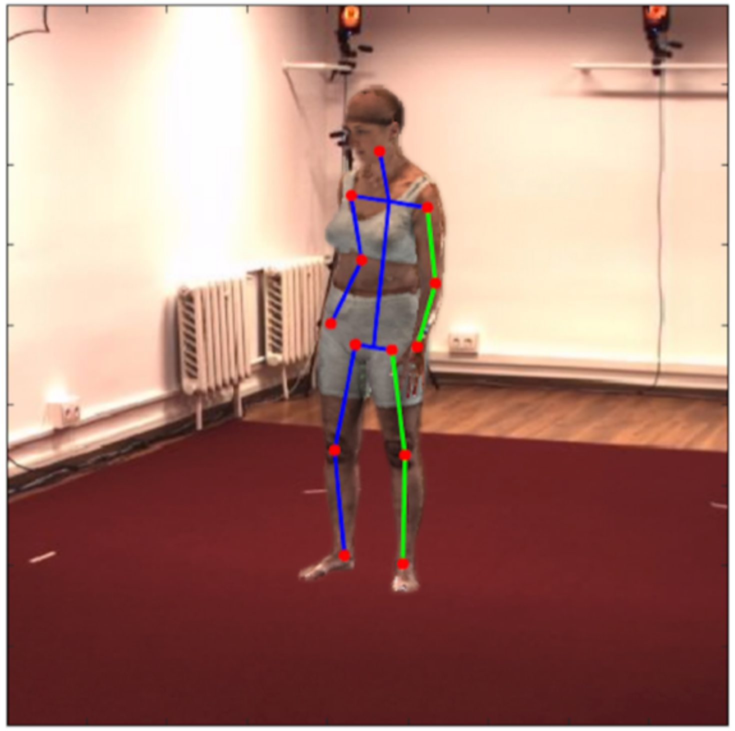
Real



Mosaic



Surreal



| Type of images | 2D source size | 3D source size | 3D pose error (mm) |
|---|---|---|---|
| Mosaic | 17,000 | 190,000 | 97.2 |
| Surreal | 0 | 190,000 | 119.5 |
| Surreal +Real | 190,000 | 190,000 | 97.8 |
| Surreal+Mosaic | 17,000 | 190,000 | 90.1 |

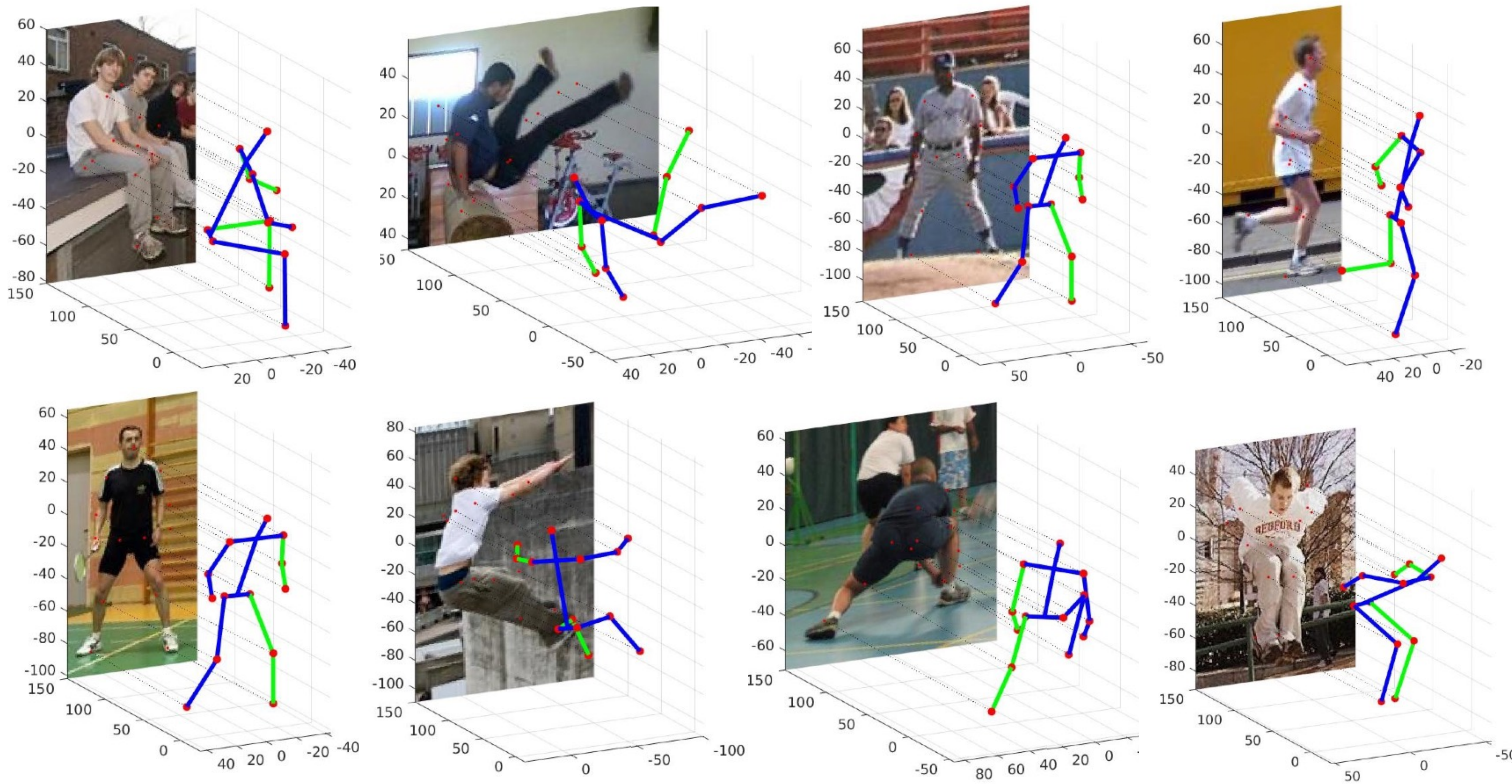
Training on **Surreal alone overfits** and does not generalize well

**Mixing** Surreal with Real images helps avoid overfitting

**Combining** Surreal and Mosaic images results in a better model

[Rogez& Schmid, Image-based Synthesis for Deep 3D Human Pose Estimation. IJCV 2018]

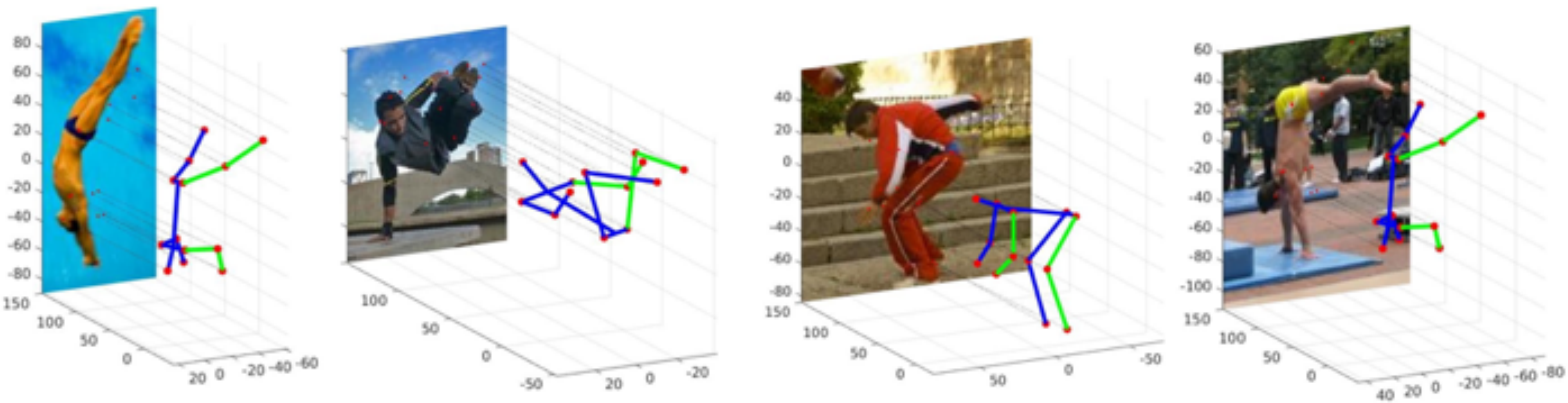
# QUALITATIVE RESULTS



- CNNs can be trained on artificially looking images and still generalize well to real images

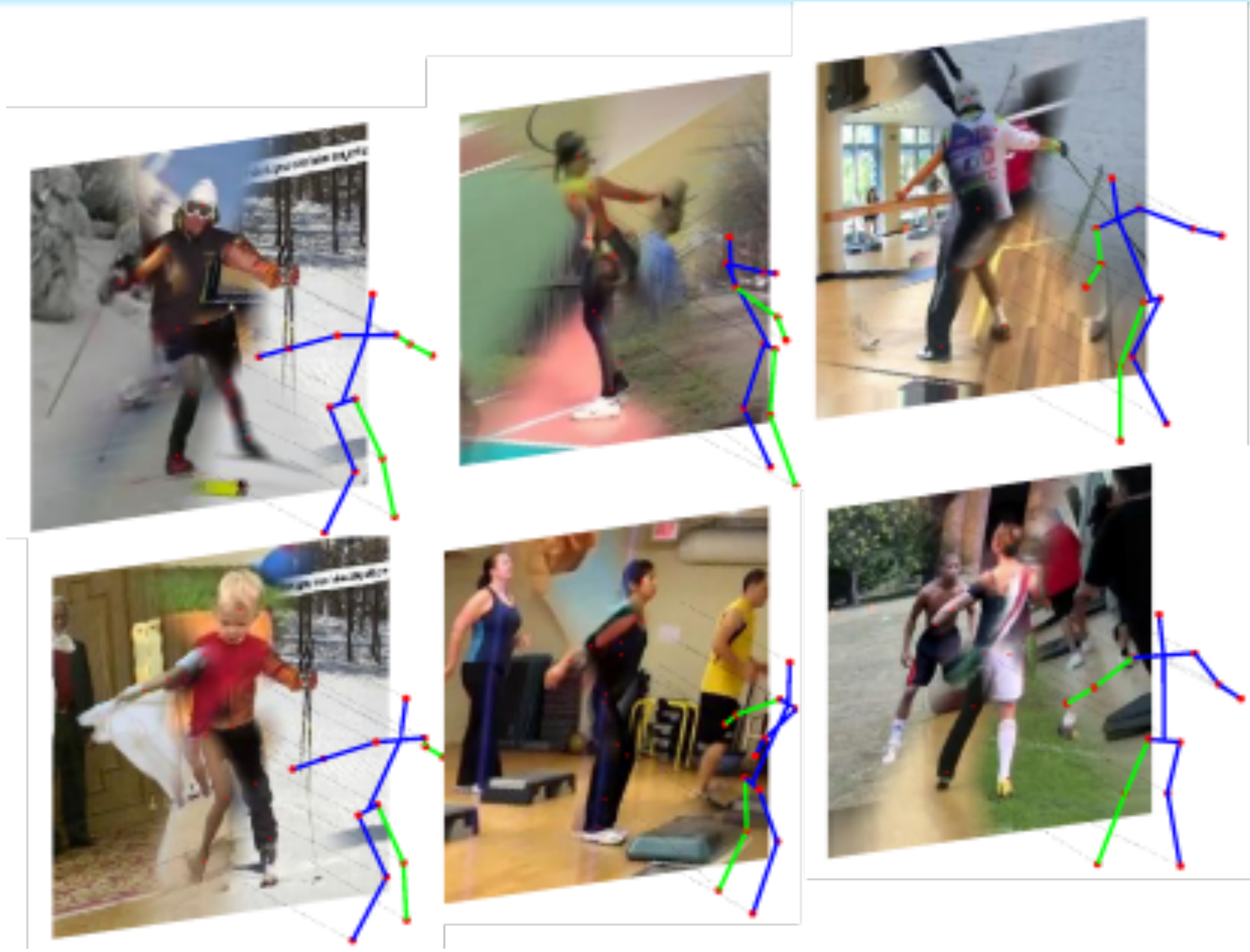
[Rogez& Schmid, Image-based Synthesis for Deep 3D Human Pose Estimation. IJCV 2018]

## Failure cases



[Rogez& Schmid, MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. NIPS'16]

# TAKE HOME MESSAGE



Data augmentation technique to synthesize (large scale) in-the-wild **images with 3D pose annotations:**

- locally **photorealistic** (no need for domain adaptation)
- **kinematically** coherent

# CLASSIFICATION: DRAWBACKS

Requires **large scale training data (images+3D pose)**

**SYNTHESIS**

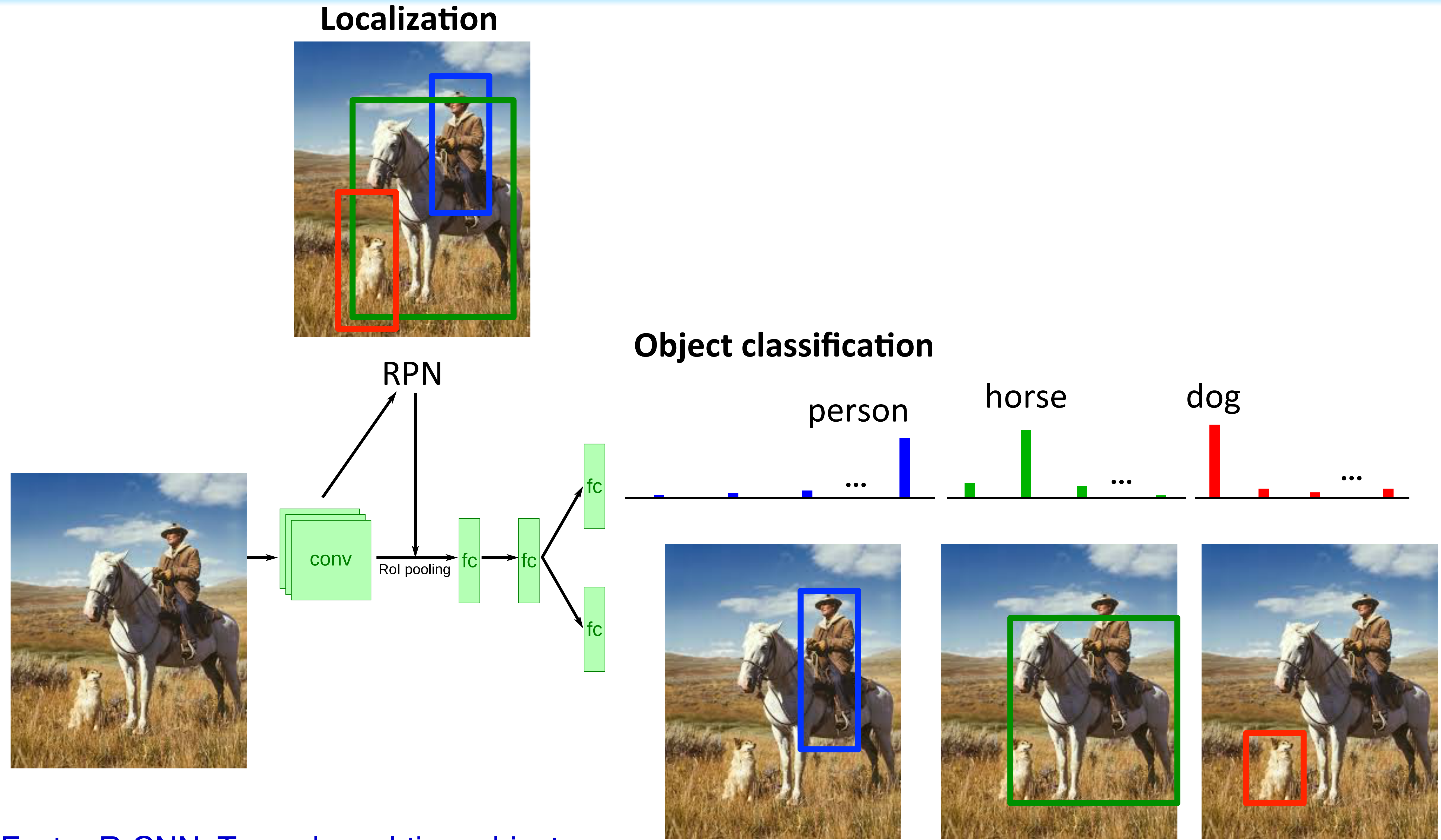
Won't work with **unseen poses**

Only **coarse pose** estimation

**CNN**

Computational **cost**

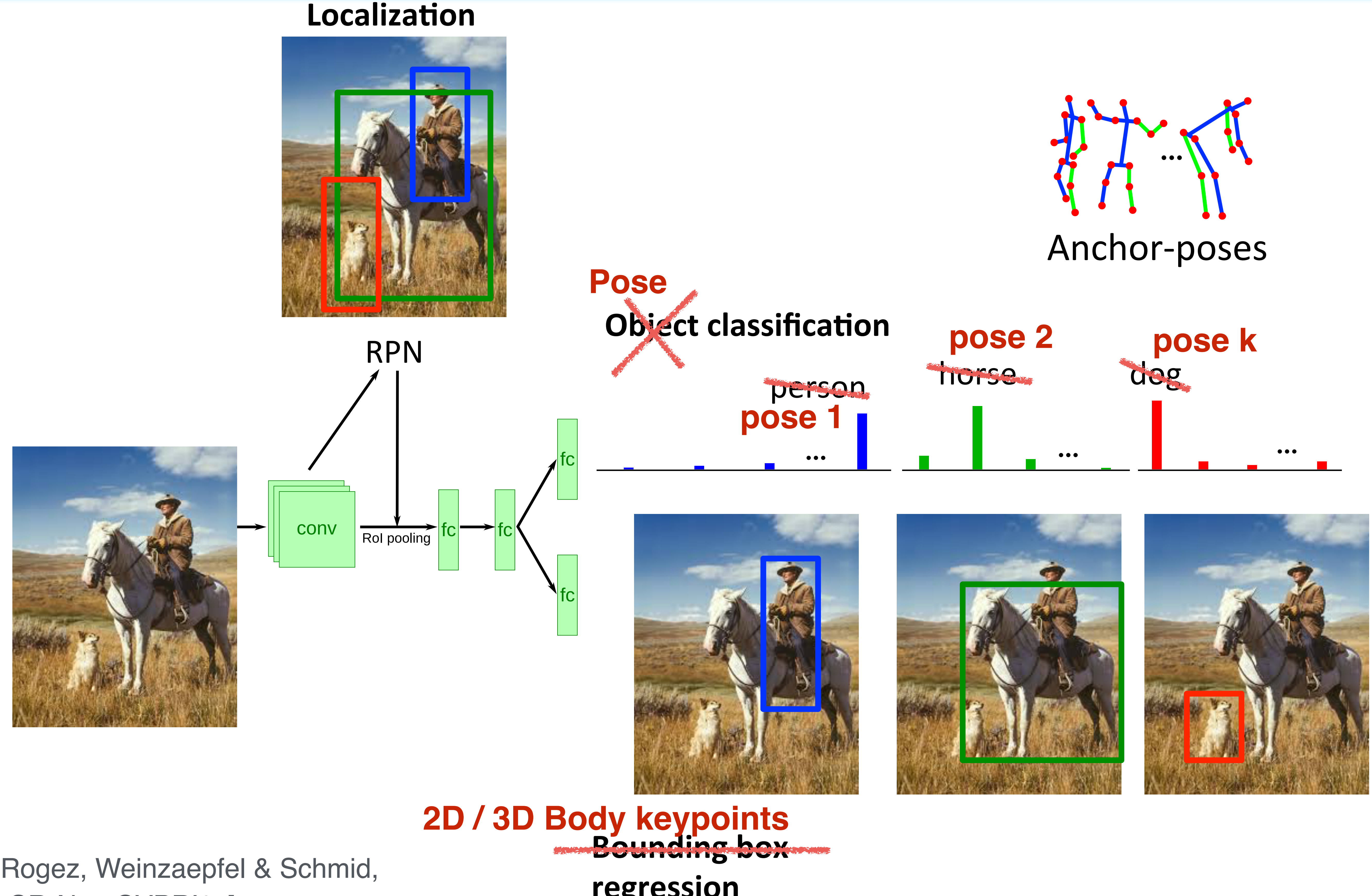
# RELATED WORK: FASTER R-CNN




**Bounding box regression**

Faster R-CNN: Towards real-time object detection with region proposal networks
[S Ren et al., NIPS 2015, PAMI 2017]

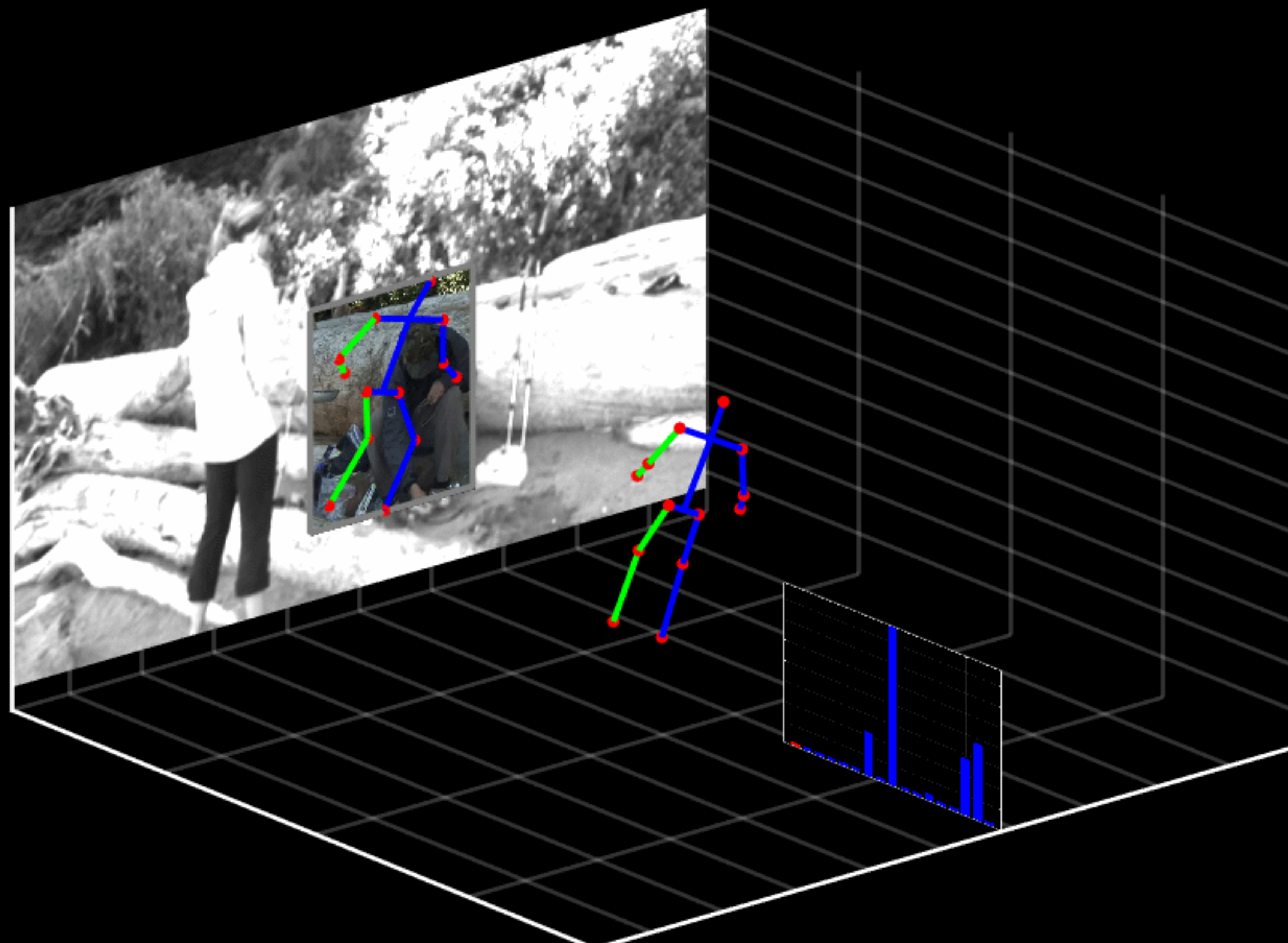
# RELATED WORK: FASTER R-CNN

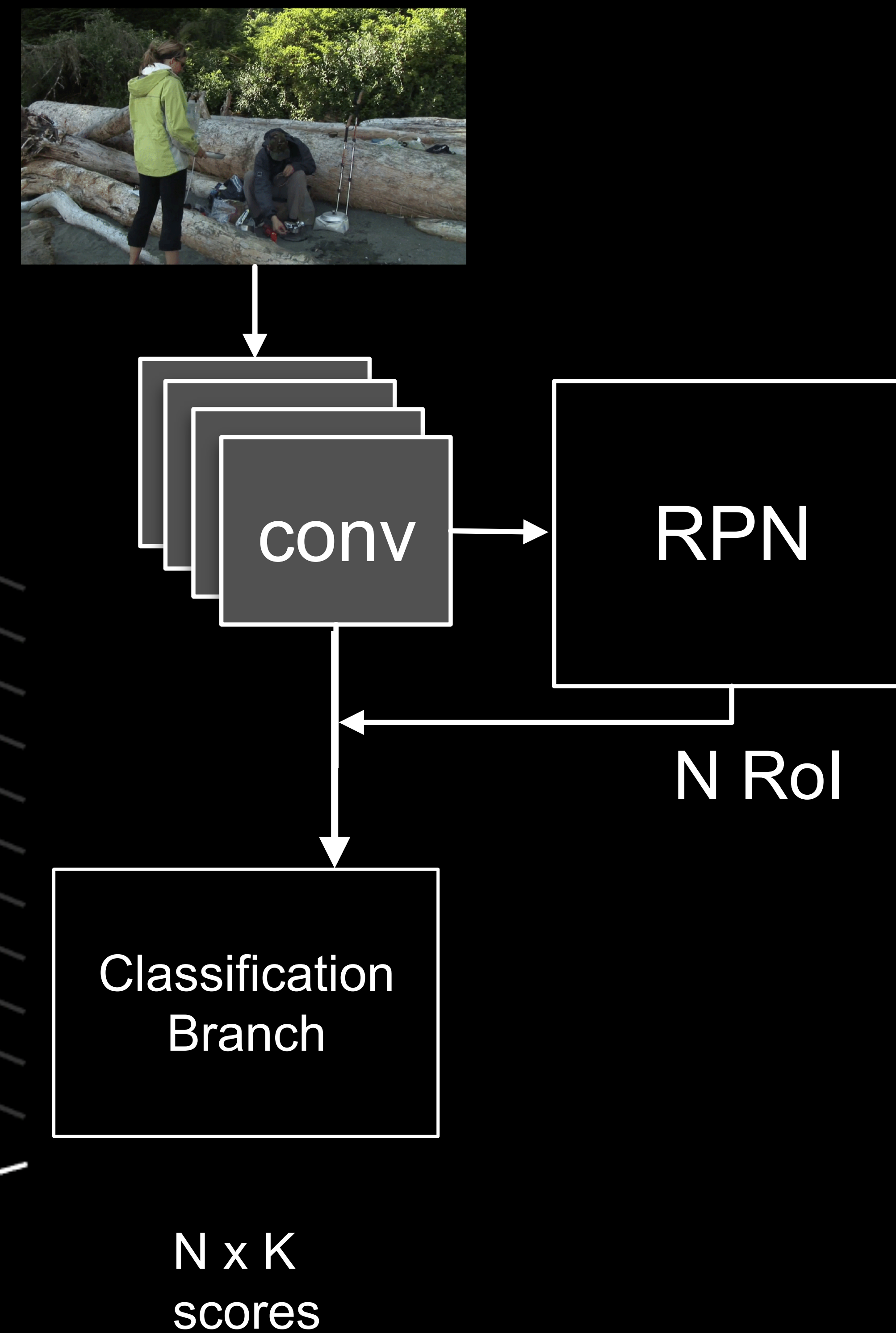



[Rogez, Weinzaepfel & Schmid, LCR-Net. CVPR'17]

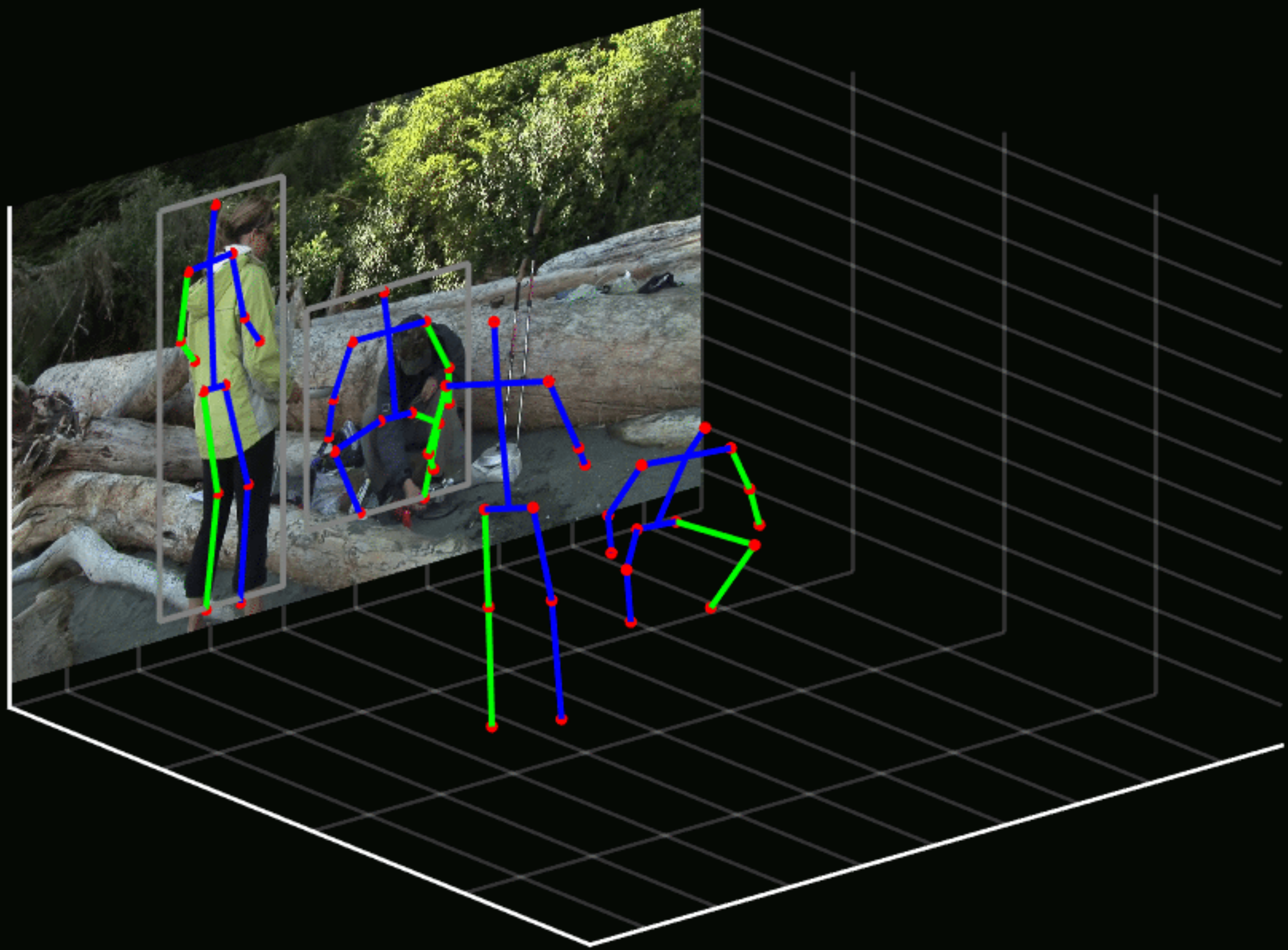
# LCR-NET: LOCALIZATION

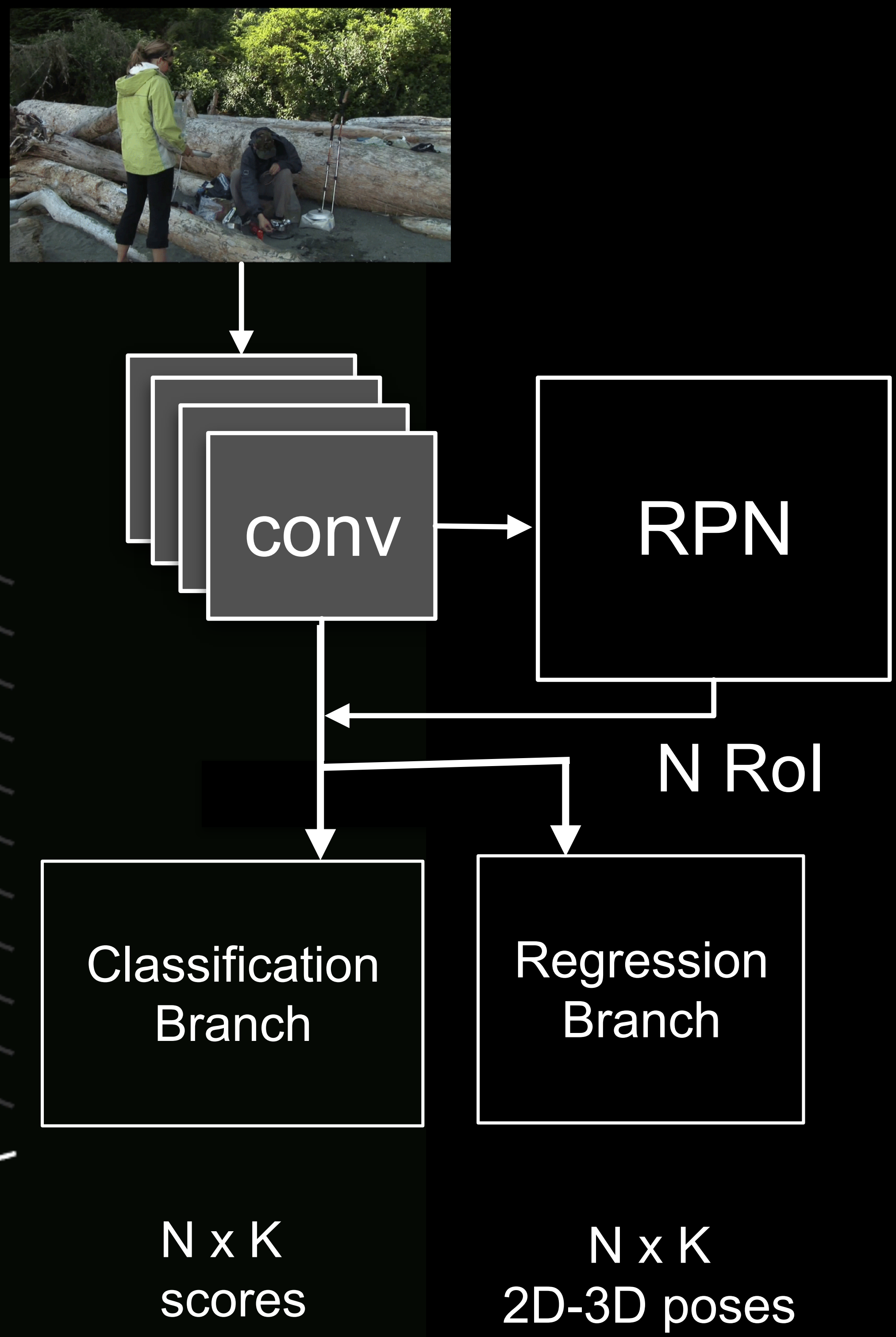
# LCR-NET: CLASSIFICATION





# LCR-NET: REGRESSION

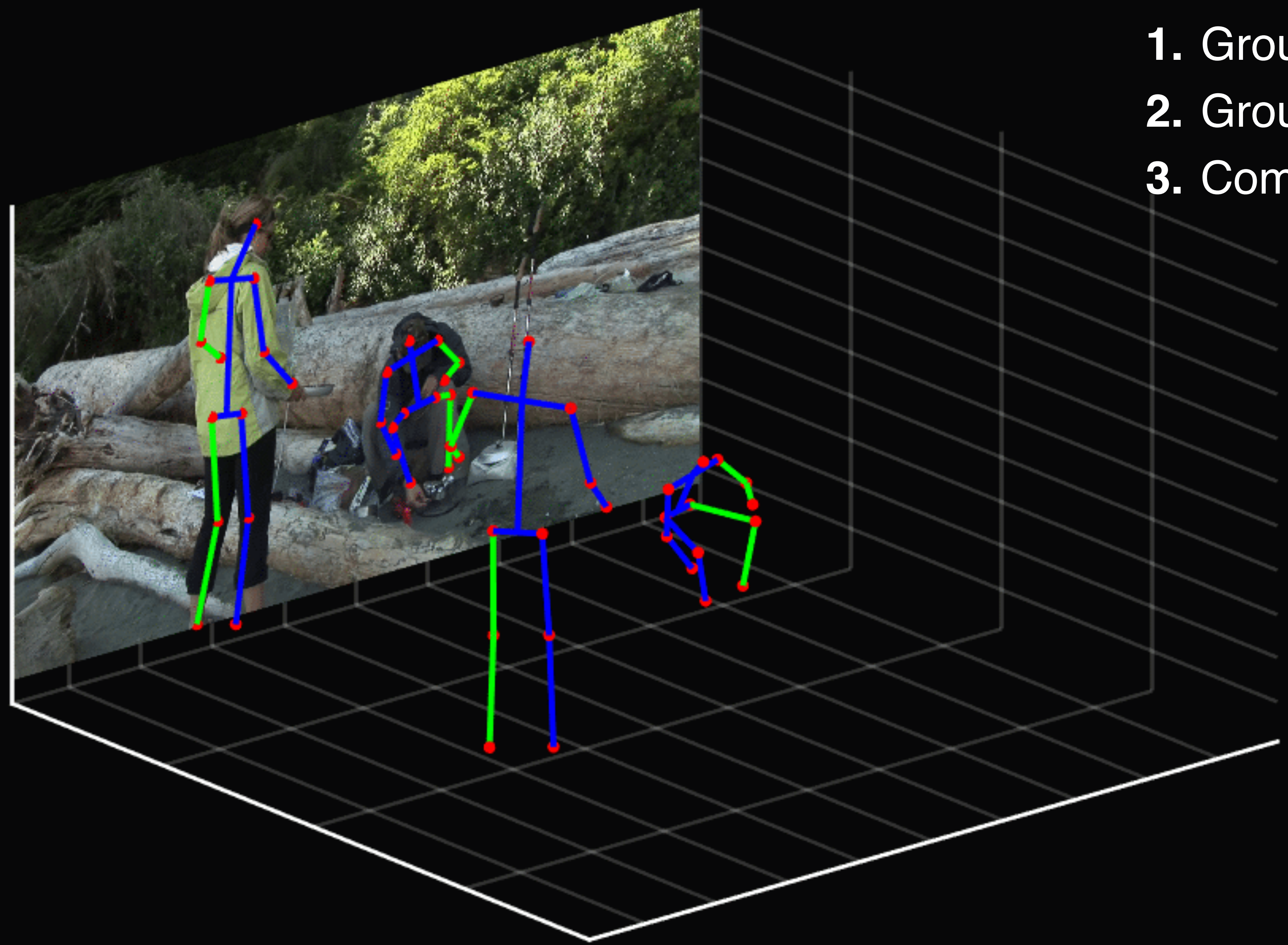




# LCR-NET: POSE PROPOSALS INTEGRATION (PPI)

N x K refined 2D/3D pose proposals + scores

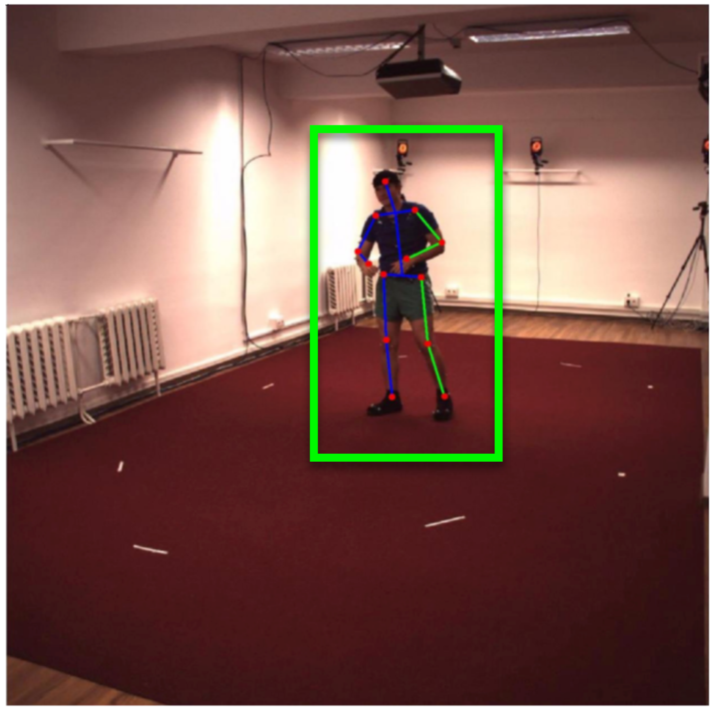
Pose estimation by NMS.



Instead we:

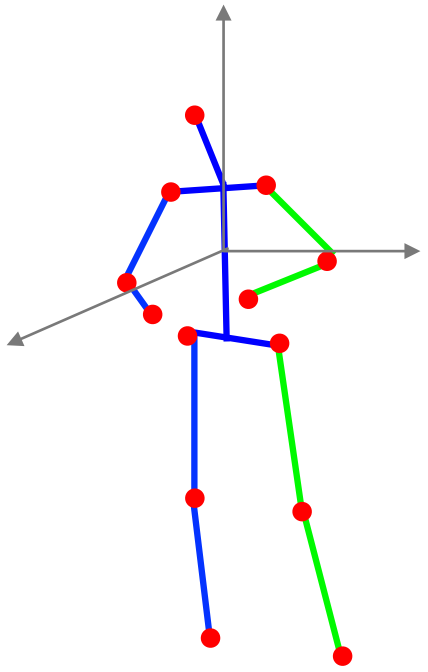
1. Group using **2D overlap** to find persons
2. Group using **3D pose** to find modes
3. Compute weighted sum

# LCR-NET: TRAINING (HUMAN 3.6M)



**Bounding box + class label**

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



**normalized 3D poses**
(aligned+orientated)

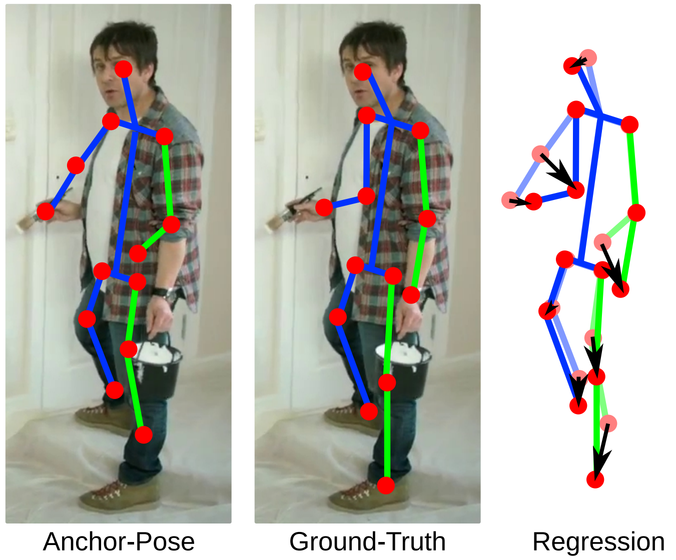
**Loss:**

$$\mathcal{L} = \mathcal{L}_{Loc} + \mathcal{L}_{Classif} + \mathcal{L}_{Reg}$$

**RPN loss** (cf FasterRCNN)

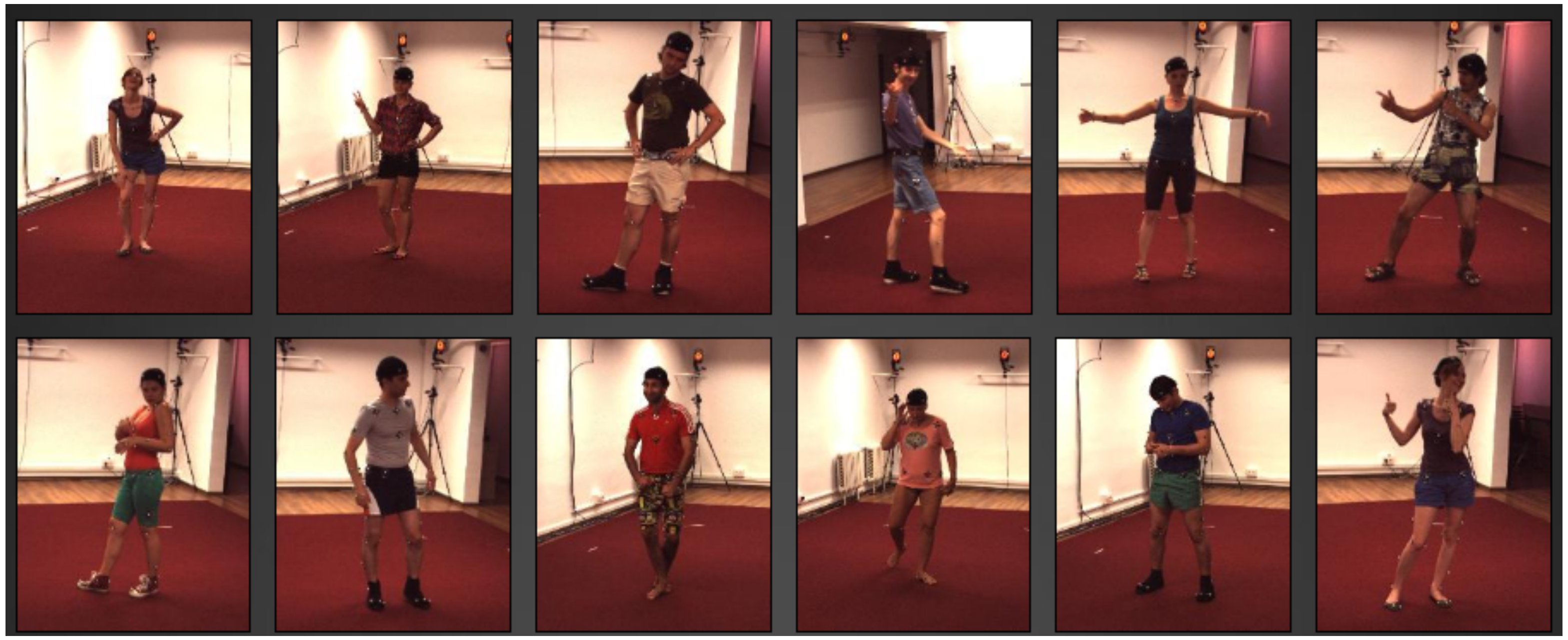
**log loss of the true class**

**L1-smooth loss**



# EVALUATION

**3D Evaluation:** Human3.6M dataset



- 300k training images with perfect bounding boxes, 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** |

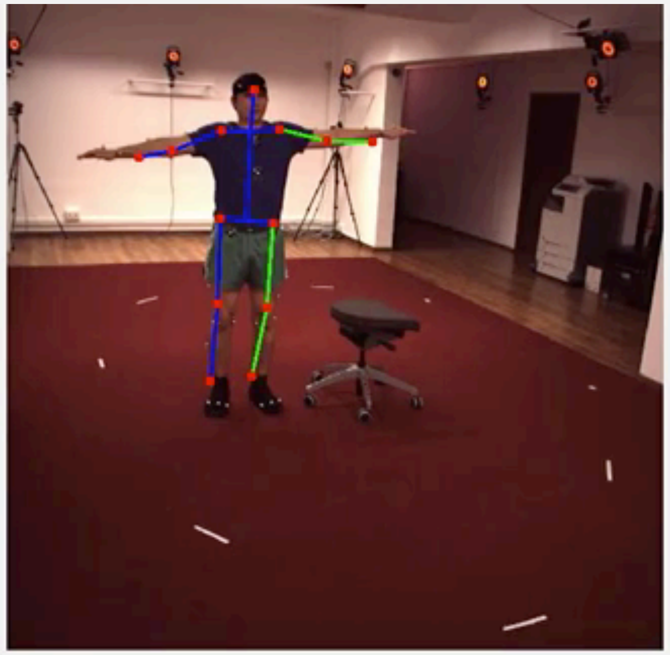
Best performance achieved for **K=100** anchor poses.

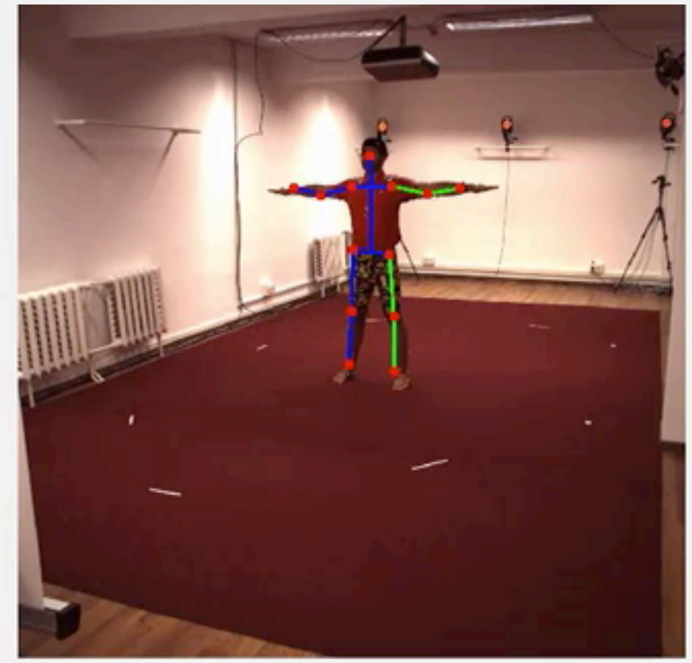
Boost with Synth. data

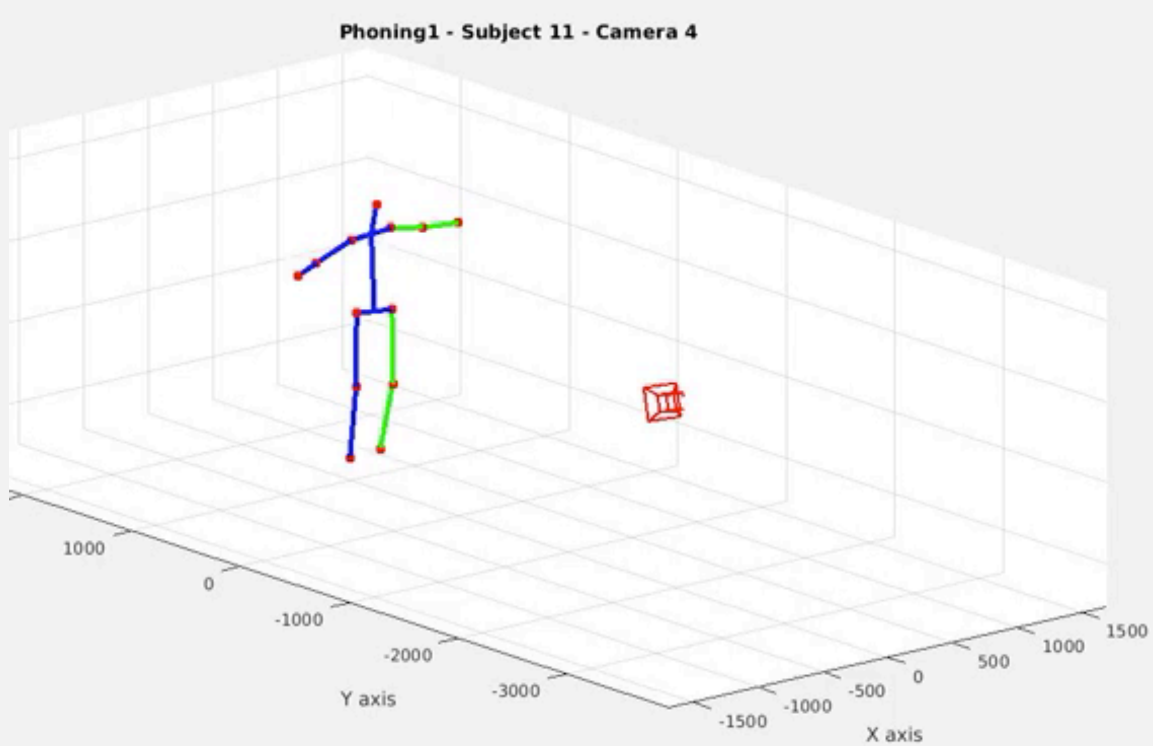
Small improvement with ResNet50

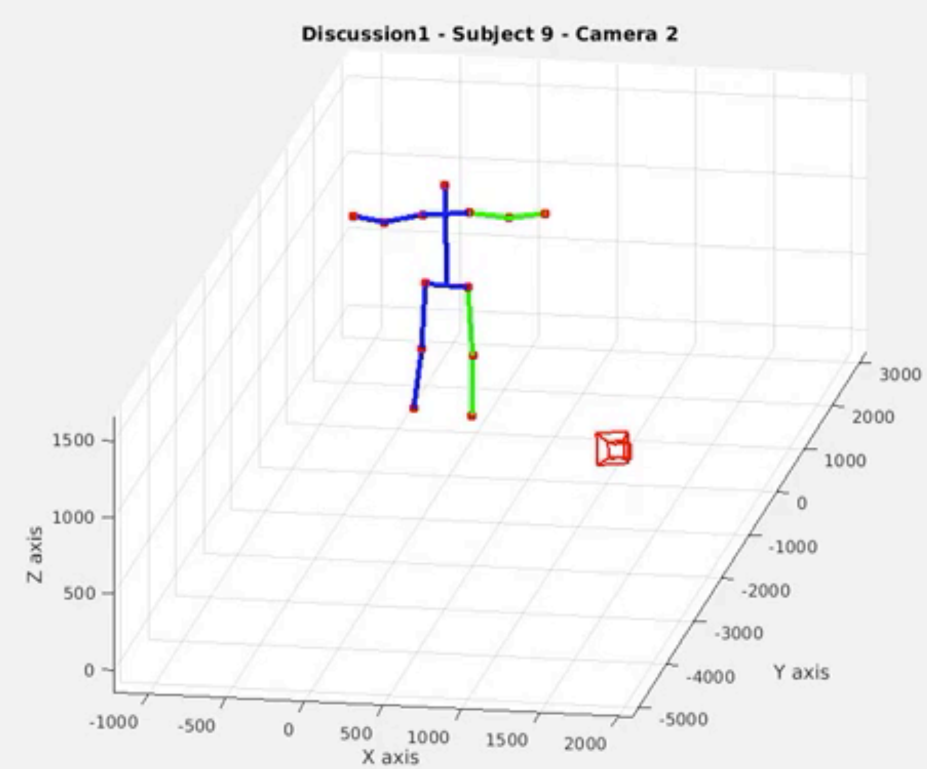
[Rogez, Weinzaepfel & Schmid, LCR-Net++. IEEE T. PAMI 2019]

**3D Evaluation:** Human3.6M dataset



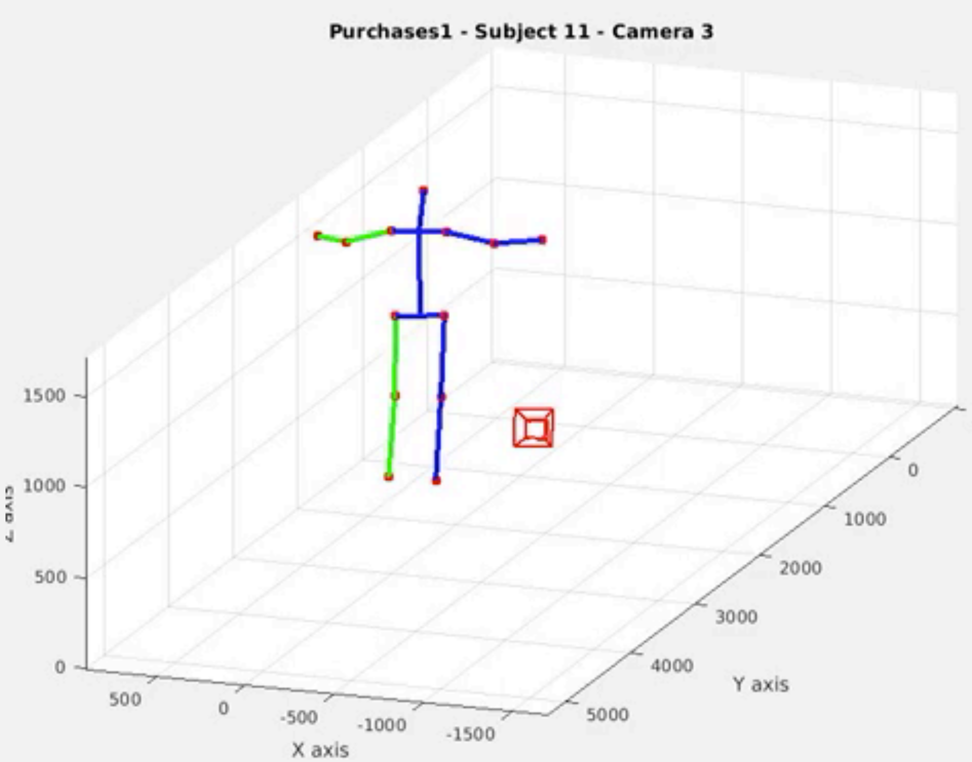


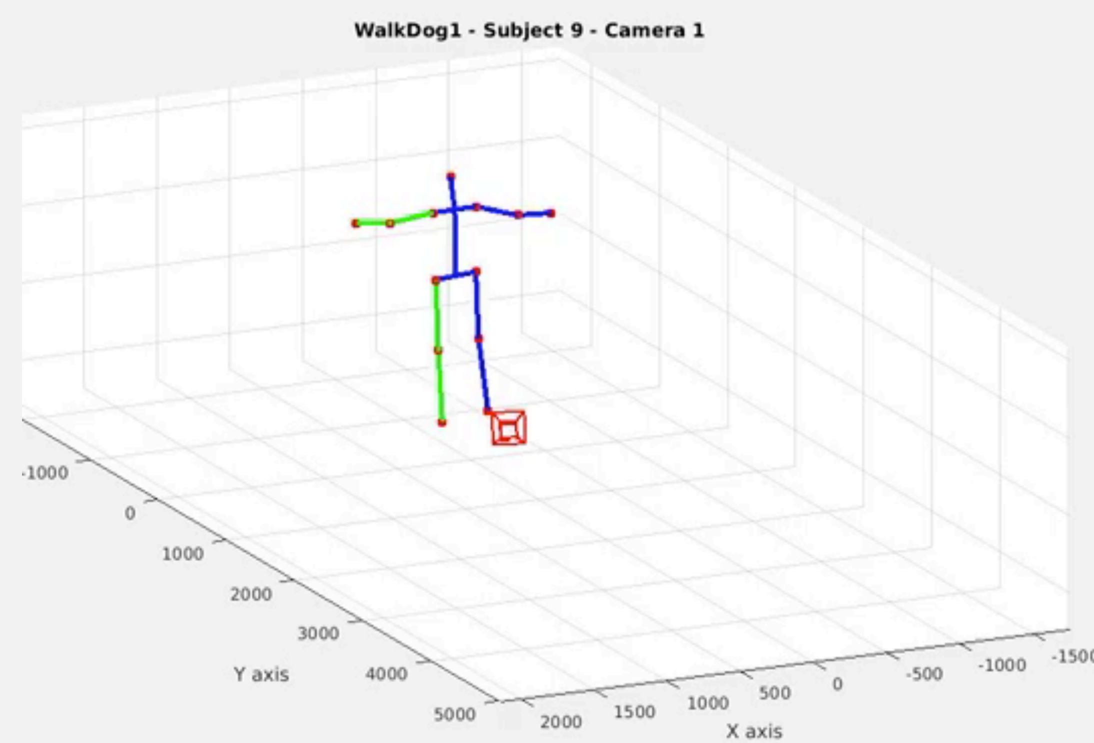


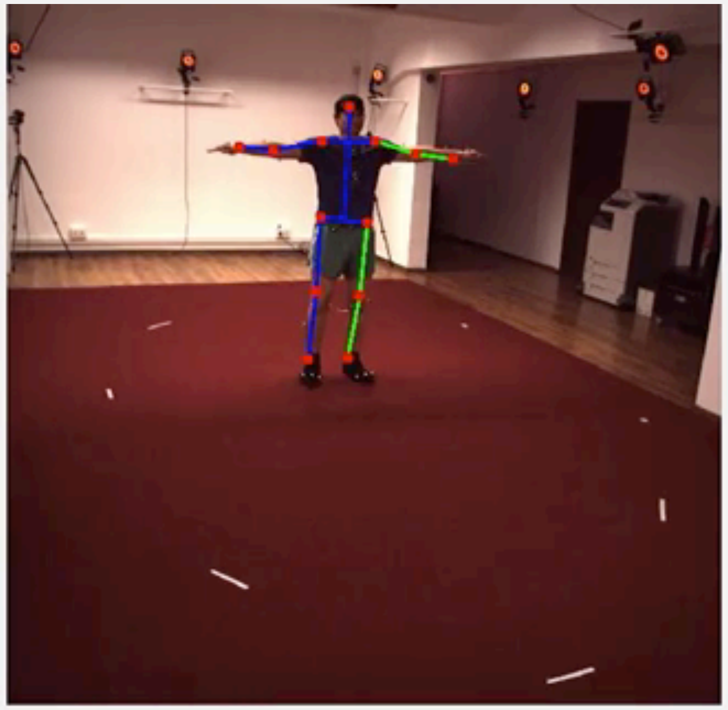
# QUALITATIVE RESULTS

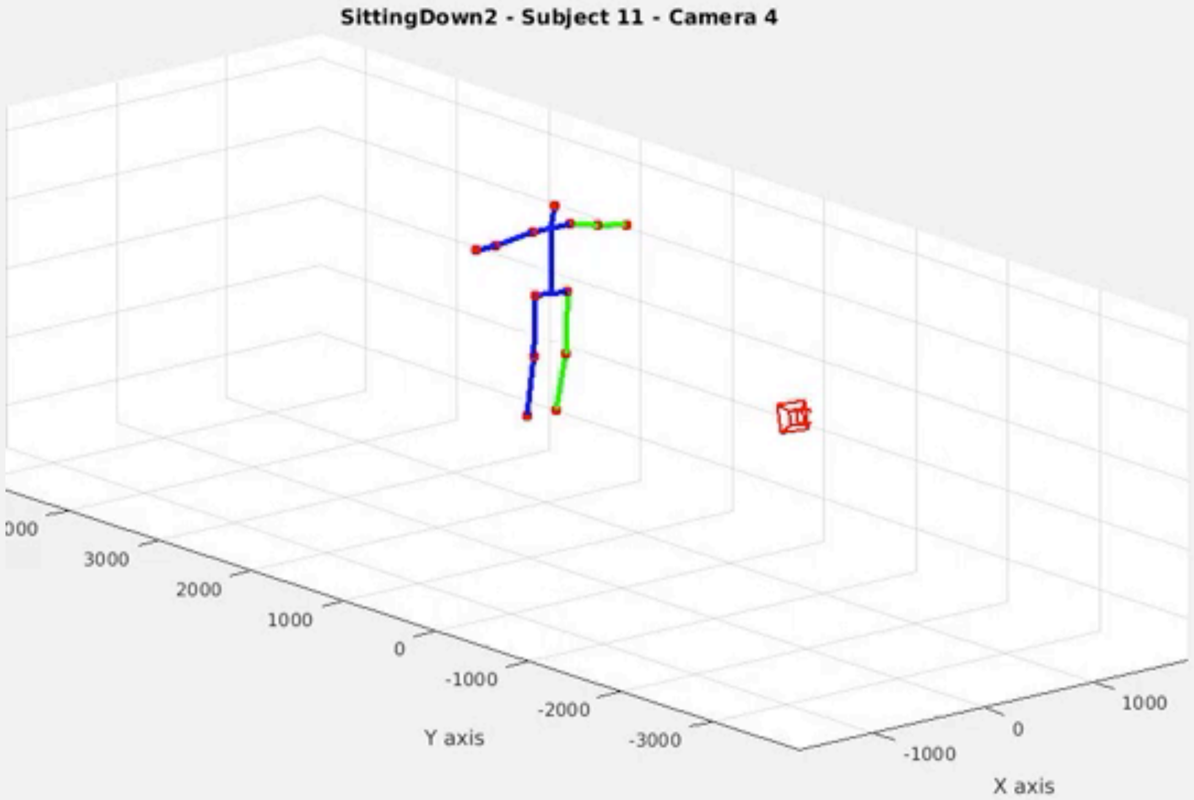
**3D Evaluation:** Human3.6M dataset



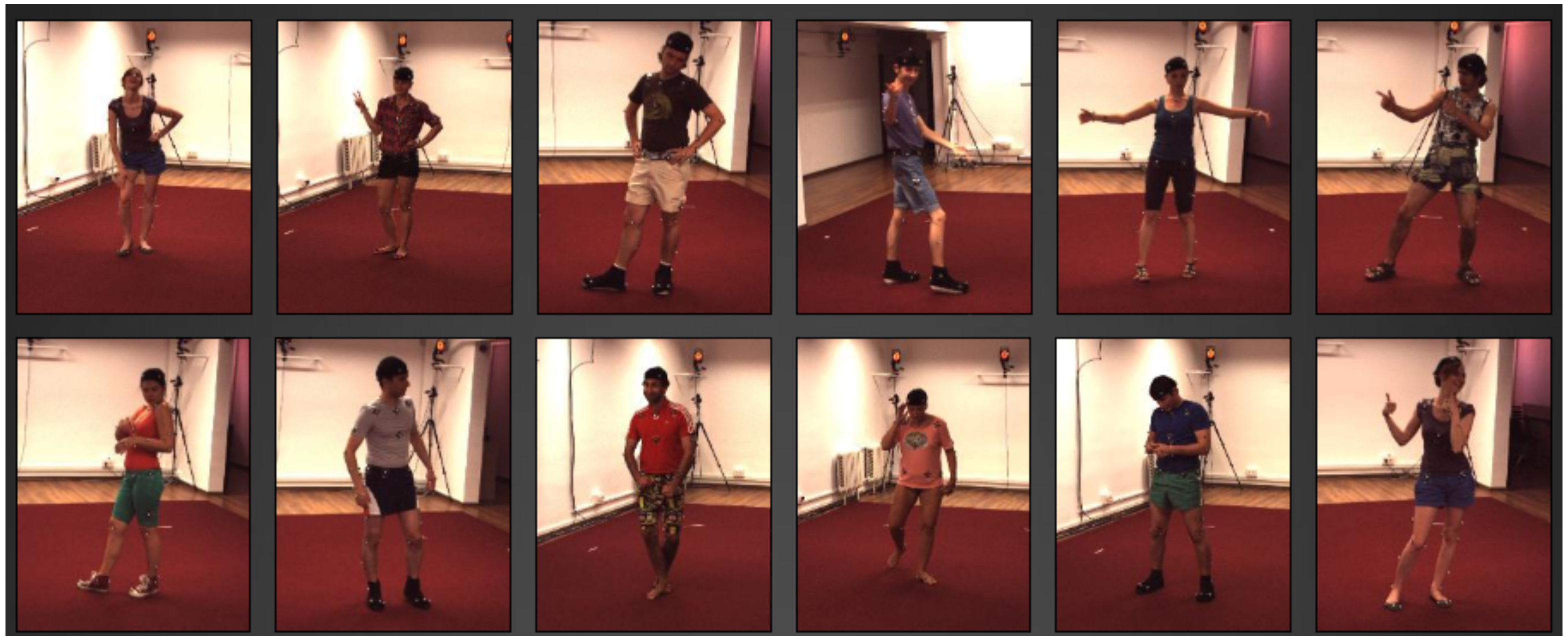
# QUALITATIVE RESULTS



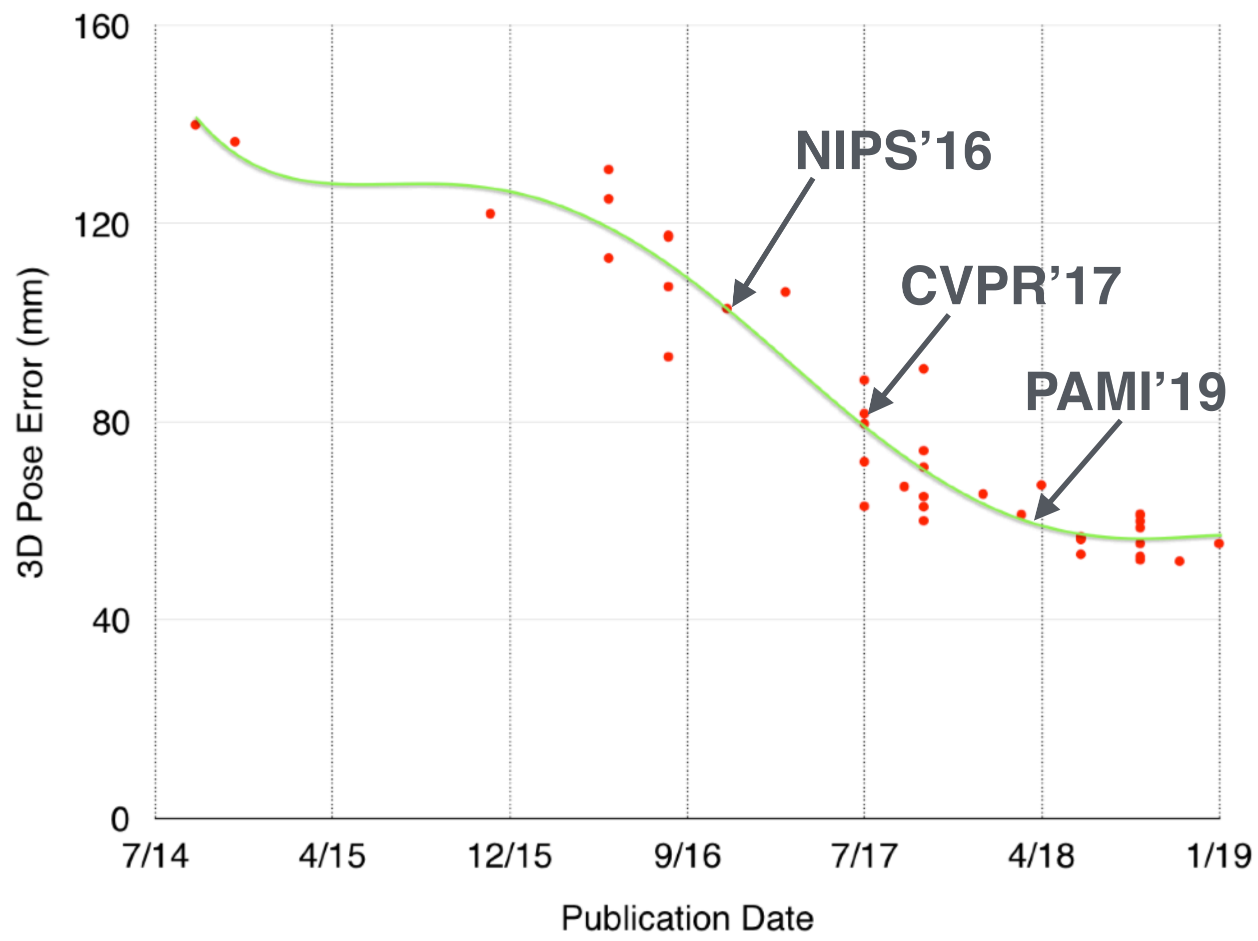


# COMPARISON WITH STATE-OF-ART

**3D Evaluation:** Human3.6M dataset



- 300k training images with perfect bounding boxes, 2D and 3D poses
- 5 subjects for training
- 2 subjects for test




Performance on H3.6M is saturating.

# EVALUATION IN THE WILD


**2D Evaluation:** MPII dataset

- 17k images with ~25k annotated 2D poses
- validation set of 1000 images

**Problems**:

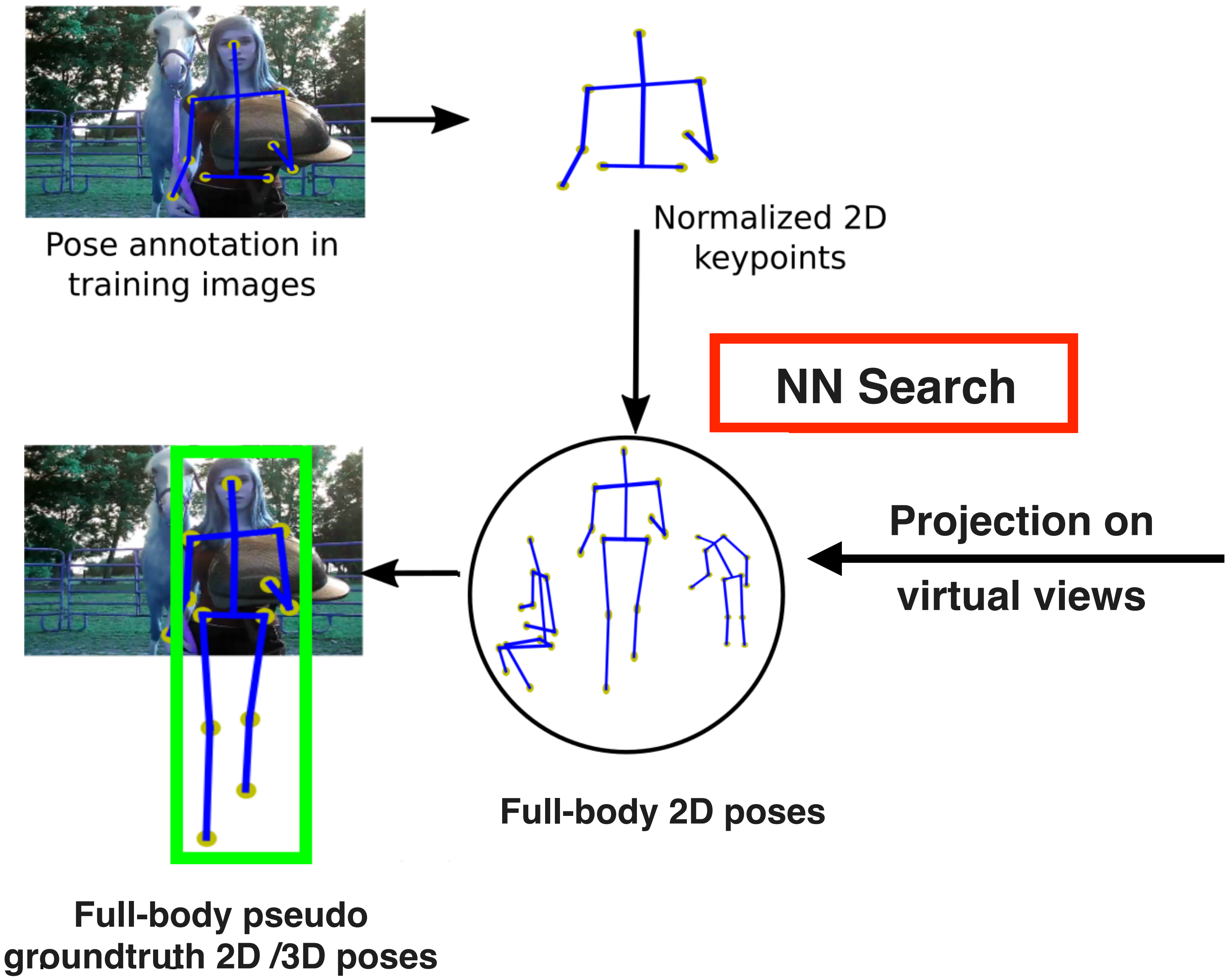
1- Many **occlusions** by objects/persons -> Mosaic data not adapted for these cases

2- People often **truncated** at image boundary -> full-body not in the image

# EVALUATION IN THE WILD

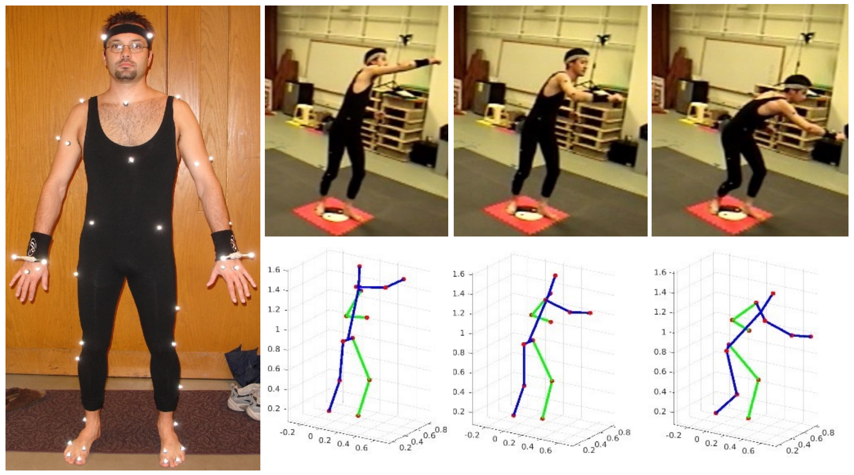
**Solution 1: Annotate images with 3D "pseudo ground truth":**

1. collect large MoCap dataset (merging 10 different datasets)
2. generate large set (8M) of 2D poses (varying camera viewpoint)
3. find NN 2D pose using labelled / visible 2D joints




**2D annotations:** LSPE (11k), MPII (17k), Coco (35k), H3.6M (17k)
**Total training set:** 160k (mirroring)

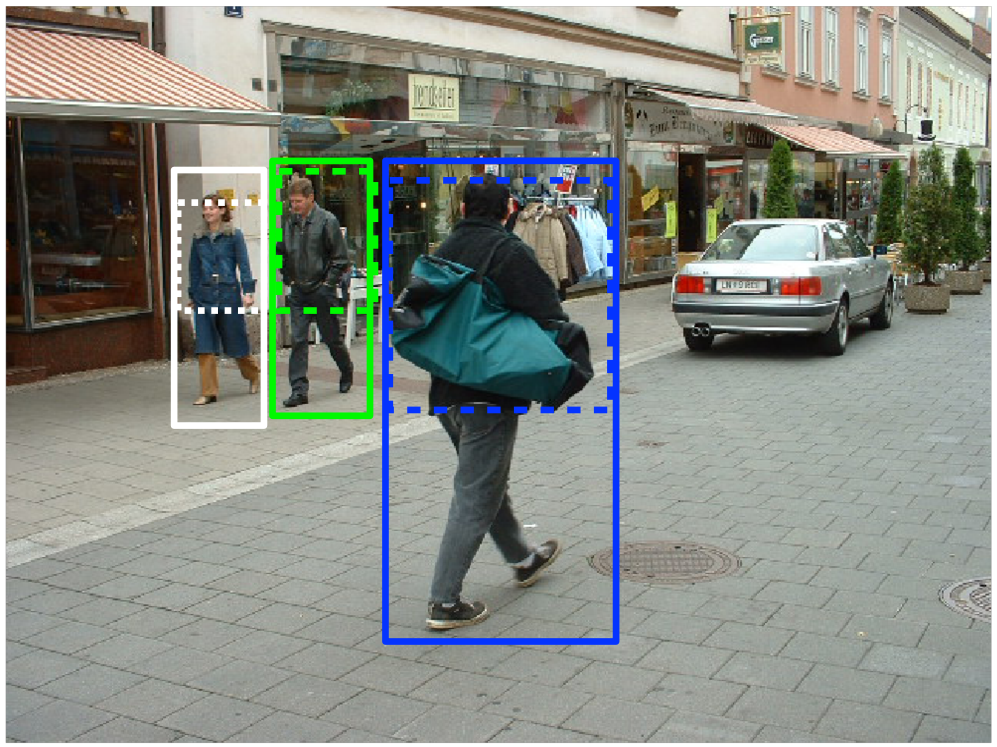
**MPI Pose Prior dataset**



**CMU MoCap dataset**

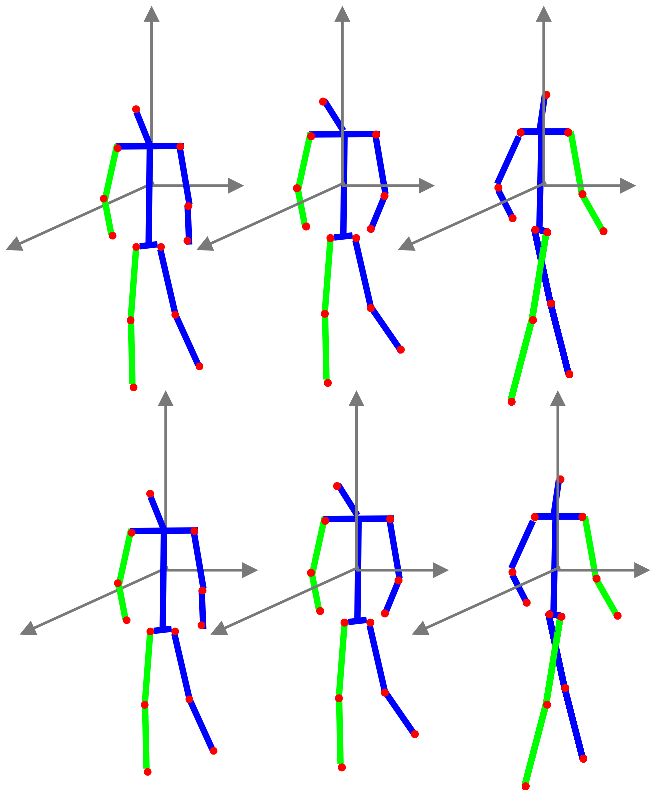
# LCR-NET: TRAINING (IN THE WILD)

Solution 2: Create "upper-body classes":



Bounding box + class label

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



normalized 3D poses
(aligned+orientated)

**Loss:**

$$\mathcal{L} = \mathcal{L}_{Loc} + \mathcal{L}_{Classif} + \mathcal{L}_{Reg}$$

RPN loss (cf FasterRCNN)

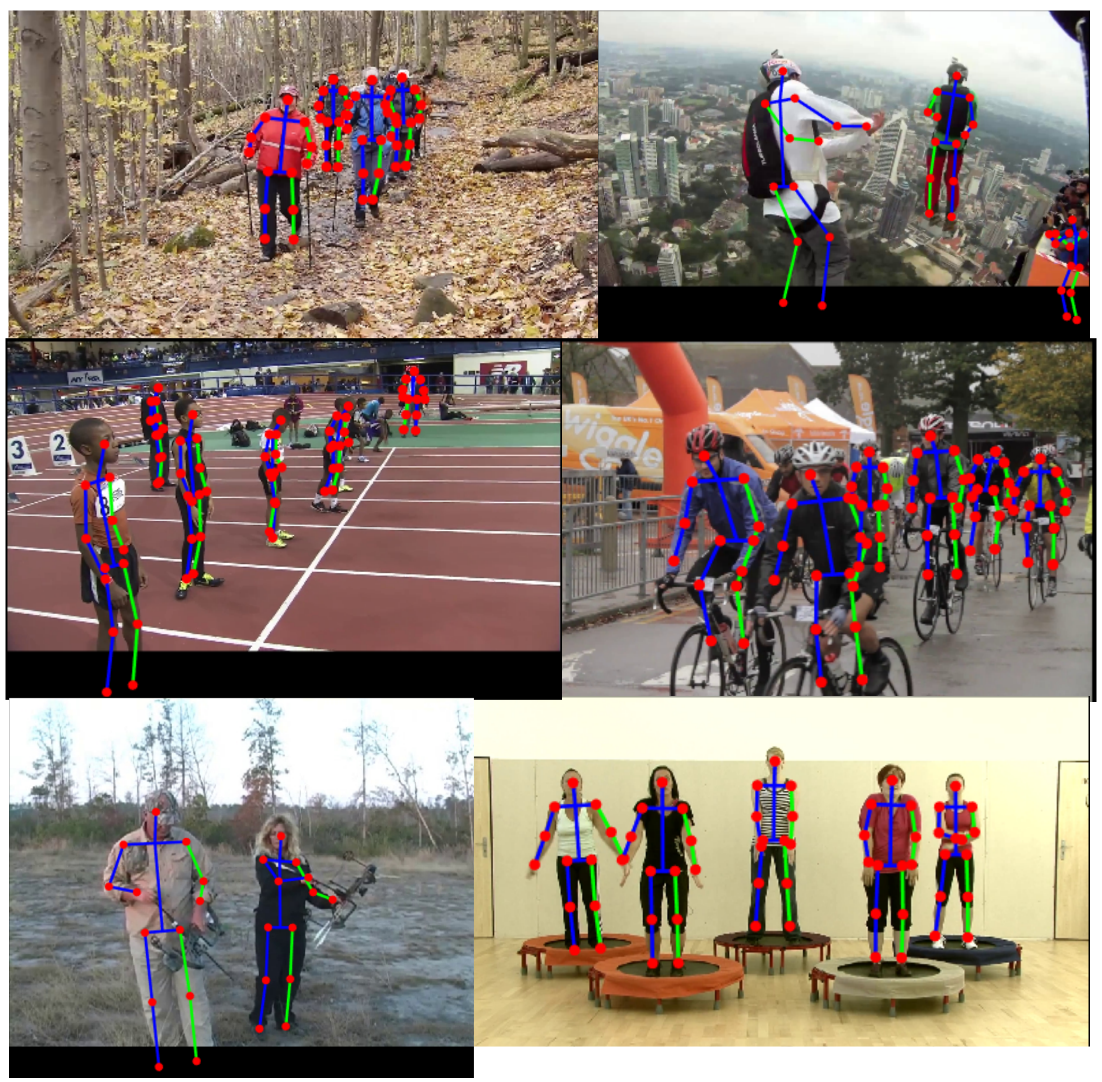
log loss of the true class

L1-smooth loss

# RESULTS IN THE WILD

**We were the first to evaluate in 4 regimes:**

- state-of-the-art results in 3D pose estimation (H3.6M)
- near state-of-the-art results for in-the-wild 2D pose estimation (MPII-single)
- competitive results in multi-person detection and 2D pose (MPII-multi)
- state-of-the-art results in multi-person 3D pose estimation (MuPoTS)

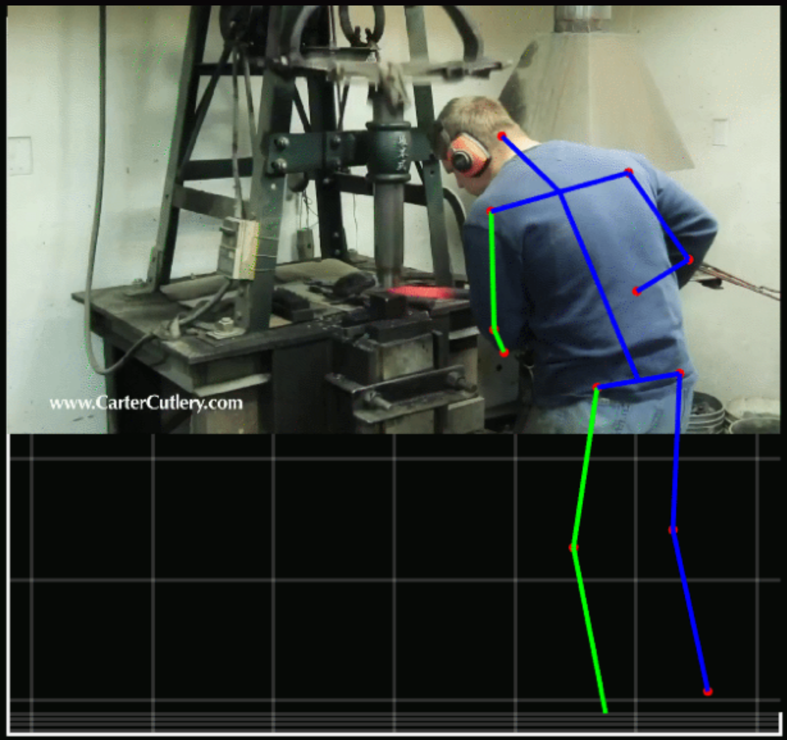


[Rogez, Weinzaepfel & Schmid, LCR-Net++. IEEE T. PAMI 2019]

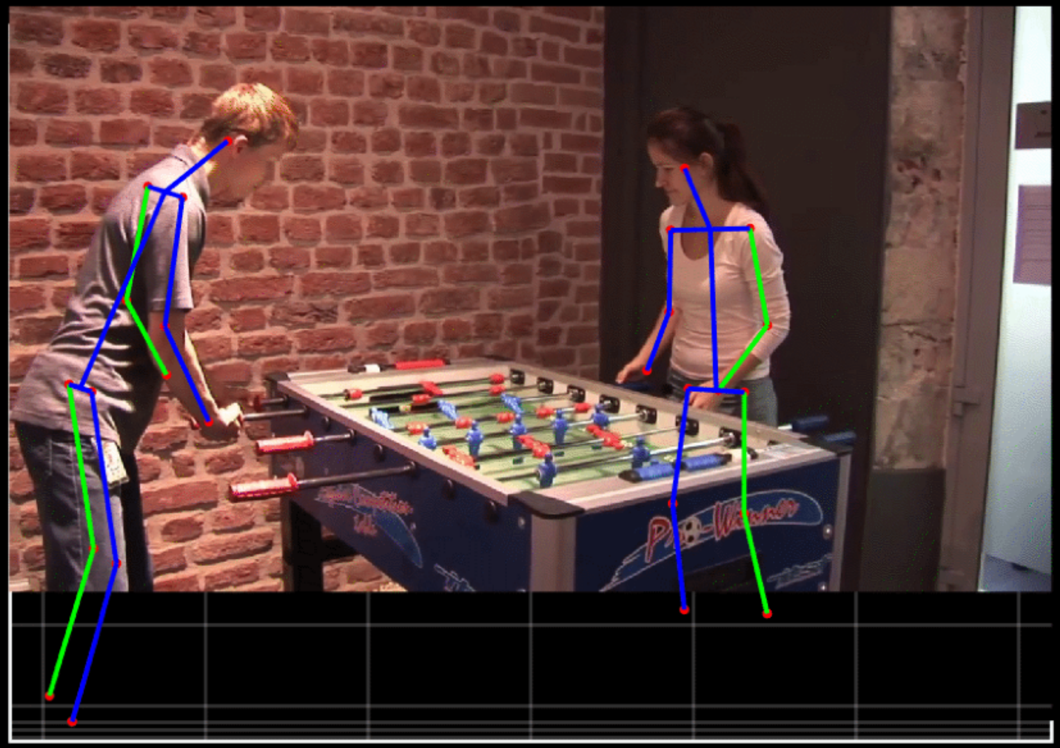
# LCR-NET CAN HANDLE VARIED POSES…

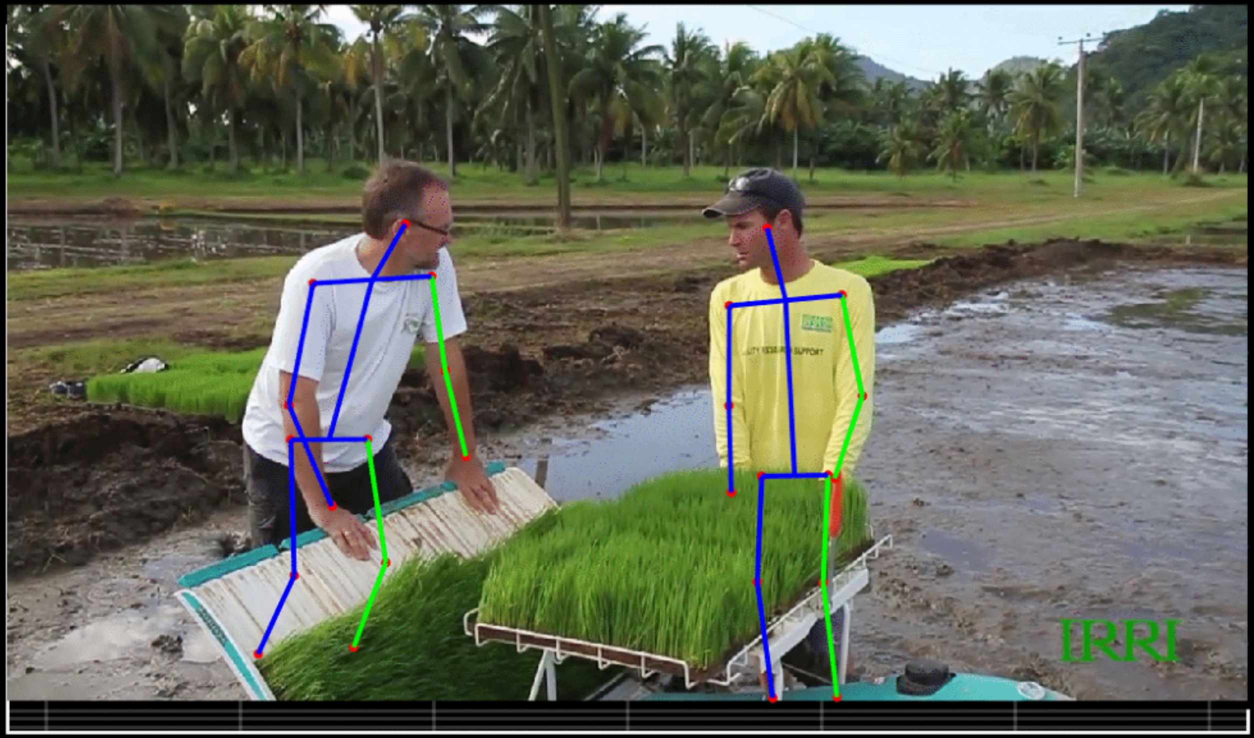
# SELF-OCCLUSIONS,

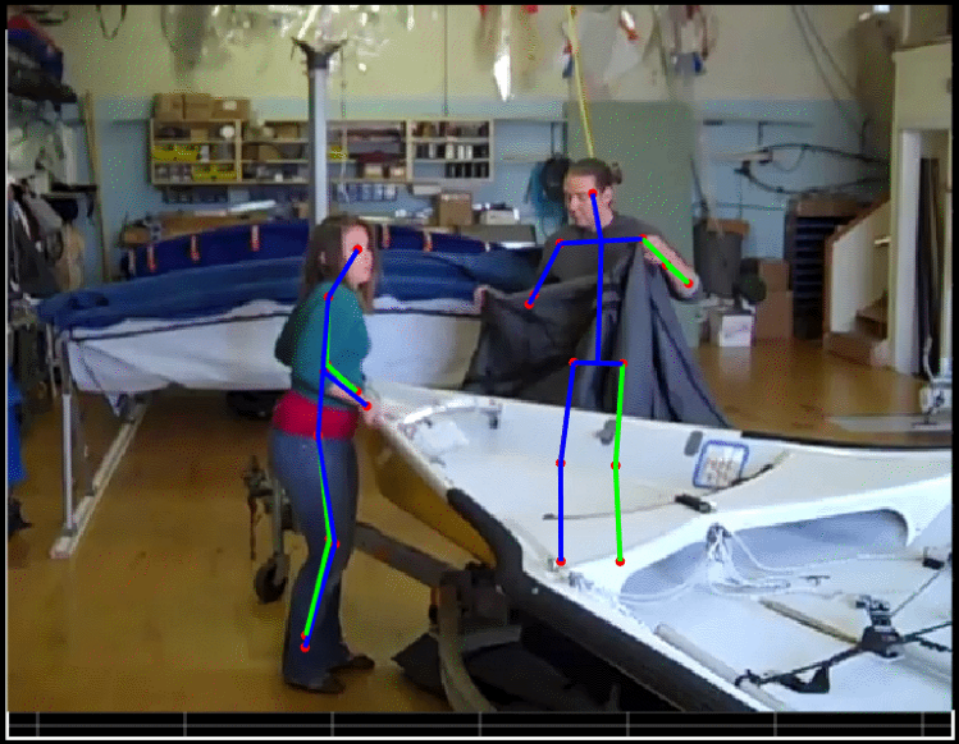


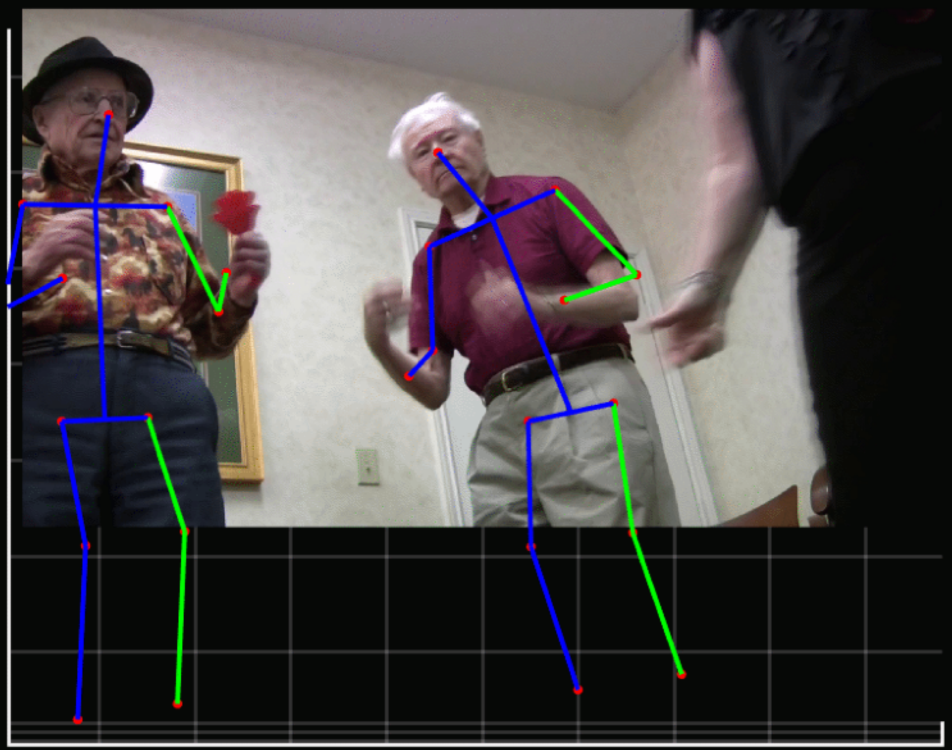
# OCCLUSIONS,

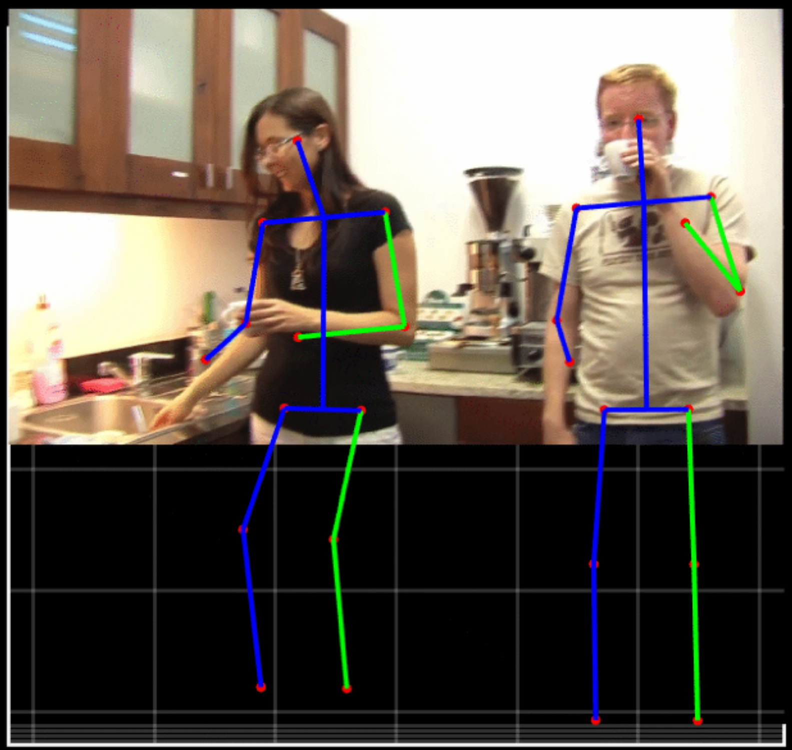


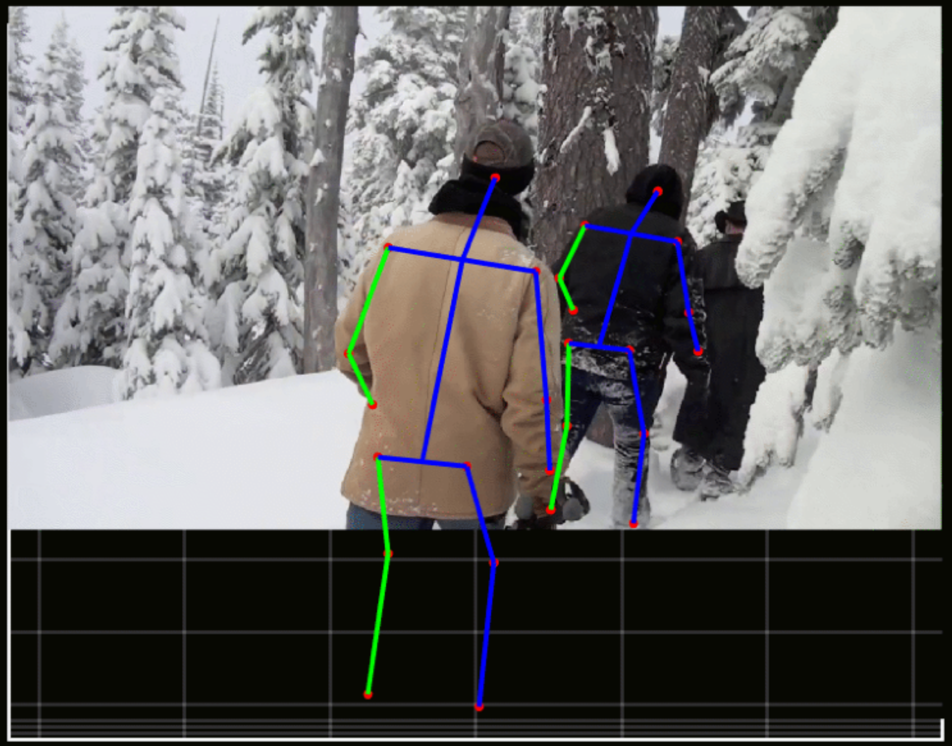


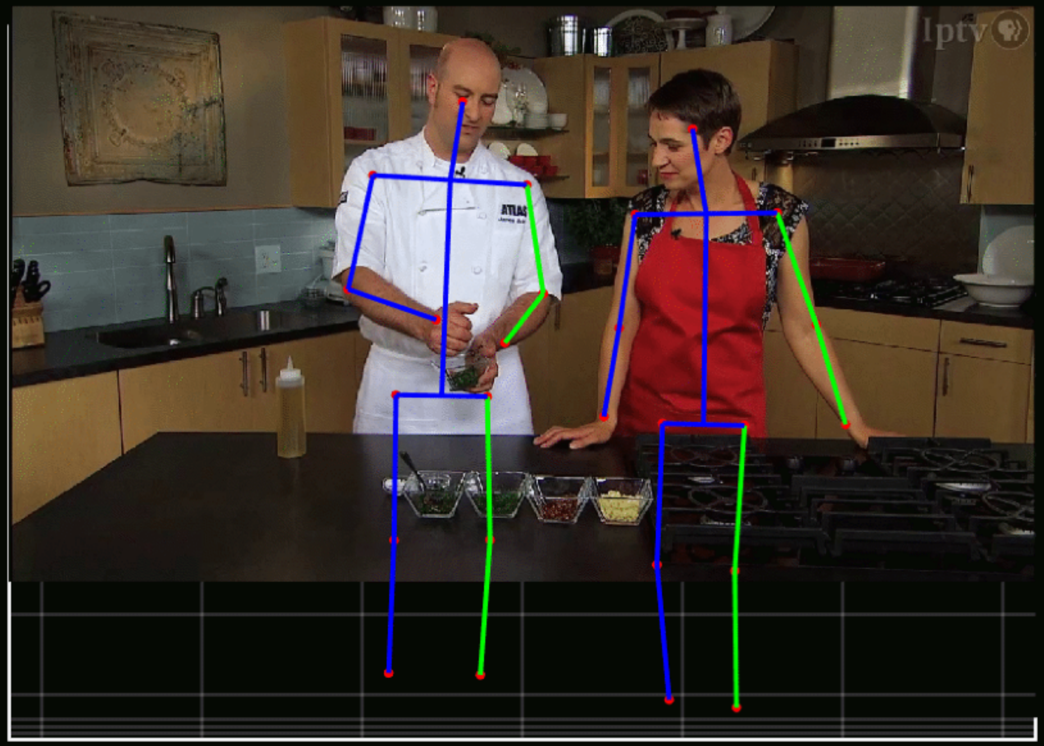



# AND TRUNCATIONS.







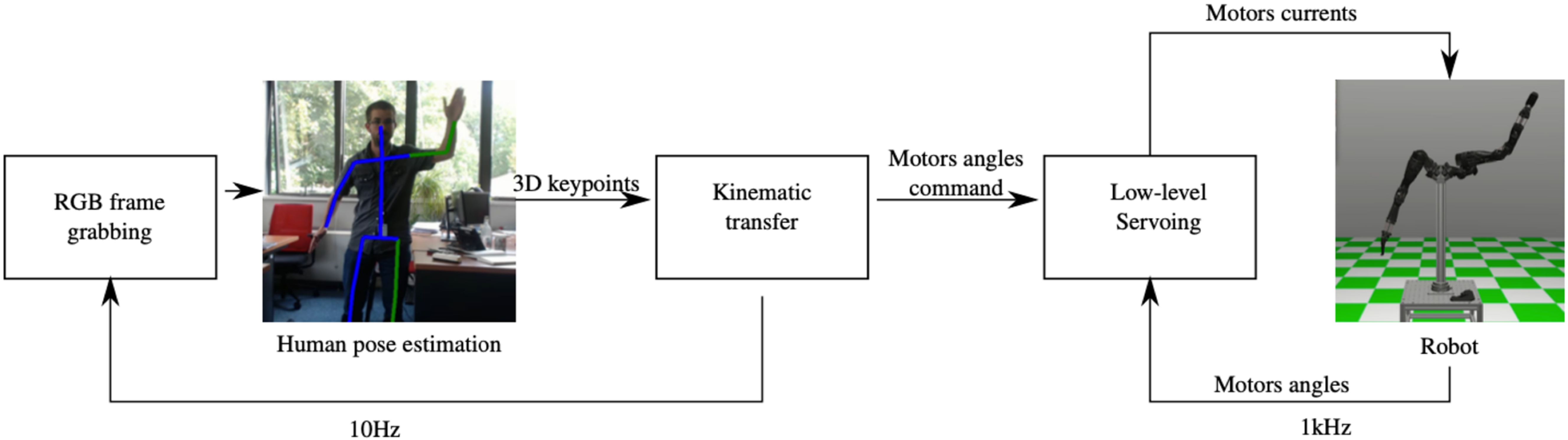


# REAL-TIME DEMO


# REAL-TIME DEMO ON PHONE

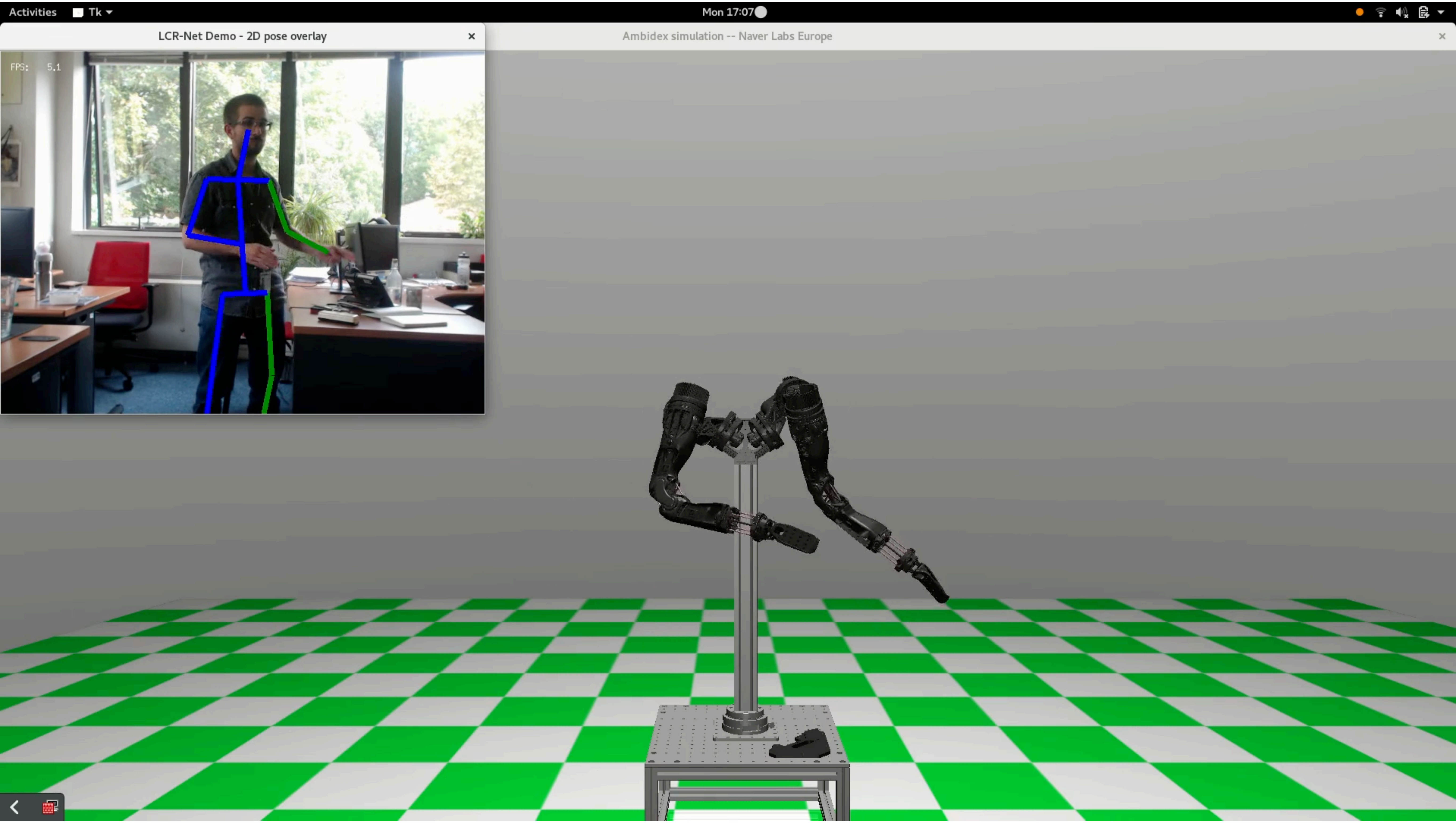
**LCR-Net** on iPhone 11

# REAL-TIME POSE RETARGETING DEMO

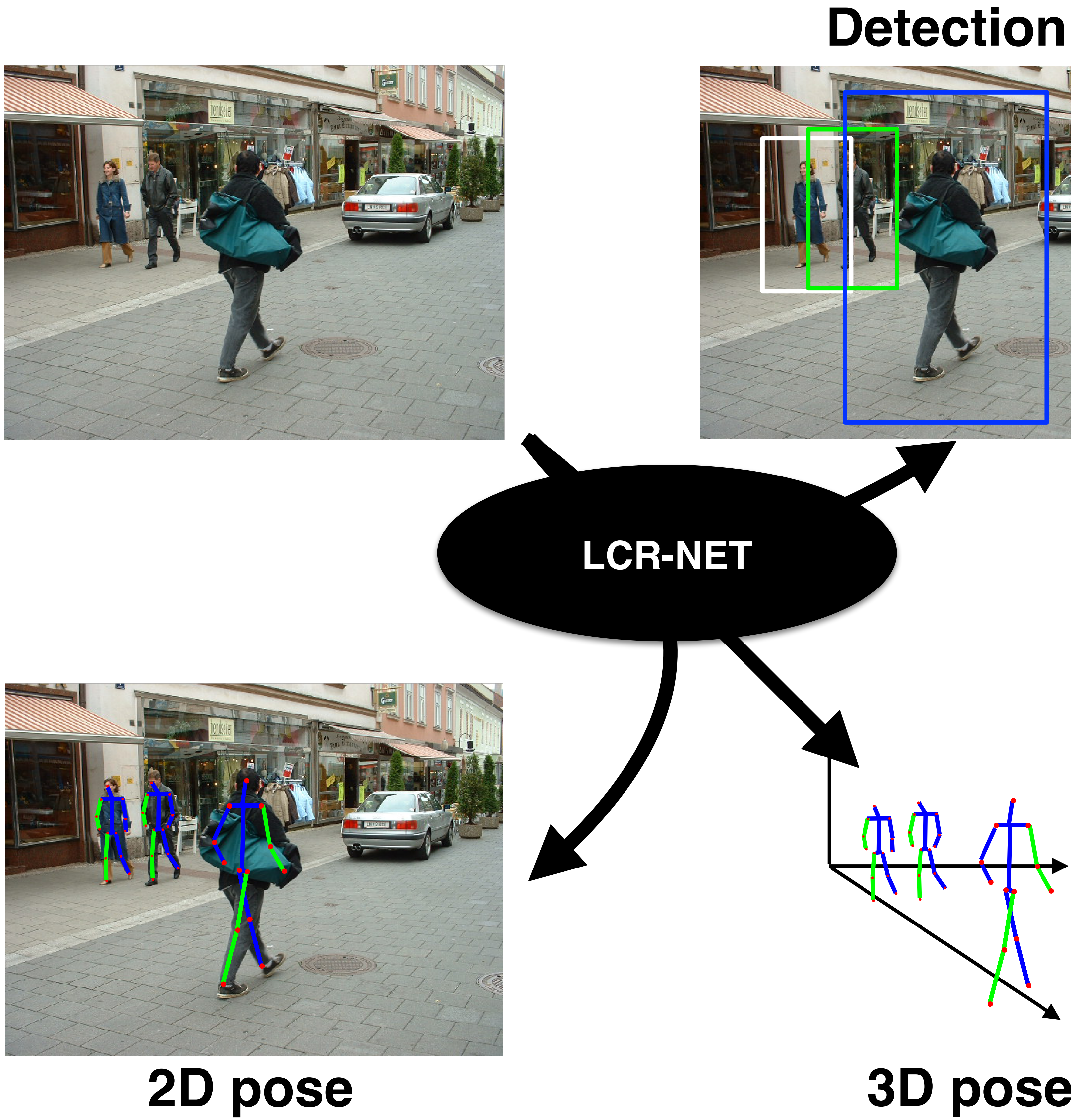



Animation of NAVER LABS robot **Ambidex** in MuJoCo simulator

# REAL-TIME POSE RETARGETING DEMO




Animation of NAVER LABS robot **Ambidex** in MuJoCo simulator



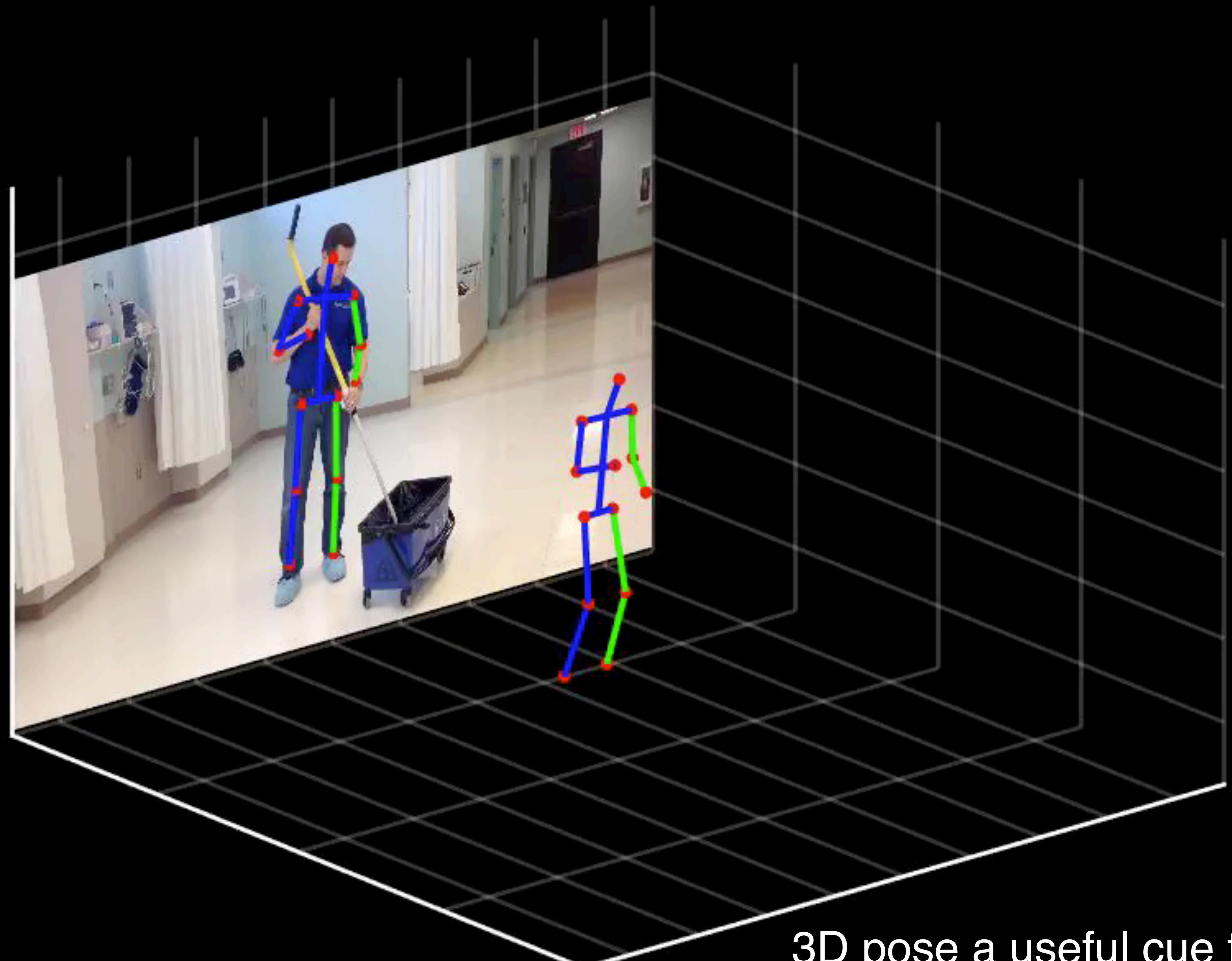

**LCR-Net** with **class-specific** regression**:**

- **reduced nb of classes** (computation)
- **refine** the 2D/3D pose
- real-time pose detection.

# OUTLINE

- Background
- Monocular 3D Human pose estimation
- Classification-based approaches
- Drawbacks and solutions
- **and beyond…**

# LCR-NET QUALITATIVE RESULTS IN VIDEO SEQUENCE



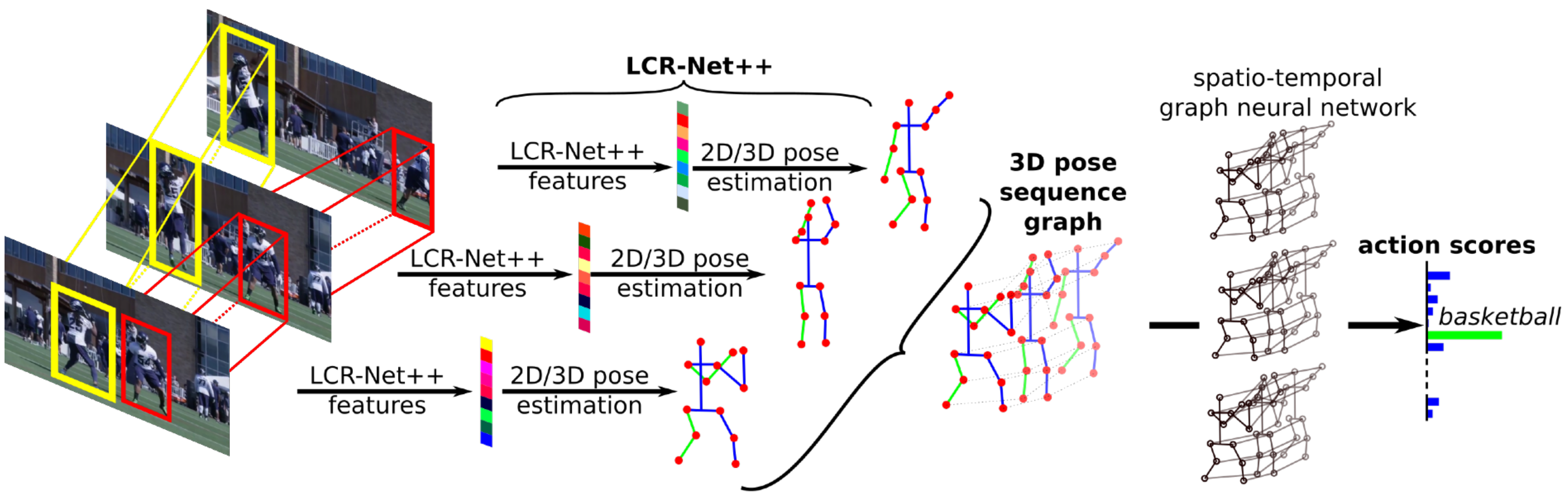
3D pose a useful cue for action recognition, especially when there's no context, e.g. mimes…

NB: Frames processed independently.

# OUT-OF-CONTEXT ACTION RECOGNITION

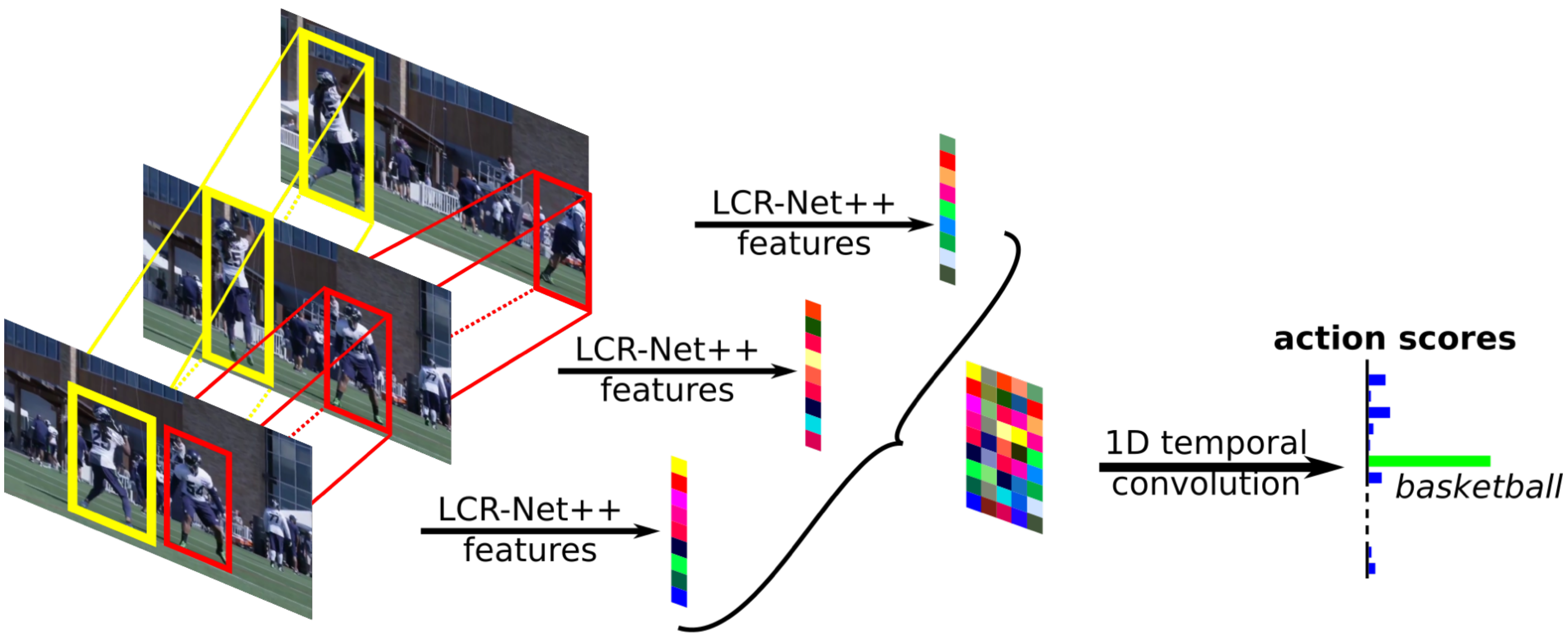
# POSE-BASED BASELINE: EXPLICIT 2D OR 3D POSES

STGCN3D / STGCN2D




[Weinzaepfel & Rogez, Mimetics: Towards understanding human actions out of context, IJCV 2021]

## SIP-Net




[Weinzaepfel & Rogez, Mimetics: Towards understanding human actions out of context, IJCV 2021]

# RESULTS ON EXISTING BENCHMARKS

| | JHMDB-1 | JHMDB | PennAction | NTU (cs) | HMB51-1 | HMDB51 | UCF101-1 | UCF101 | Kinetics |
|---|---|---|---|---|---|---|---|---|---|
| PoTion [7] | 59.1 | 57.0 | - | - | 46.3 | 43.7 | 60.5 | 65.2 | 16.6 |
| Zolfaghari *et al.* [54] (pose only) | 45.5 | - | - | 67.8 | 36.0 | - | 56.9 | - | - |
| MultiTask [28] (uses RGB) | - | - | **97.4** | 74.3 | - | - | - | - | - |
| STGCN [48] (OpenPose) | 25.2 | 25.4 | 71.6 | **79.8** | 38.6 | 34.7 | 54.0 | 50.6 | 30.7 |
| STGCN2D | 23.2 | 23.2 | 85.5 | 69.4 | 36.5 | 32.7 | 49.2 | 44.4 | 11.9 |
| STGCN3D | 53.1 | 50.5 | 89.2 | 75.0 | 39.8 | 41.0 | 48.5 | 51.1 | 10.6 |
| SIP-Net | **66.4** | **62.4** | 93.5 | 64.8 | **50.7** | **51.2** | **66.1** | **66.0** | **32.8** |

- explicit 2D and even more 3D poses suffer from noise in-the-wild
- implicit poses are more robust
- lack details on hands/fingers/faces + fine-grained classes

[Weinzaepfel & Rogez, Mimetics: Towards understanding human actions out of context, IJCV 2021]

# RESULTS ON MIMETICS

- Flow is less biased than RGB
- Implicit poses perform better

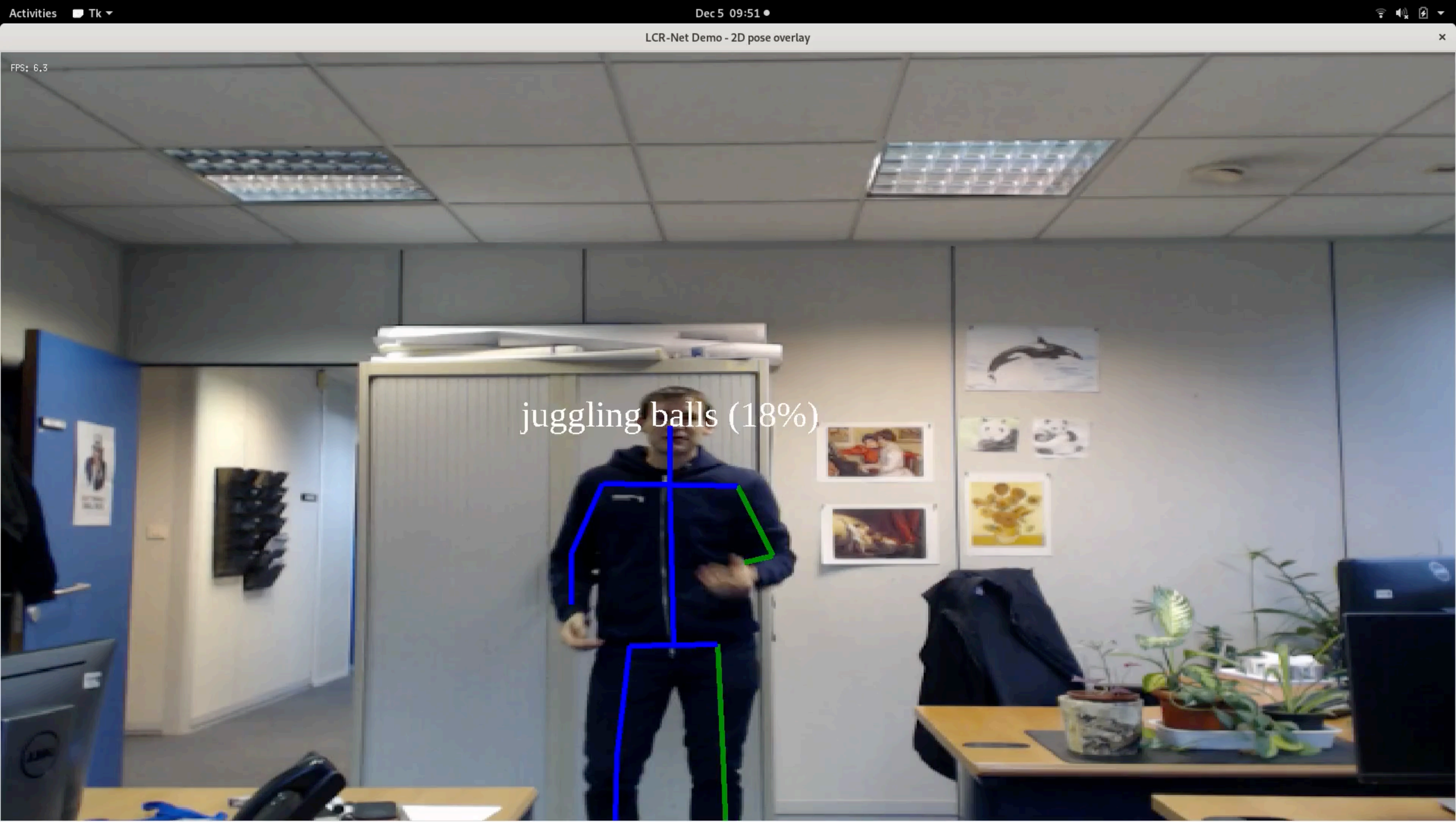
| | top-1 | top-5 | mAP |
|---|---|---|---|
| RGB (3D-ResNeXt-101) | 8.6 | 20.1 | 15.6 |
| Flow (3D-ResNeXt-101) | 11.8 | 29.6 | 21.1 |
| RGB+Flow (late fusion) | 10.5 | 26.9 | 19.1 |
| STGCN [48] (OpenPose) | 12.6 | 27.4 | 20.7 |
| STGCN2D | 9.0 | 20.5 | 15.4 |
| STGCN3D | 5.8 | 13.8 | 11.3 |
| SIP-Net | **14.2** | **32.0** | **22.7** |

SIP-Net: playing piano
RGB: massage back
Flow: playing piano

SIP-Net: shooting goal
RGB: playing badminton
Flow: dancing ballet

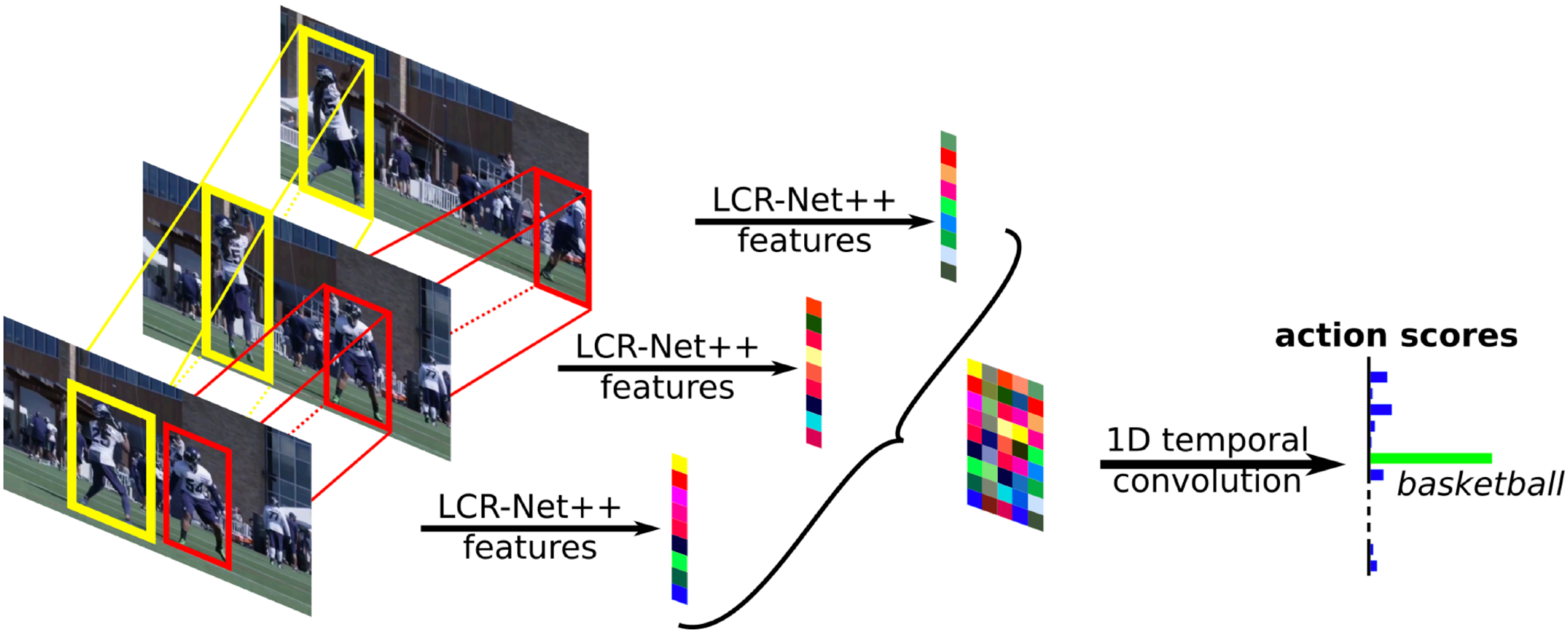
[Weinzaepfel & Rogez, Mimetics: Towards understanding human actions out of context, IJCV 2021]

# MIMED ACTION RECOGNITION DEMO




Real-time multi-person mimed action recognition demo

# TAKE HOME MESSAGE




- State-of-the-art action recognition methods are biased towards context (scene and objects).
- The implicit pose features of **LCR-Net** can be used for **out of-context** action recognition.
- More details are required on **hands and faces.**
- **3D poses are too noisy** to be used for action recognition as-is.

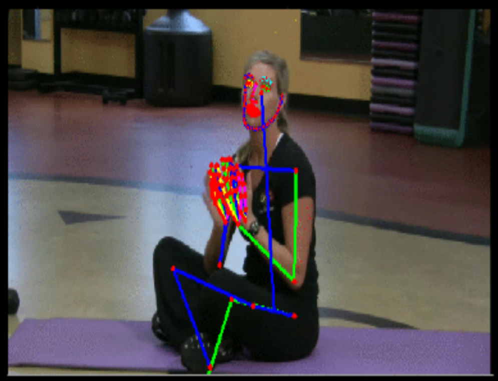
# BEYOND SIMPLE CLASSIFICATION

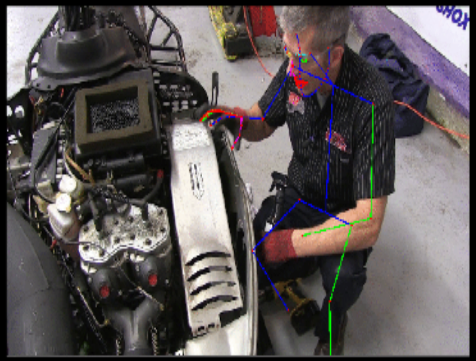
- Holistic approach requires annotations for all the body parts.

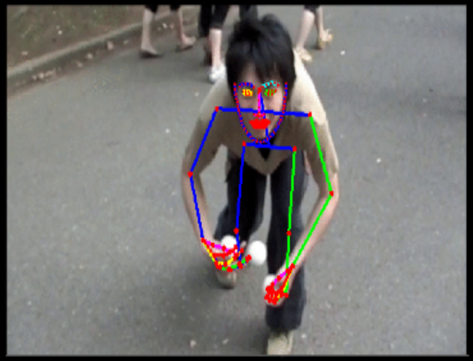
What if we want to classify poses including bodies, hands, faces?

- How to handle mis-detections and improve performances in videos (while keeping real-time performance).s

# WHOLE BODY POSE ESTIMATION

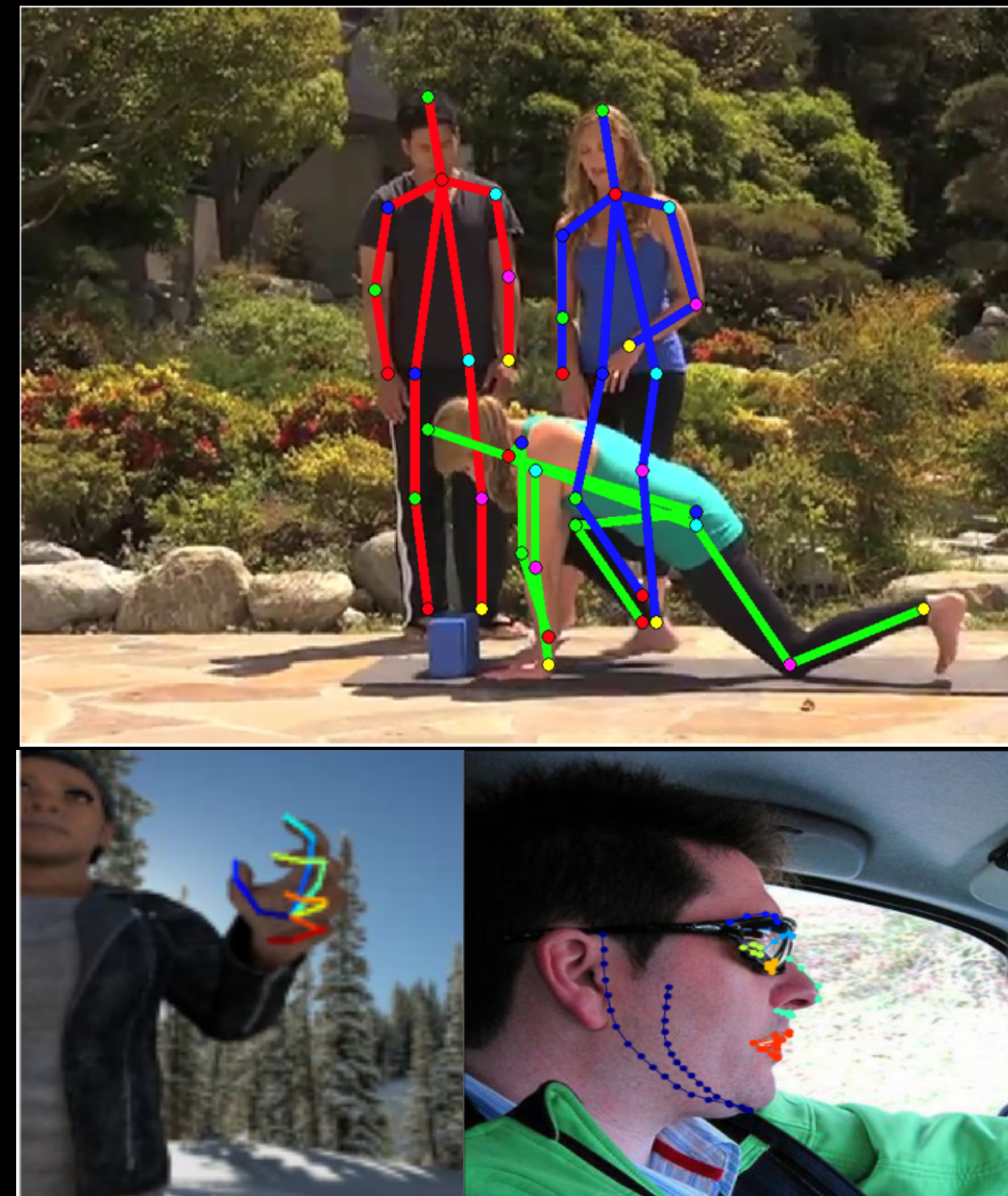






# NO IN-THE WILD WHOLE BODY POSE DATASET

Whole-body pose in controlled environment



Part-specific datasets in the wild

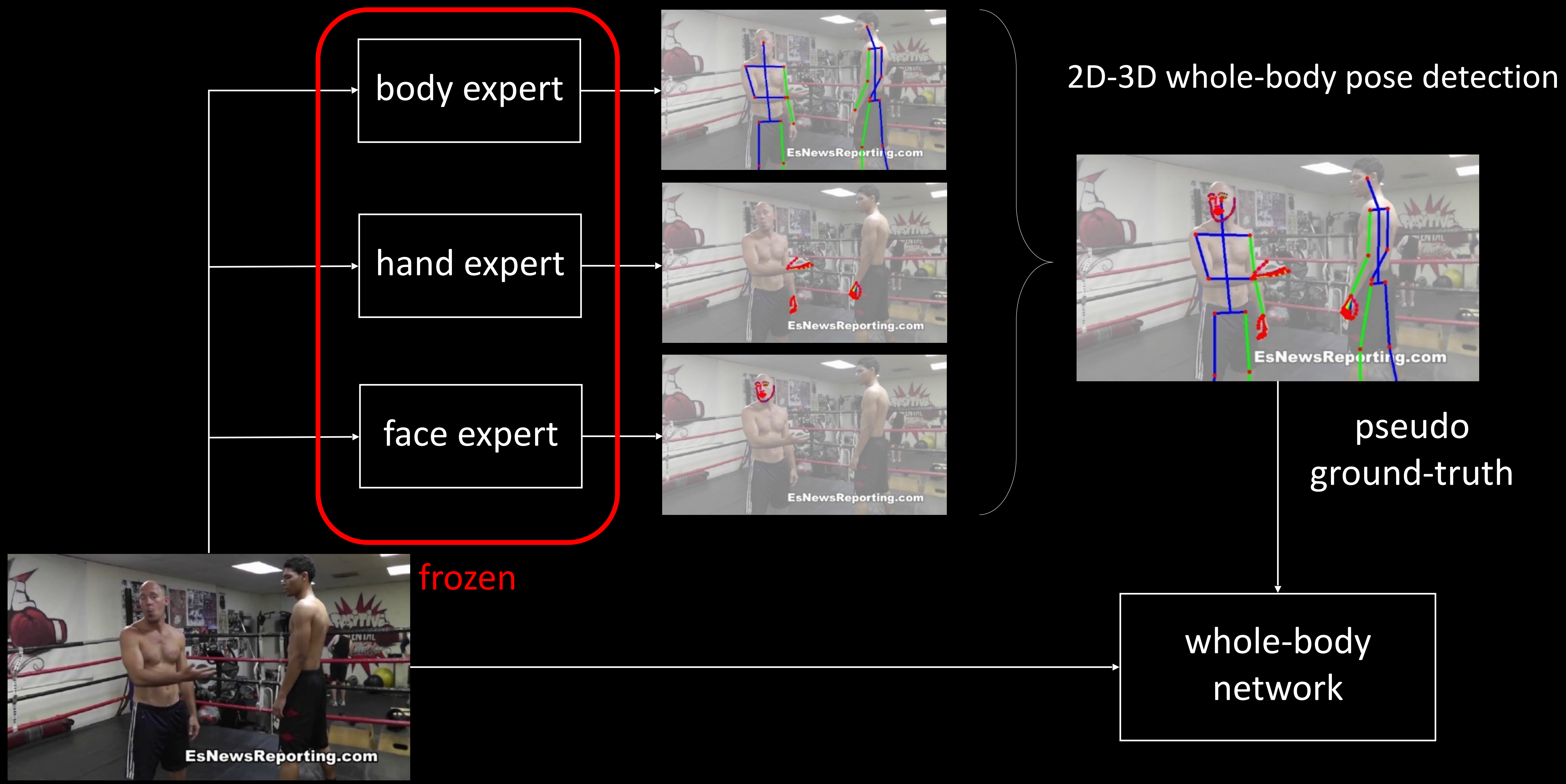
# WHOLE-BODY NETWORK

2D-3D
whole-body
ground-truth


Whole-body
network

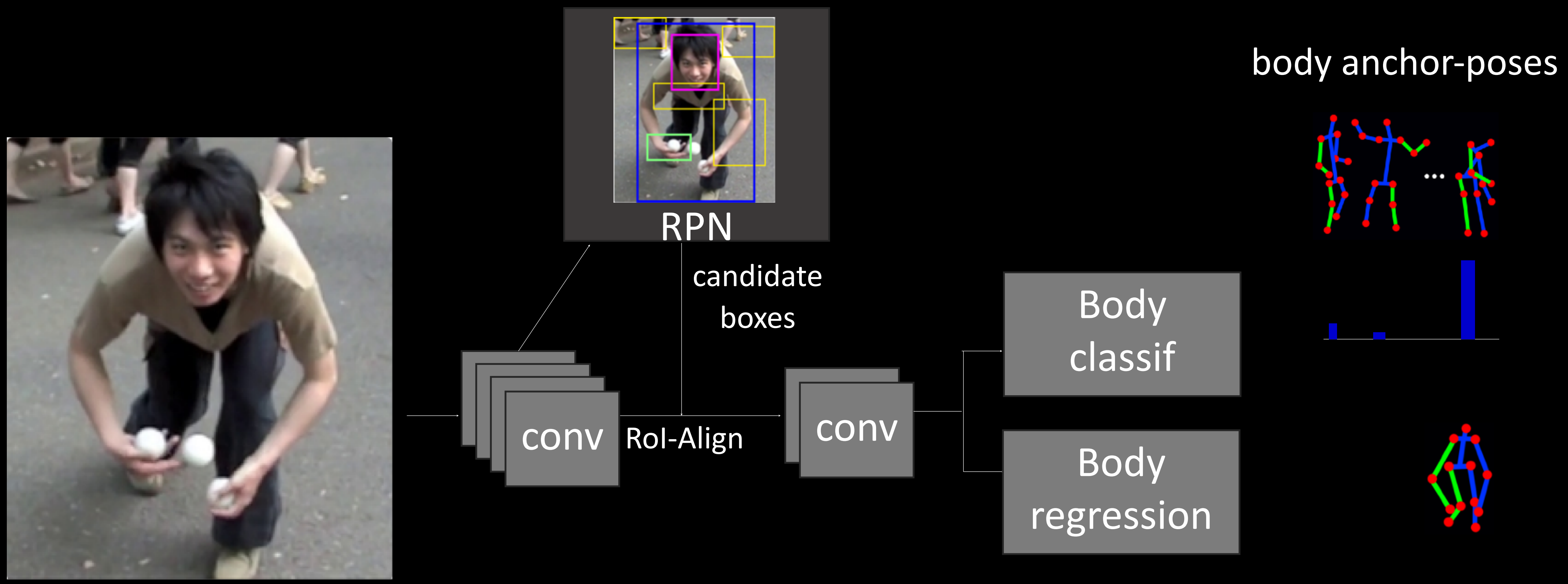
[Weinzaepfel, Bregier, Combaluzier, Leroy & Rogez, DOPE: Distillation of Part Experts for whole body pose estimation, ECCV 2020]

# DISTILLATION OF PART EXPERTS

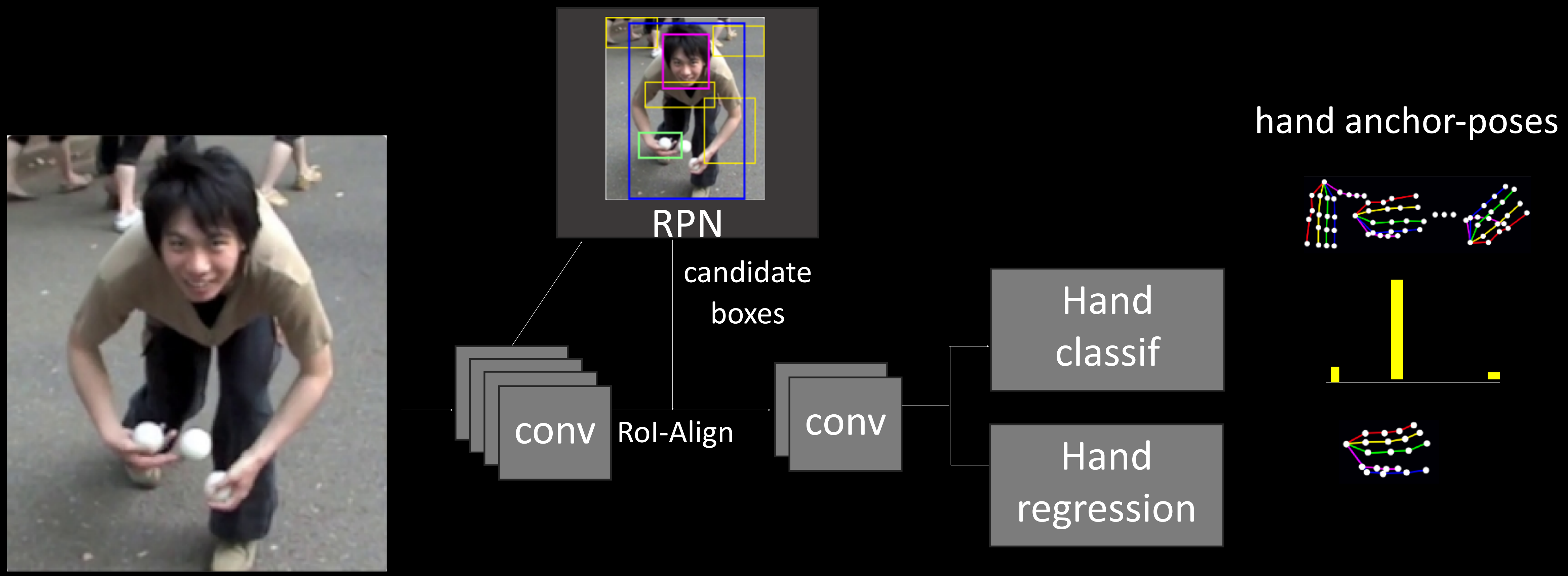



[Weinzaepfel, Bregier, Combaluzier, Leroy & Rogez, DOPE: Distillation of Part Experts for whole body pose estimation, ECCV 2020]

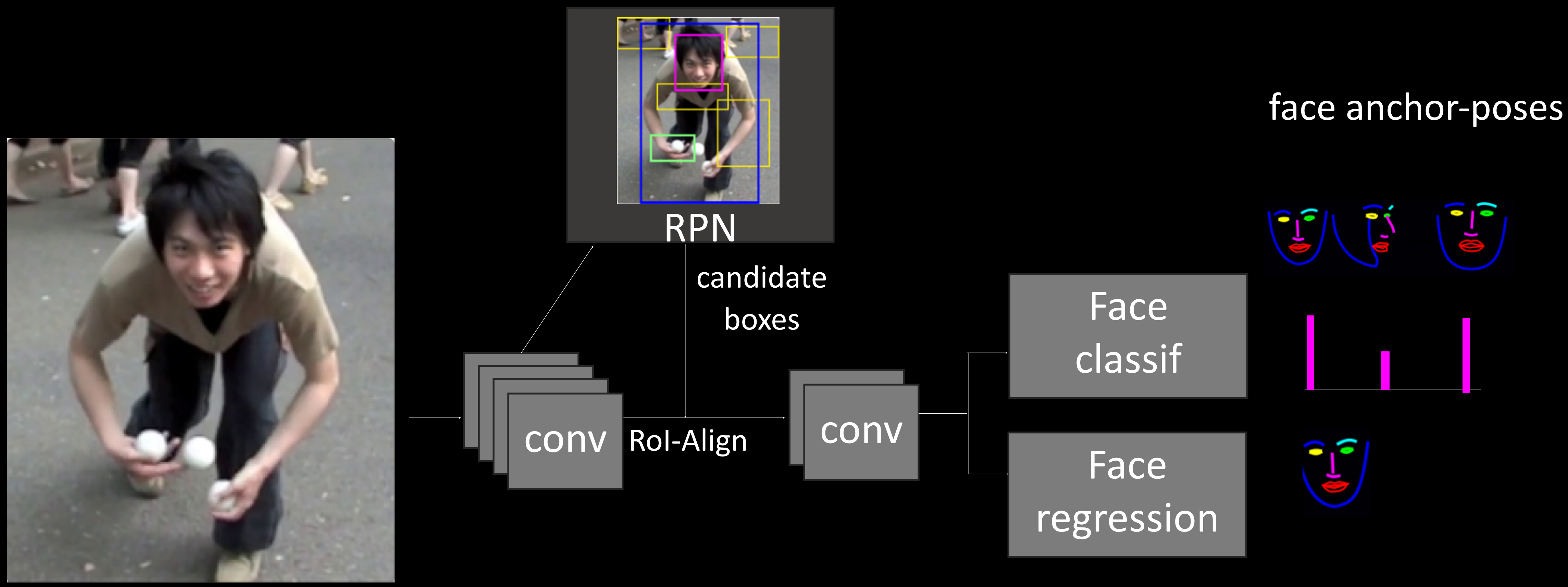
# DETECTION ARCHITECTURE



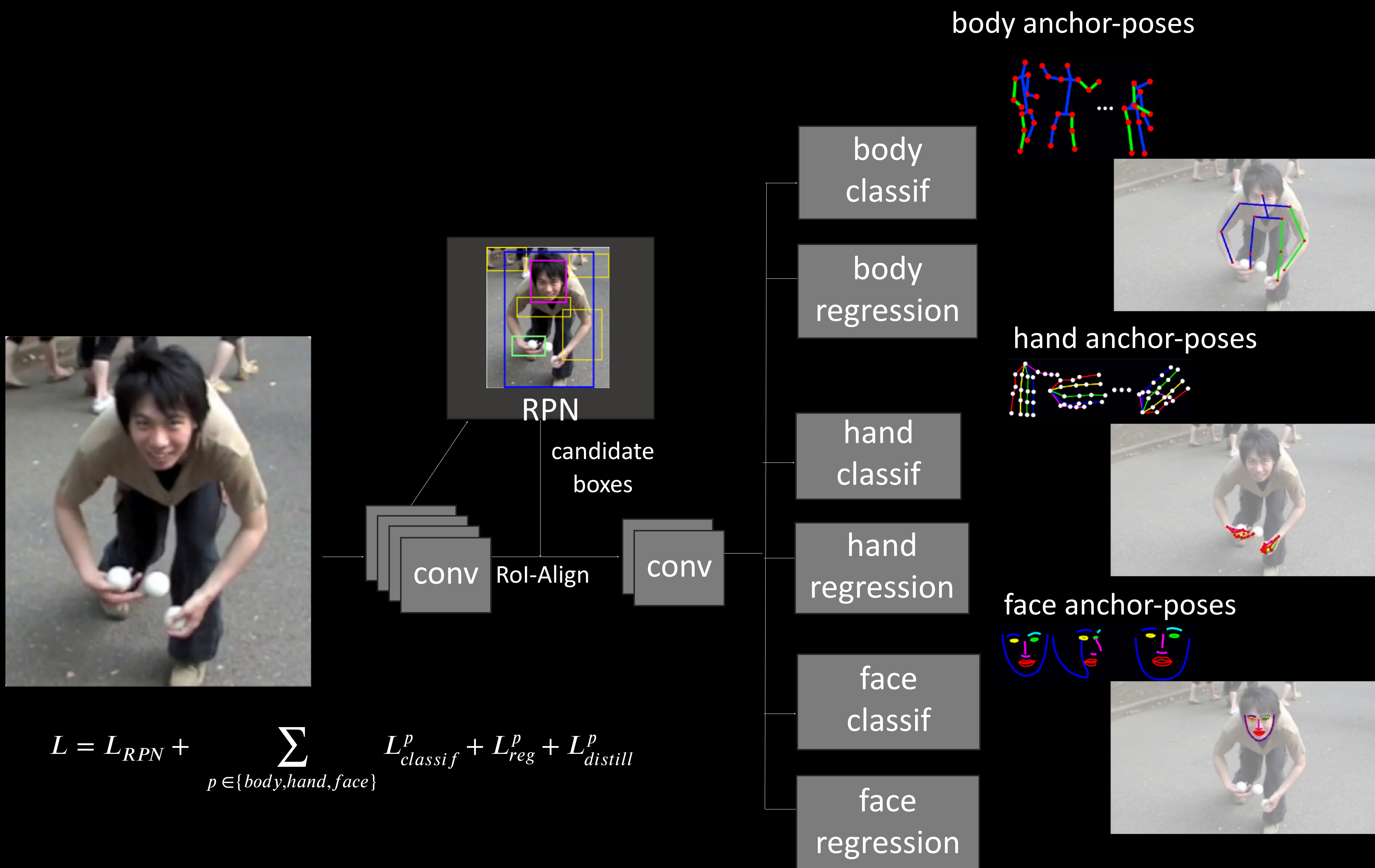
# DETECTION ARCHITECTURE



# DETECTION ARCHITECTURE



# DETECTION ARCHITECTURE



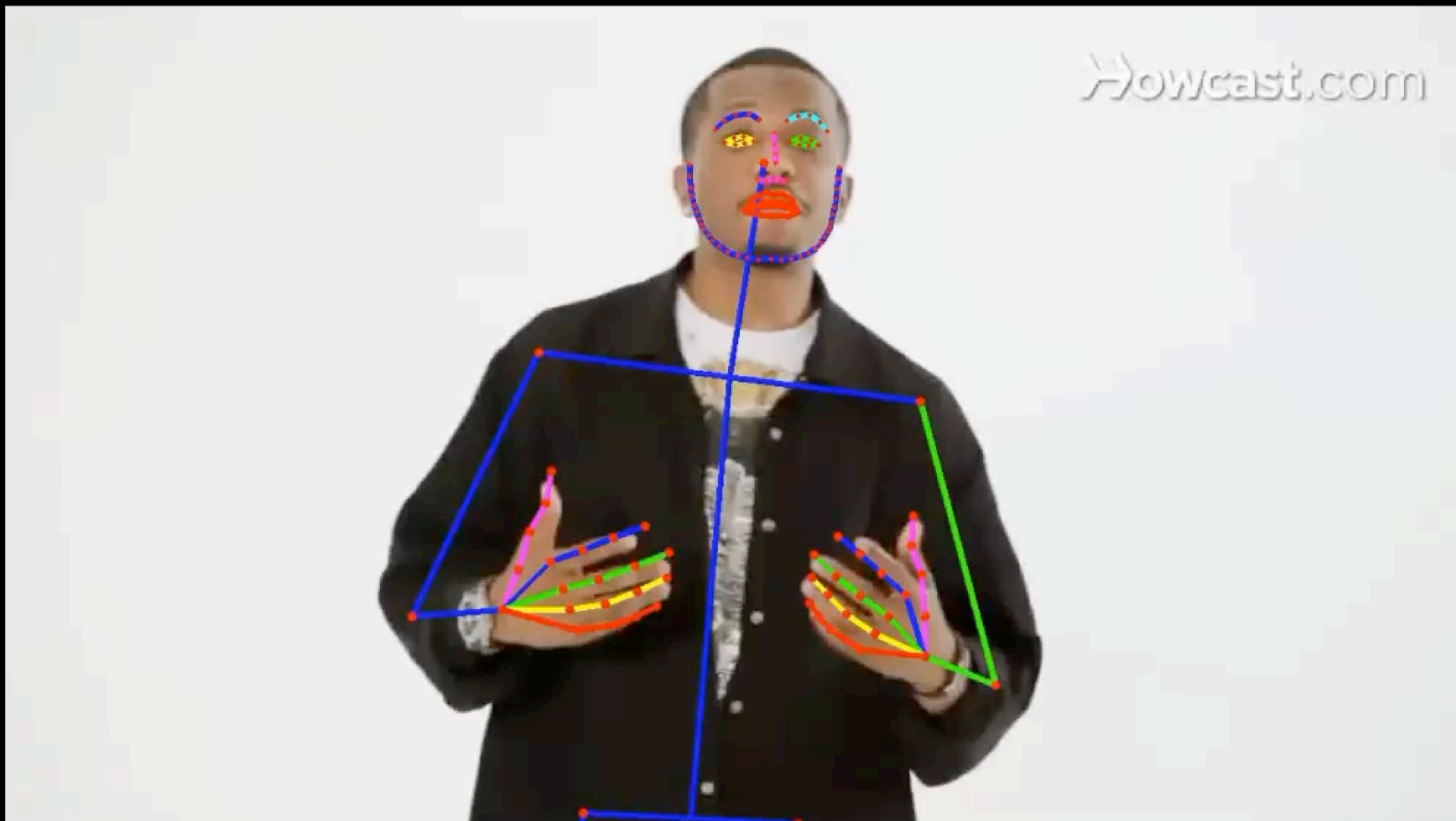

$$L = L_{RPN} + \sum_{p \in \{body, hand, face\}} L^{p}_{classif} + L^{p}_{reg} + L^{p}_{distill}$$

# EXPERIMENTAL RESULTS

| | 2D body pose (MPII, PCKh@0.5) | 3D body pose (MuPoTs, PCK3D) | 3D hand pose (RenderedHand, AUC) | face landmarks (Menpo, AUC) |
|---|---|---|---|---|
| Body expert | **89.6** | 66.8 | - | - |
| Hand expert | - | - | **87.1** | - |
| Face expert | - | - | - | 73.9 |
| Ignoring unannotated parts | 88.3 | 66.6 | 81.1 | 61.7 |
| **DOPE** | 88.8 | **67.2** | 84.9 | **75.0** |

[Weinzaepfel, Bregier, Combaluzier, Leroy & Rogez, DOPE: Distillation of Part Experts for whole body pose estimation, ECCV 2020]

DOPE: Distillation Of Part Experts for whole-body 3D pose estimation in the wild




Input: RGB image
(processed frame-by-frame)
Output: 2D-3D whole-body poses
(body, hands, face)

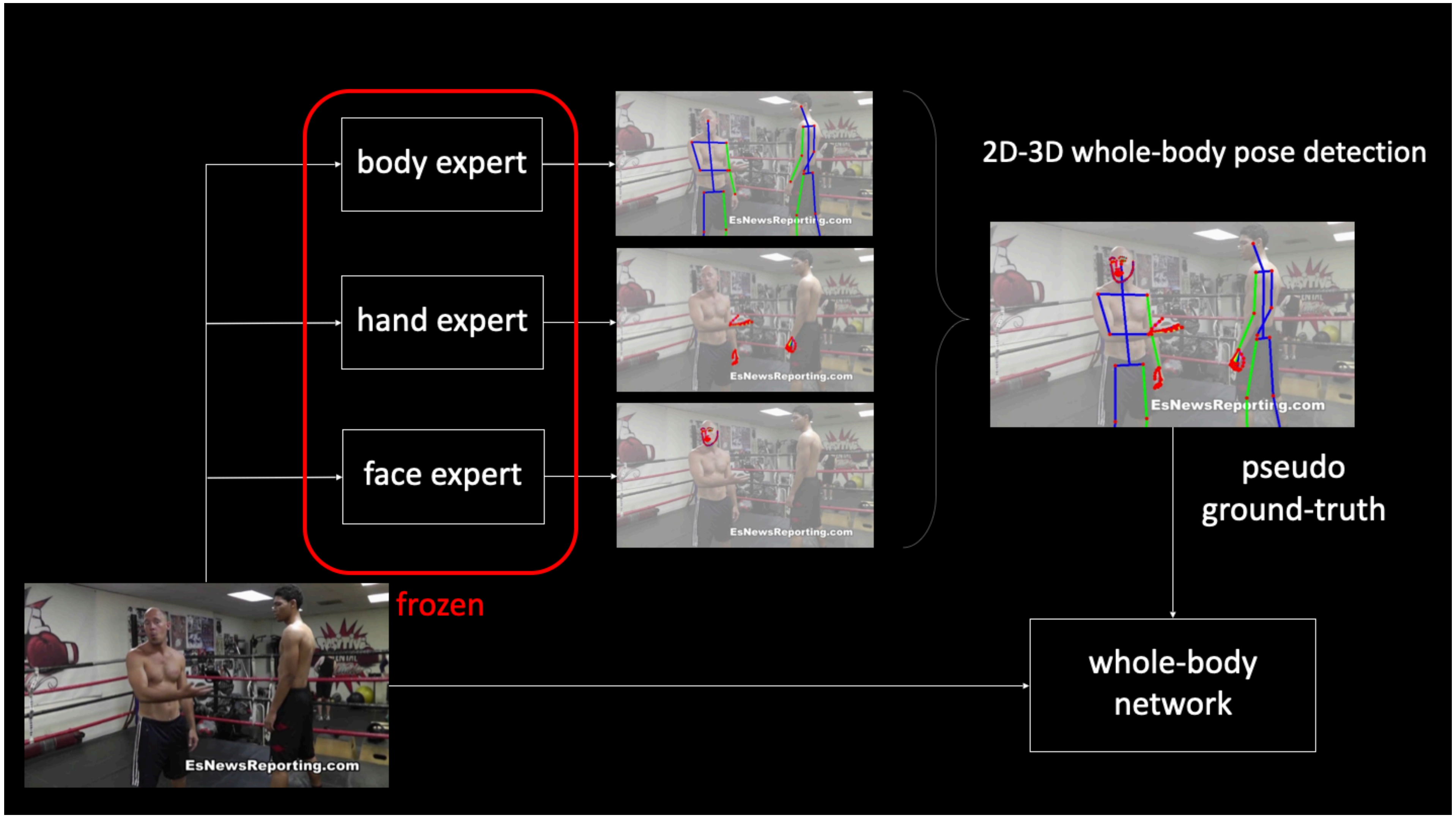
NAVER LABS Europe

[Weinzaepfel, Bregier, Combaluzier, Leroy & Rogez, DOPE: Distillation of Part Experts for whole body pose estimation, ECCV 2020]

# REAL-TIME DEMO

Frame-by-frame processing on a laptop with GTX 1080 (only 2D is shown for clarity)

# TAKE HOME MESSAGE




Lack of annotated data can be solved using **a teacher-student approach**

A **single DOPE model** can performs **multiple tasks** with a comparable network capacity as one single expert.

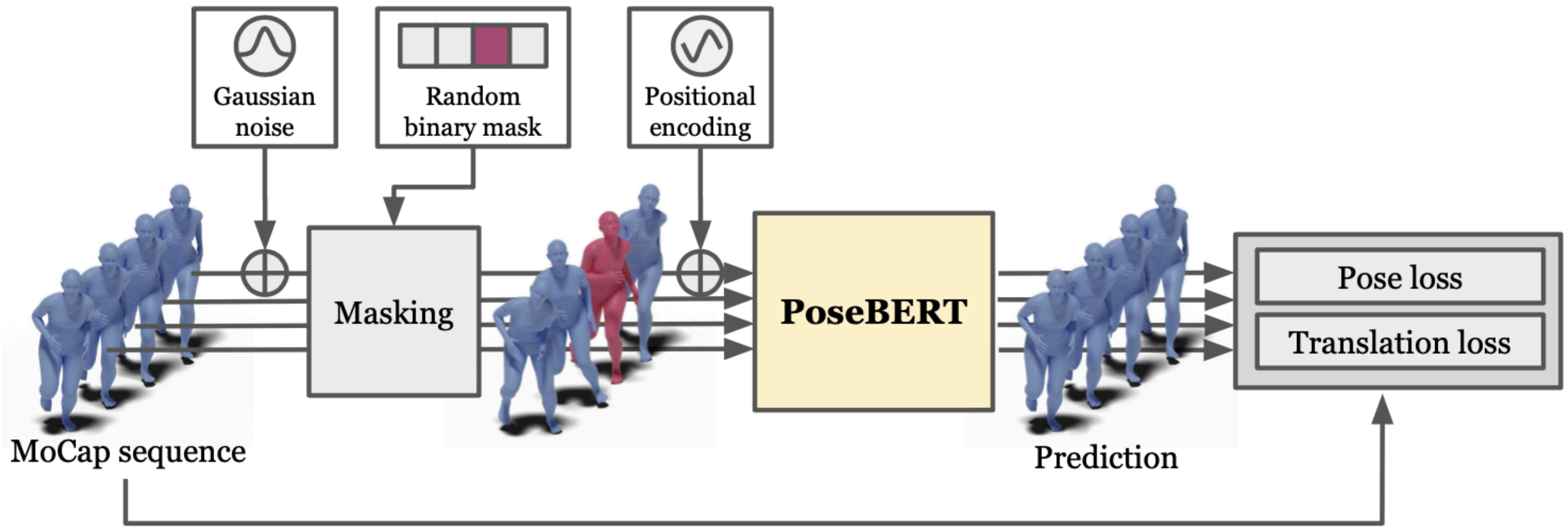
## BEYOND SIMPLE CLASSIFICATION

How to handle mis-detections and improve performances in videos?

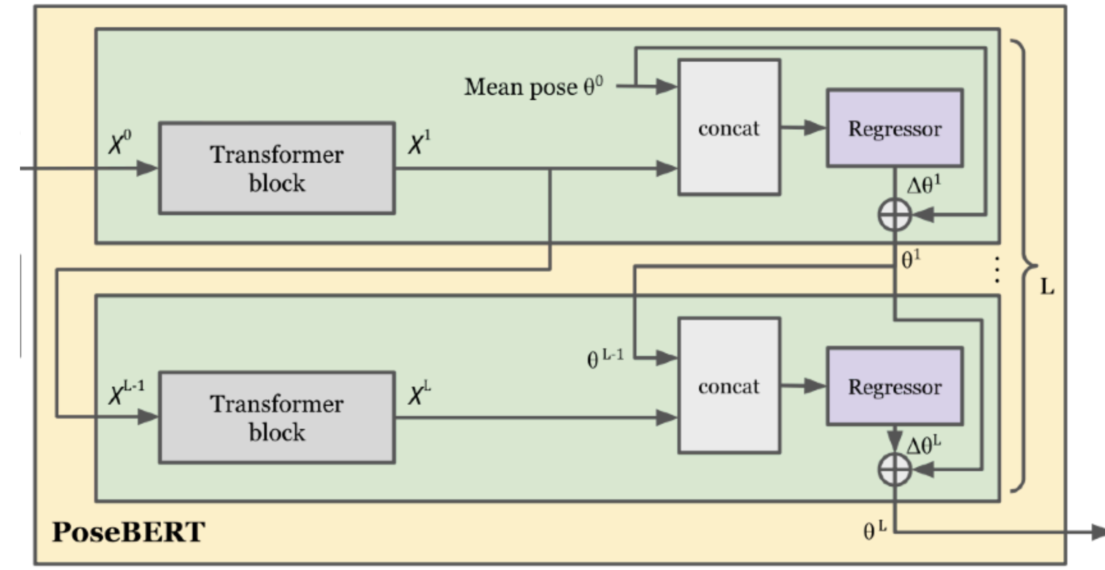
# IMPROVING 3D POSE ESTIMATION IN VIDEOS

Idea: Consider body pose as the tokens of body language and get inspiration from NLP technique **BERT** (Bi-directionnal Encoder Representations from Transformers)

Proposal: Train a BERT-like model on massive amounts of pose sequences from MoCap data.




[Baradel, Groueix, Weinzaepfel, Bregier, Kalantidis & Rogez, Leveraging MoCap data for Human Mesh Recovery, 3DV 2021]



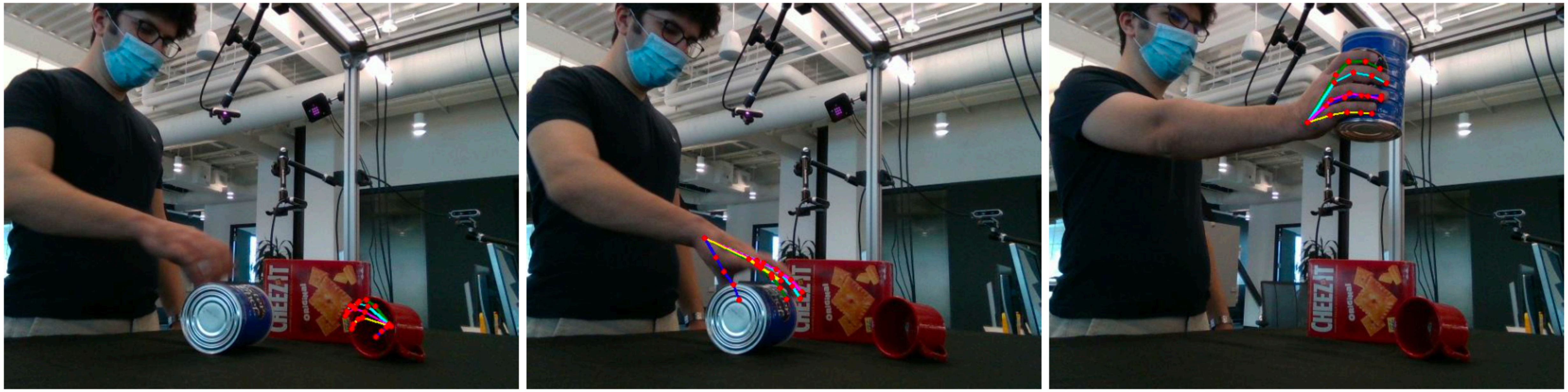
# POSEBERT RESULTS ON MUPOTS

(a) PoseBERT input (LCR-Net++).

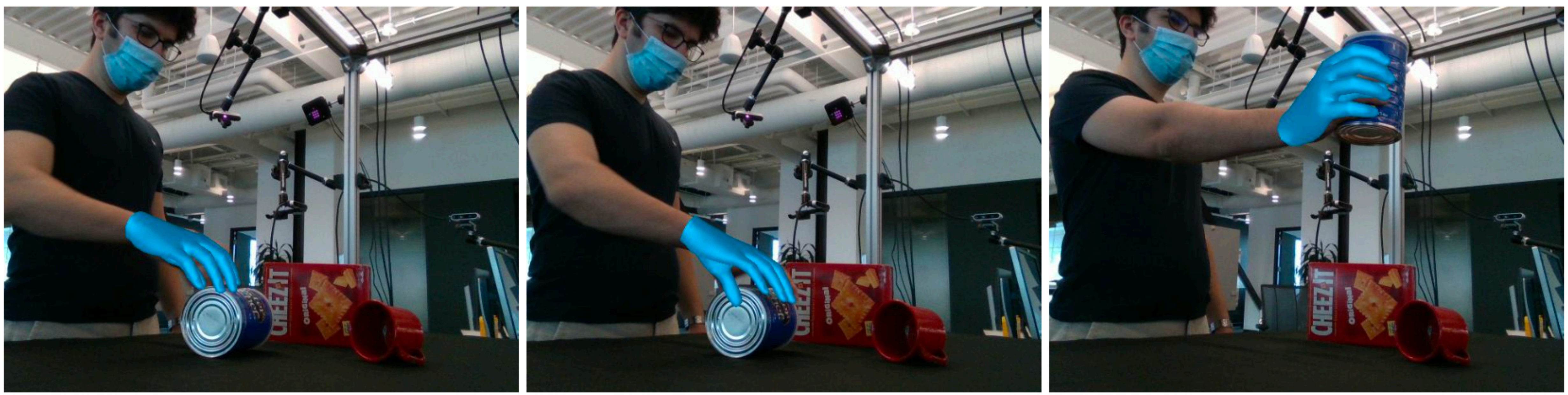
(b) PoseBERT output.

| Method | MPJPE ↓ | PA-MPJPE ↓ | Accel ↓ |
|---|---|---|---|
| LCR-Net++ [1] | 153.76 | 105.23 | 37.98 |
| *(matched groundtruths only)* | *136.79* | *85.53* | *32.86* |
| *(miss detections replaced by nearest detection)* | *139.36* | *86.42* | *28.25* |
| *(+ Savitzky-Golay filtering)* | *138.54* | *86.50* | *16.10* |
| + **PoseBERT** | **126.62 (↓ 27.14)** | **82.53 (↓ 22.70)** | **12.78 (↓ 25.20)** |

# POSEBERT RESULTS ON DEX-YCB DATASET




(a) PoseBERT input (LCR-Net hand expert).

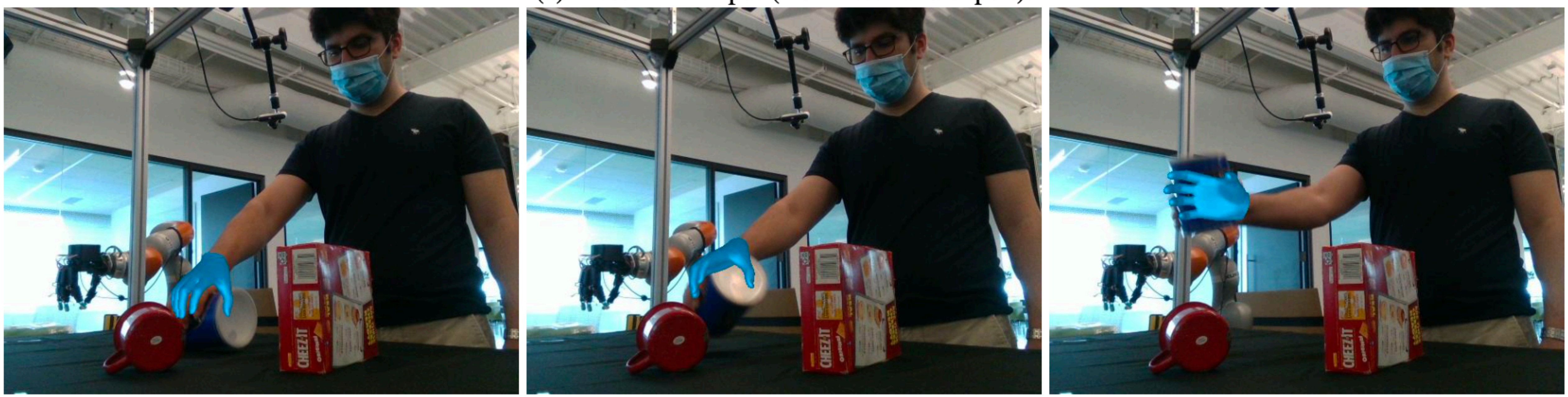



(b) PoseBERT output.

| Detection | Regression | MPJPE ↓ | PA-MPJPE ↓ | Accel ↓ |
|---|---|---|---|---|
| × | HR-Net [56] | 17.34 | 6.83 | 12.77 |
| **(Ground-truth)** | **+ PoseBERT** | **14.05** (↓ **3.29**) | **4.09** (↓ **2.79**) | **3.62** (↓ **9.15**) |
| ✓ | LCR-Net - Hand expert [1] | 46.31 | 16.15 | 33.44 |
| | *(matched groundtruths only)* | *34.51* | *10.07* | *39.81* |
| | *(miss detections replaced by nearest detection)* | *40.73* | *11.10* | *27.43* |
| | **+ PoseBERT** | **29.21** (↓ **17.1**) | **6.88** (↓ **9.27**) | **4.52** (↓ **28.92**) |

# POSEBERT RESULTS ON DEX-YCB DATASET

(a) PoseBERT input (LCR-Net hand expert).



(b) PoseBERT output.

| Detection | Regression | MPJPE ↓ | PA-MPJPE ↓ | Accel ↓ |
|---|---|---|---|---|
| × | HR-Net [56] | 17.34 | 6.83 | 12.77 |
| **(Ground-truth)** | **+ PoseBERT** | **14.05** (↓ **3.29**) | **4.09** (↓ **2.79**) | **3.62** (↓ **9.15**) |
| ✓ | LCR-Net - Hand expert [1] | 46.31 | 16.15 | 33.44 |
| | *(matched groundtruths only)* | *34.51* | *10.07* | *39.81* |
| | *(miss detections replaced by nearest detection)* | *40.73* | *11.10* | *27.43* |
| | **+ PoseBERT** | **29.21** (↓ **17.1**) | **6.88** (↓ **9.27**) | **4.52** (↓ **28.92**) |

# RESULTS WHEN PLUGGED ON OTHER ALGORITHMS

| | 3DPW | MPI-INF | MUPOTS | AIST |
|---|---|---|---|---|
| SPIN | 59.6 | 68.0 | 83.0 | 76.2 |
| **+ PoseBERT** | **57.3 ↓ 2.3** | **64.3 ↓ 3.7** | **80.9 ↓ 2.1** | **74.6 ↓ 1.6** |
| VIBE | 56.5 | 65.4 | 83.4 | 76.0 |
| **+ PoseBERT** | **54.9 ↓ 1.6** | **64.4 ↓ 1.0** | **81.0 ↓ 2.4** | **74.5 ↓ 1.5** |
| MoCap-SPIN | 55.6 | 66.7 | 81.0 | 75.7 |
| **+ PoseBERT** | **52.9 ↓ 2.7** | **63.8 ↓ 2.9** | **79.9 ↓ 1.1** | **74.1 ↓ 1.6** |
| ROMP | 91.1 | - | - | - |
| **+ PoseBERT** | 90.2 ↓ **0.9** | - | - | - |
| LCRNET++ | 68.8 | - | - | - |
| **+ PoseBERT** | 58.5 ↓ **10.3** | - | - | - |

*The reported metric is the PA-MPJPE*

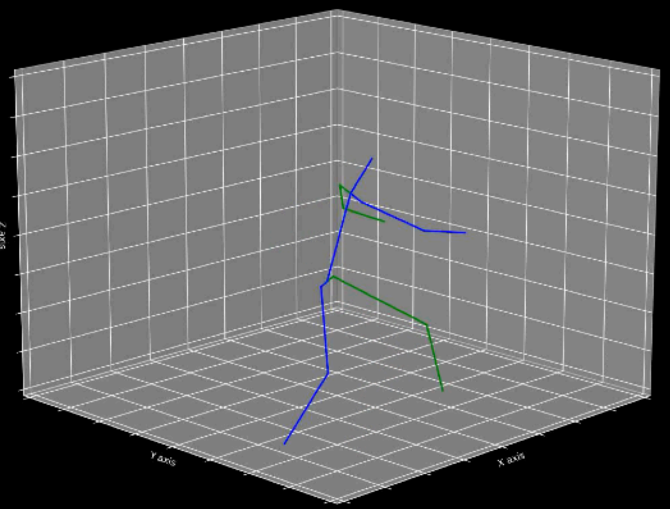
- ➢ Always improve performance
- ➢ More robust estimation
- ➢ Low computation overhead

# QUALITATIVE RESULTS OF SPIN + POSEBERT

Online version run at 15fps on a GPU Tesla T4
Demo code will be released soon

[Baradel, Groueix, Weinzaepfel, Bregier, Kalantidis & Rogez,
Leveraging MoCap data for Human Mesh Recovery, 3DV 2021]

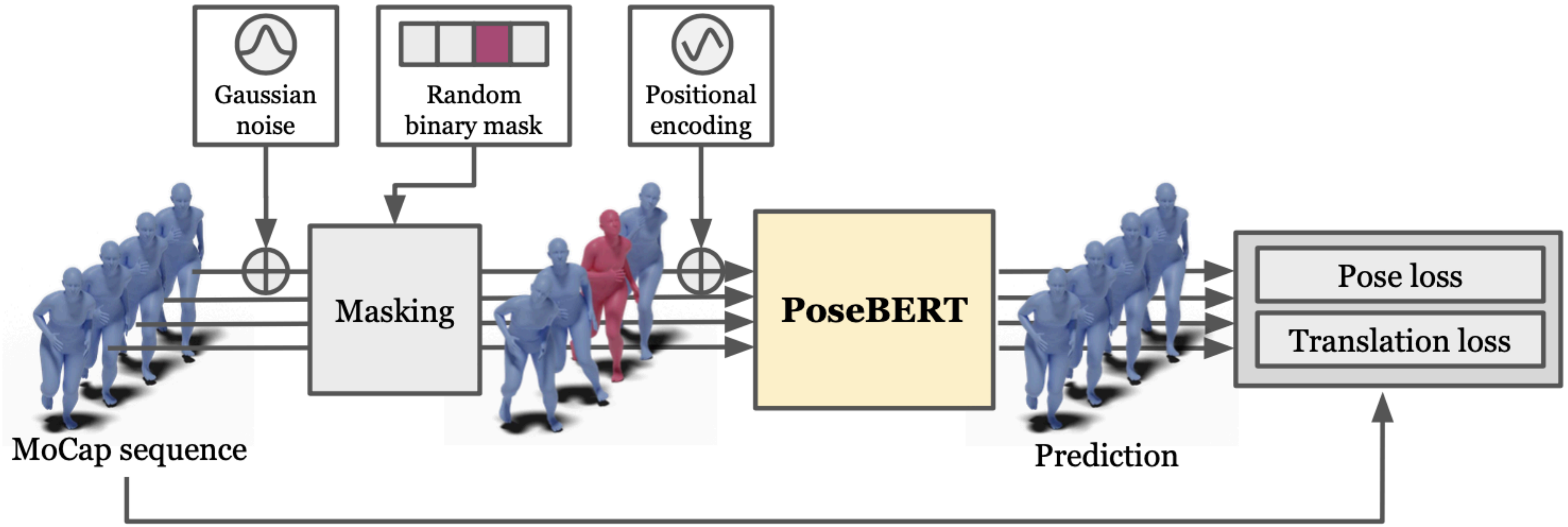
# QUALITATIVE RESULTS OF LCR-NET + POSEBERT



# APPLICATION: ANIMATION OF A ROBOTIC GRIPPER

[Baradel, Groueix, Weinzaepfel, Bregier, Kalantidis & Rogez,
Leveraging MoCap data for Human Mesh Recovery, 3DV 2021]

# TAKE HOME MESSAGE




- **PoseBERT** is a plug and play module trained with **masking** on MoCap data only
- **PoseBERT** can:
  - Correct  noisy 3D poses
  - Complete missing detections
  - Directly output SMPL paprameters

# REFERENCES

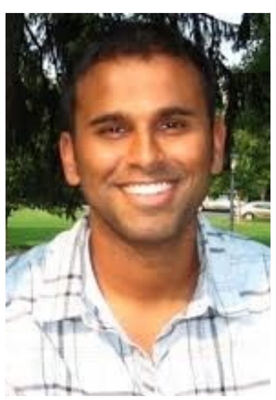
- **Pose estimation by classification: showcase**
  - Rogez, Rihan, Ramalingam, Orrite and Torr, CVPR'08
  - Rogez, Rihan, Orrite and Torr, IJCV'12
  - Rogez, Supancic and Ramanan, CVPR'15
  - Rogez, Supancic and Ramanan, ICCV'15
- **Mocap-guided data augmentation for 3D pose in the wild**
  - Rogez and Schmid, NIPS'16
  - Rogez and Schmid, IJCV'19
- **LCR-Net:**
  - Rogez, Weinzaepfel and Schmid, CVPR'17
    Rogez, Weinzaepfel and Schmid, IEEE PAMI 2019
- **Action recognition / MIMETICS**
  - Weinzaepfel and Rogez, IIJV 2021
- **DOPE**
  - Weinzaepfel, Bregier, Combaluzier, Leroy and Rogez, ECCV 2020
  - Armagan et al., ECCV 2020
- **PoseBERT:**
  - Baradel, Groueix, Weinzaepfel, Bregier, Kalantidis and Rogez, 3DV 2021

# COLLABORATORS












# 3D Human Sensing from monocular visual data using classification techniques

Grégory Rogez

Thanks for your attention!

NAVER LABS Europe