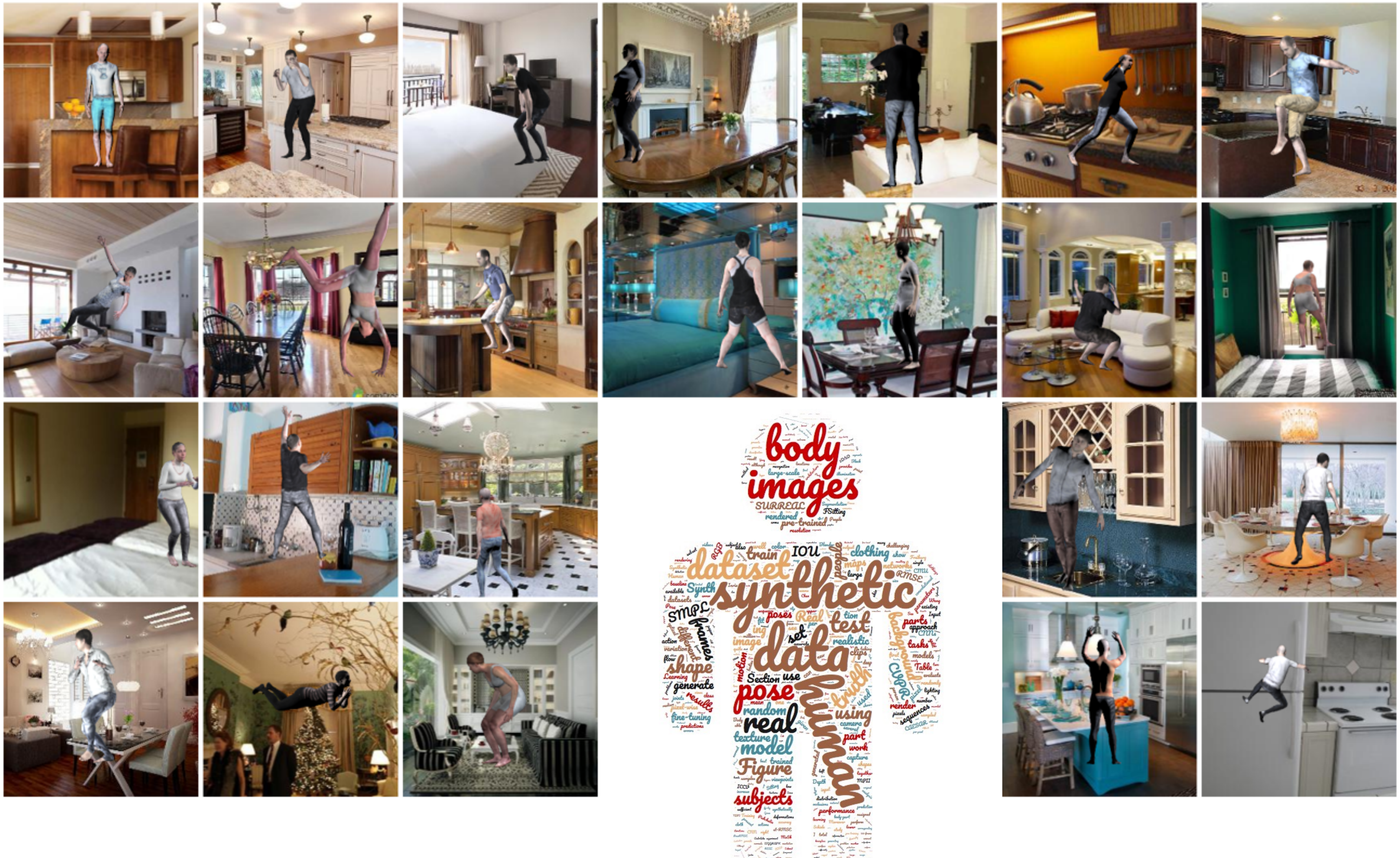


Learning from Synthetic Humans

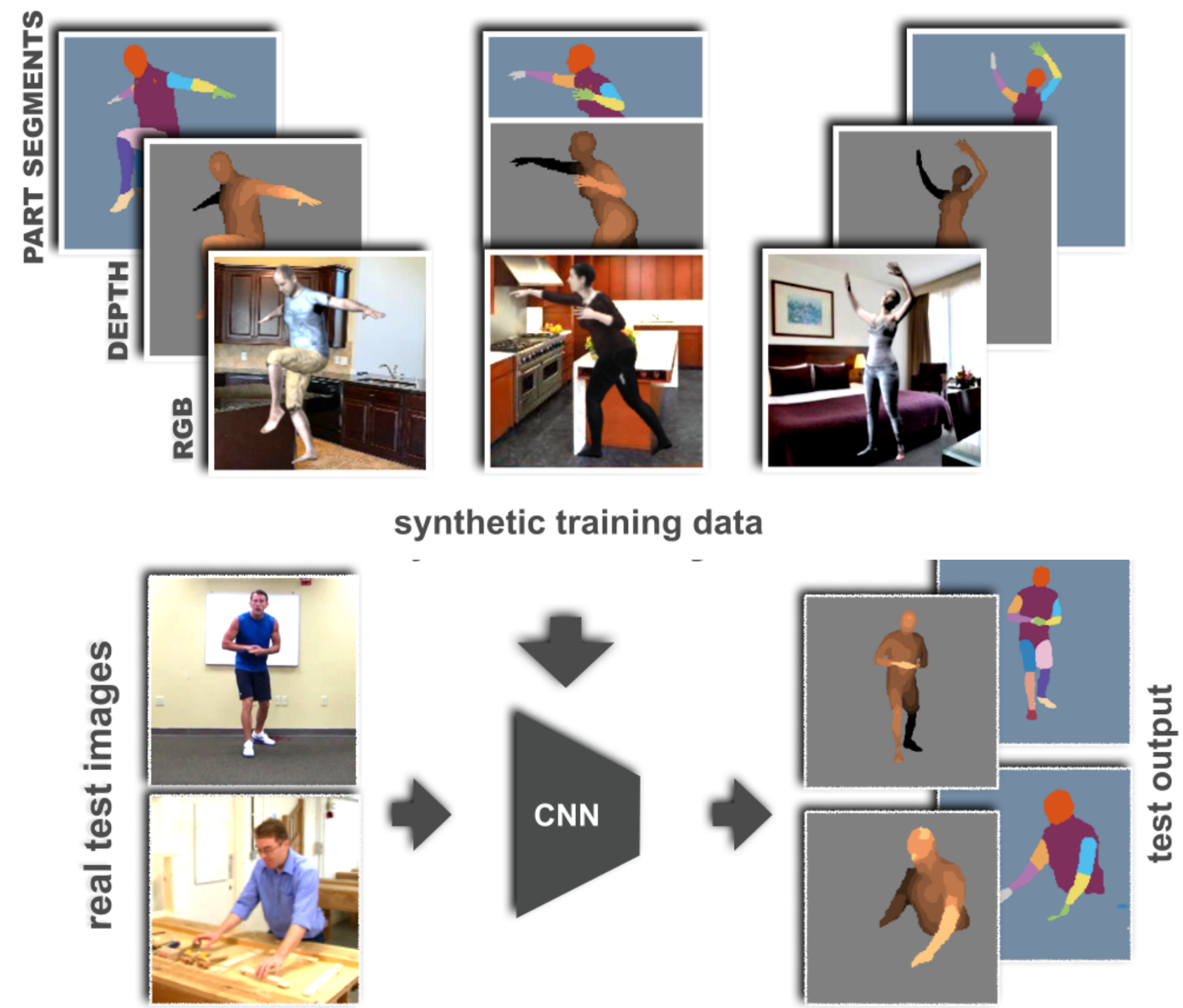
Gül Varol[1], Javier Romero[2], Xavier Martin[1], Naureen Mahmood[2],
Michael J. Black[2], Ivan Laptev[1], Cordelia Schmid[1]

[1] INRIA [2] MPI

INVENTEURS DU MONDE NUMÉRIQUE

Goal

- Generating **synthetic** but **photo-realistic** videos of **people** for training CNNs.
- Demonstrating advantages of this data for training:
    1. Human parts **segmentation**
    2. Human **depth** estimation



Motivation

- The annotation for 2D human pose is expensive to collect and difficult to extend.
- Manual labeling of 3D human pose, depth and motion is impractical.
- Synthetic data comes with rich ground truth.

Challenges

- Domain adaptation

- Occlusion

- Multi-person

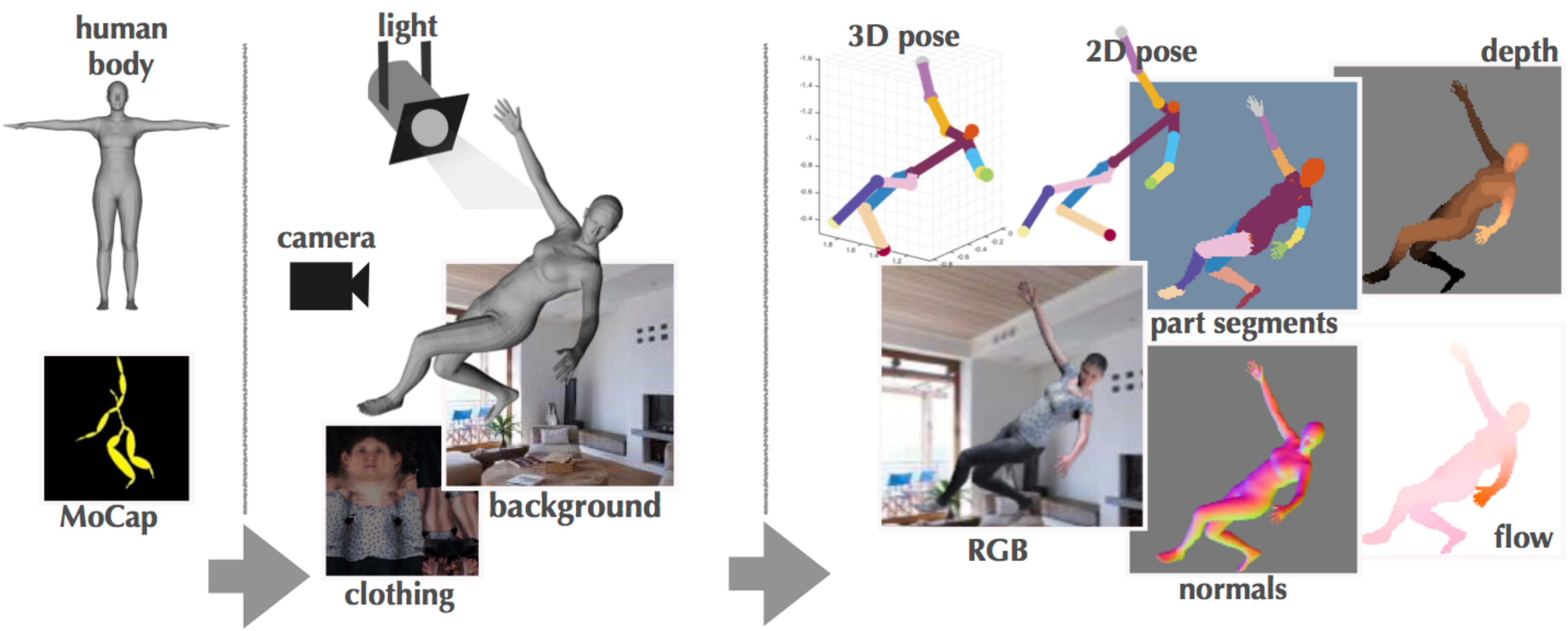
- Object interaction

- Extreme poses

SURREAL Dataset

Synthetic hUmans foR REAL tasks

A body with *random* 3D shape is configured in a *random* pose and a 2D image is rendered from a *random* camera with *random* lighting by compositing the human model with *random* texture on top of a *random* static scene image.



Together with the RGB image, 2D/3D pose, surface normals, optical flow, depth image, and segmentation map for body parts are generated.

SURREAL Dataset

- CAESARS dataset for human body shapes
- LSUN dataset for static background images
- CAESARS dataset and another collection of 3D scans for body textures (clothes)
- CMU dataset for MoCap sequences (marker data)

	#subjects	#sequences	#clips	#frames
Train	115	1,964	55,001	5,342,090
Test	30	703	12,528	1,194,662
Total	145	2,607	67,582	**6,536,752**

SURREAL Dataset

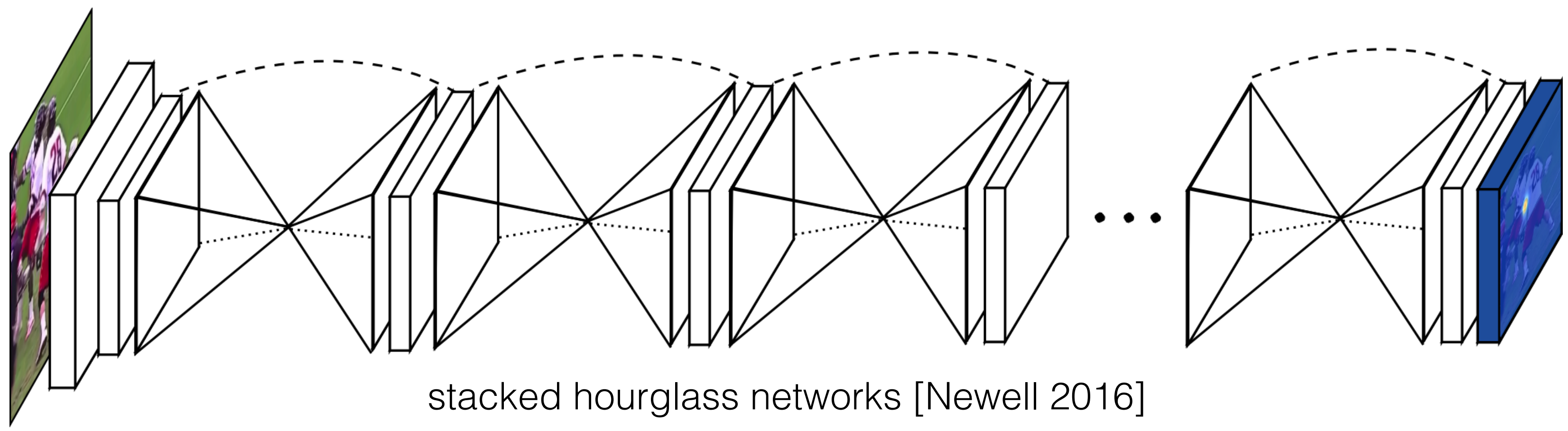
https://www.youtube.com/watch?v=SJ0vw6CzS7U

Tasks

- Human parts segmentation
- Human depth estimation

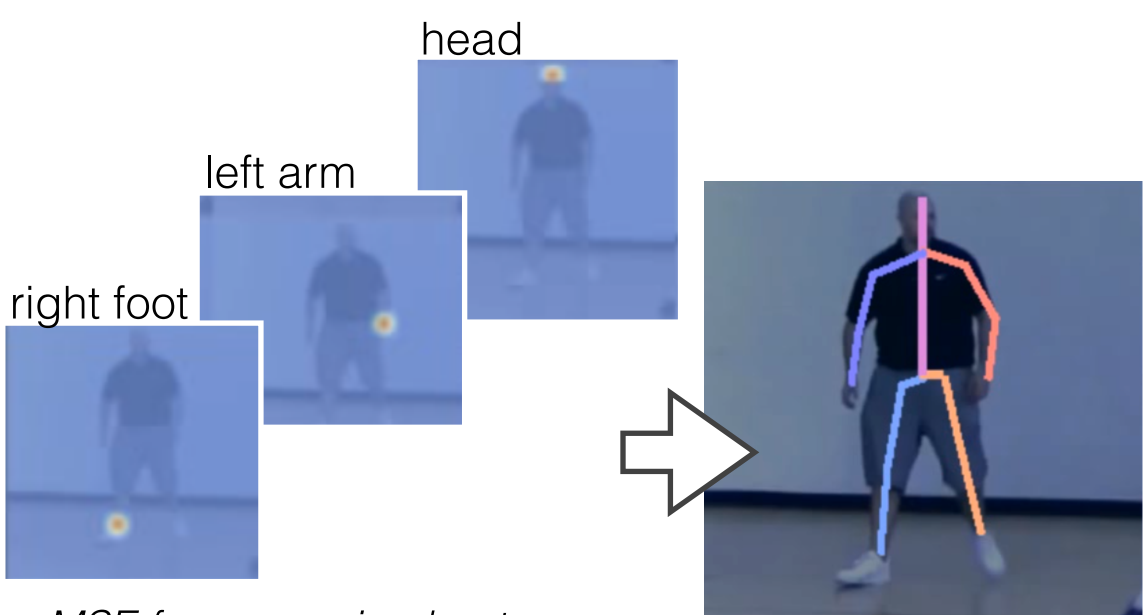
Approach - Segmentation

We build on the stacked hourglass network architecture introduced originally for 2D pose estimation problem, extend it for segmentation.



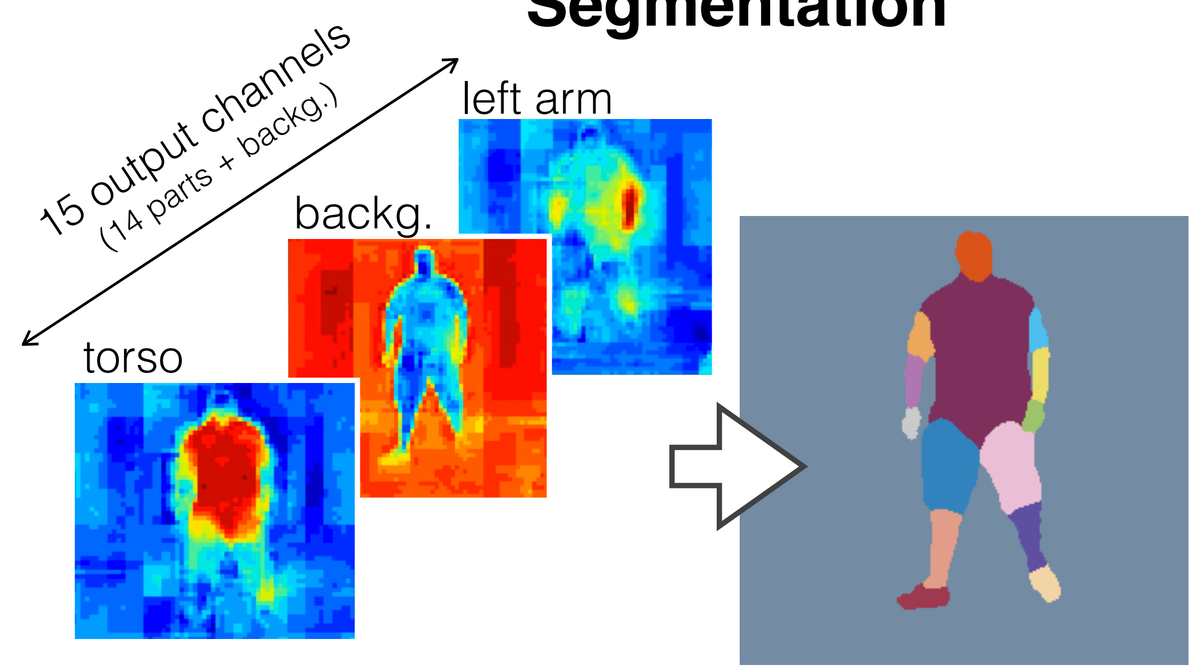
stacked hourglass networks [Newell 2016]

2D pose



MSE for regressing heatmaps

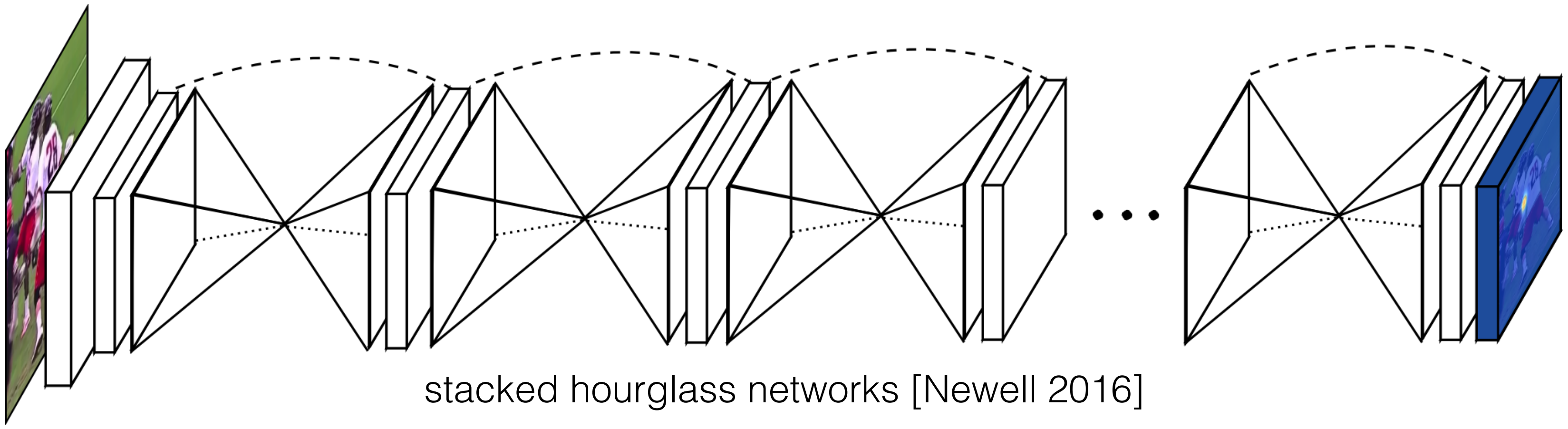
Segmentation



Softmax error for classifying pixels as one of the parts

Approach - Depth

Depth is continuous. However we are interested in the global pose of the person instead of the precise surface. We discretize depth of a person in 20 values and pose depth estimation as a classification problem.

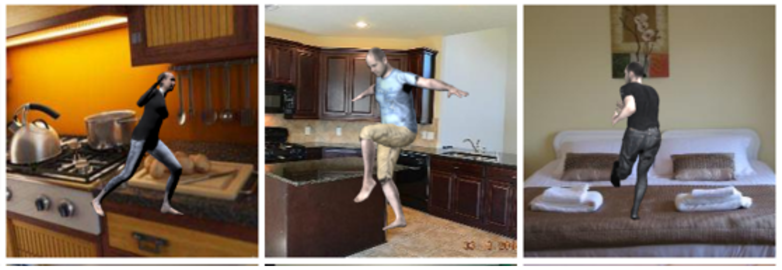


stacked hourglass networks [Newell 2016]

We align depth maps so that the pelvis depth falls on the center of the axis and quantize the depth into 19 bins (9 behind and 9 in front of the pelvis).

Experiments - Datasets

- SURREAL
  - validation on synthetic test set for **segmentation** and **depth**



- Freiburg Sitting People
  - **segmentation** dataset

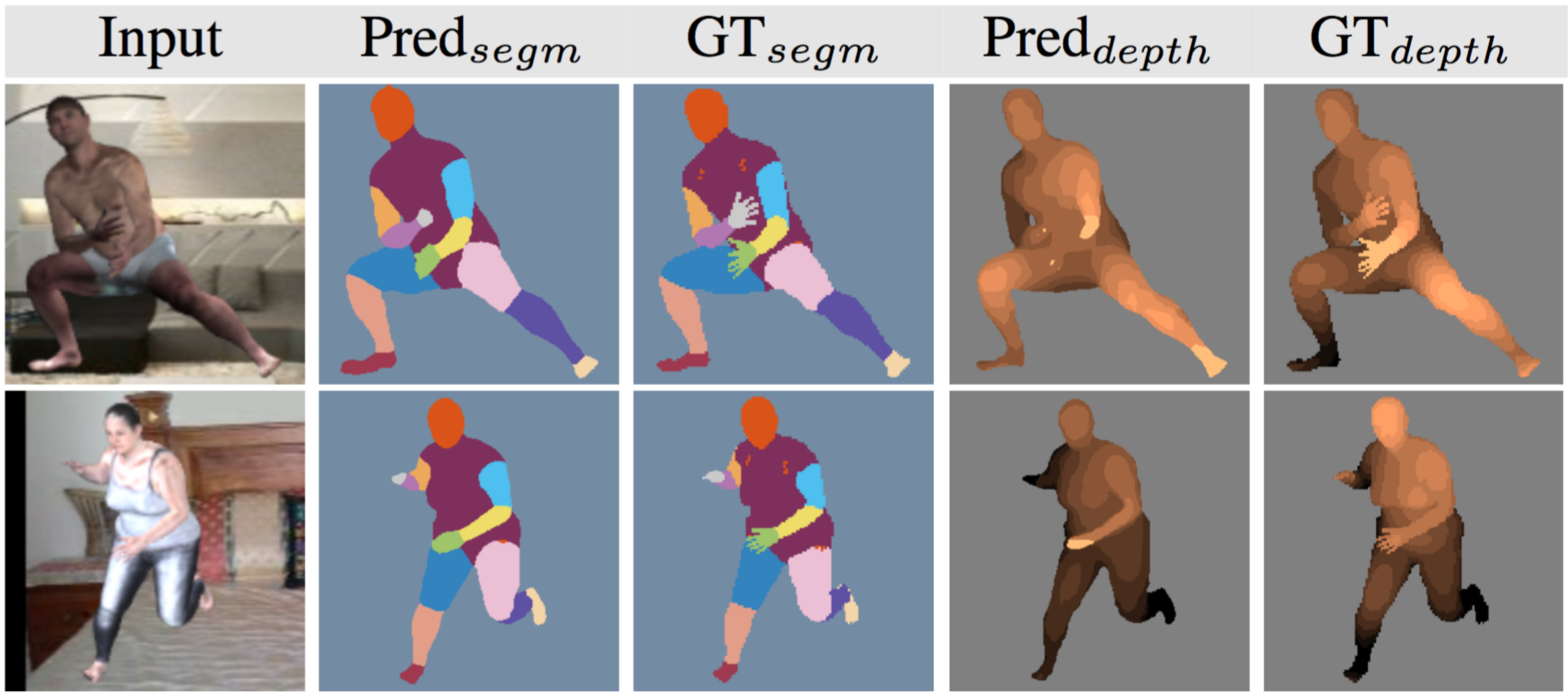
- Human3.6M
  - MoCap dataset with RGB videos
  - we generate ground truth for **segmentation** and **depth**

- MPII Human Pose
  - 2D pose dataset
  - no ground truth
  - qualitative results for **segmentation** and **depth**

Experiments - Evaluation Metrics

- Segmentation
  - Pixel accuracy
  - IOU (intersection over union)
- Depth
  - RMSE (root mean squared error)
  - st-RMSE (scale and translation invariant RMSE)
  - pose-RMSE (RMSE evaluated on joint locations)
  - st-pose-RMSE

Experiments - SURREAL Dataset

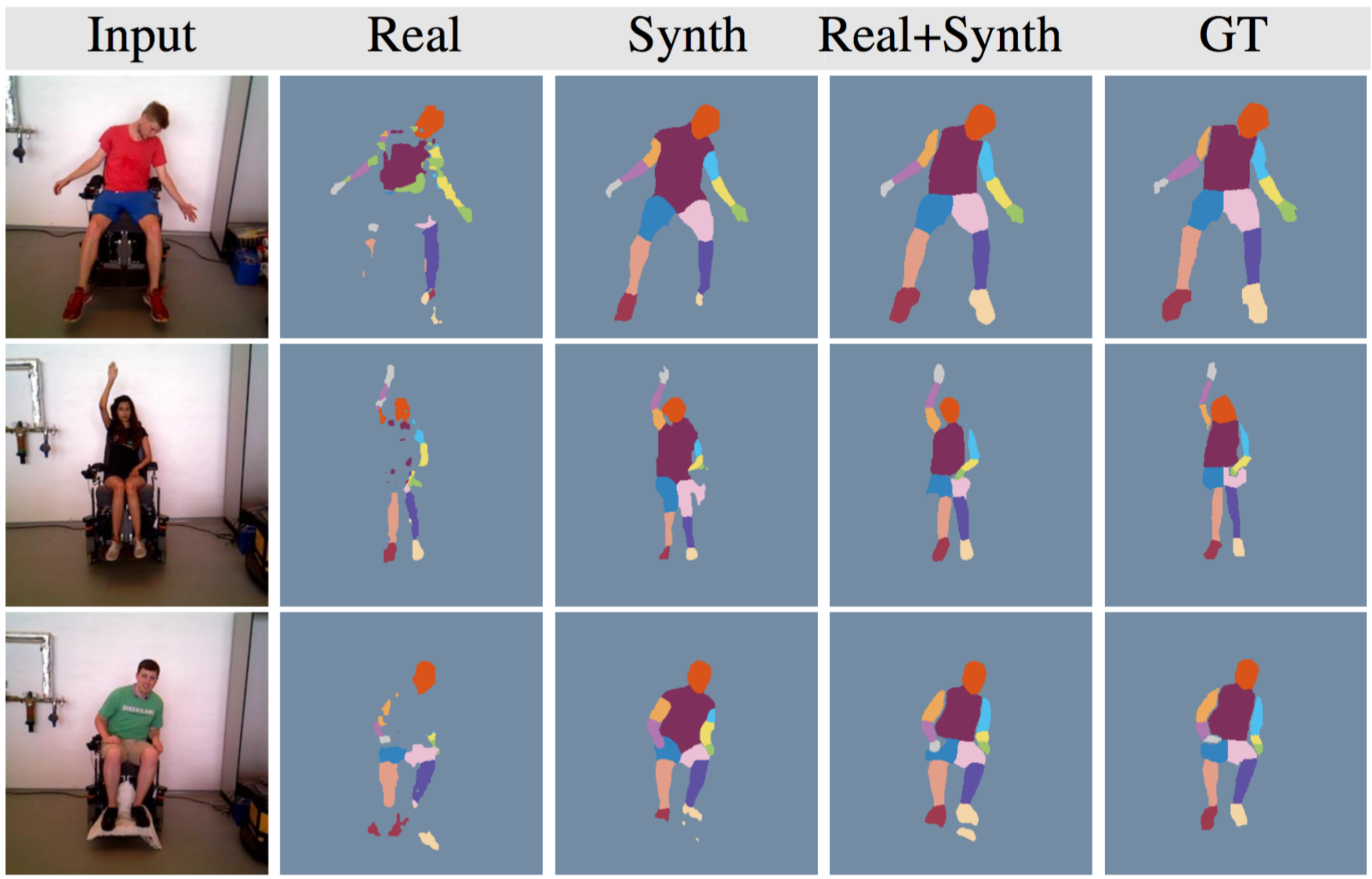


Segmentation	
IOU	69.13 %
Accuracy	80.61 %

Depth	
RMSE	72.9 mm
st-RMSE	56.3 mm

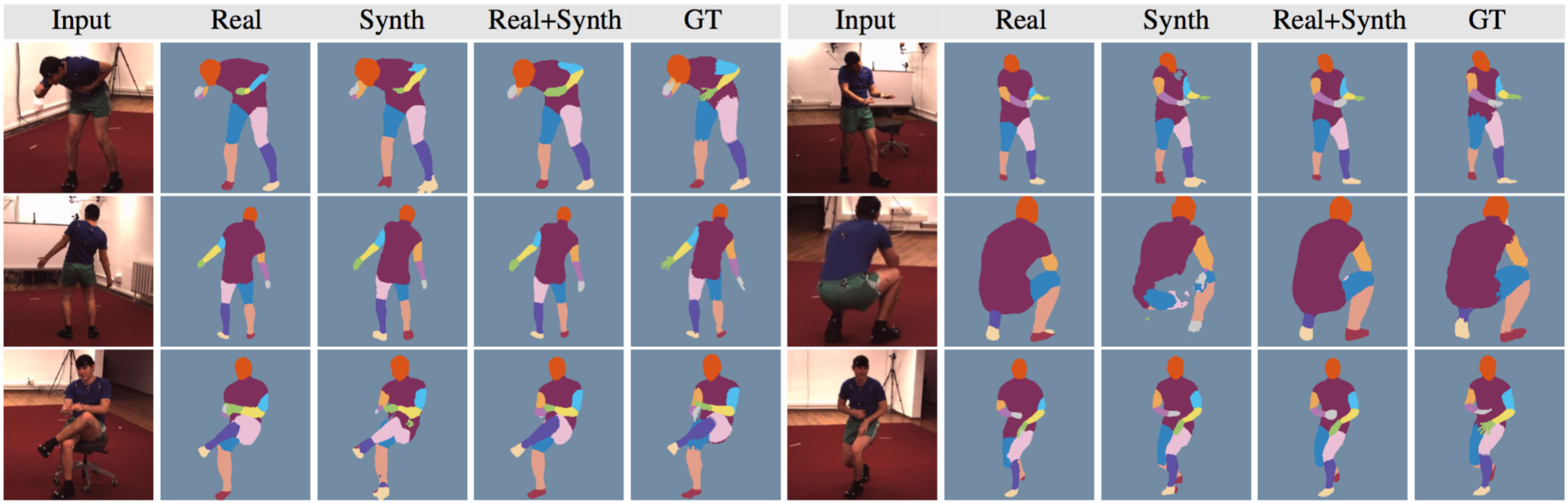
Varol et al. Learning from Synthetic Humans, CVPR'17.

Experiments - Freiburg Sitting People Dataset



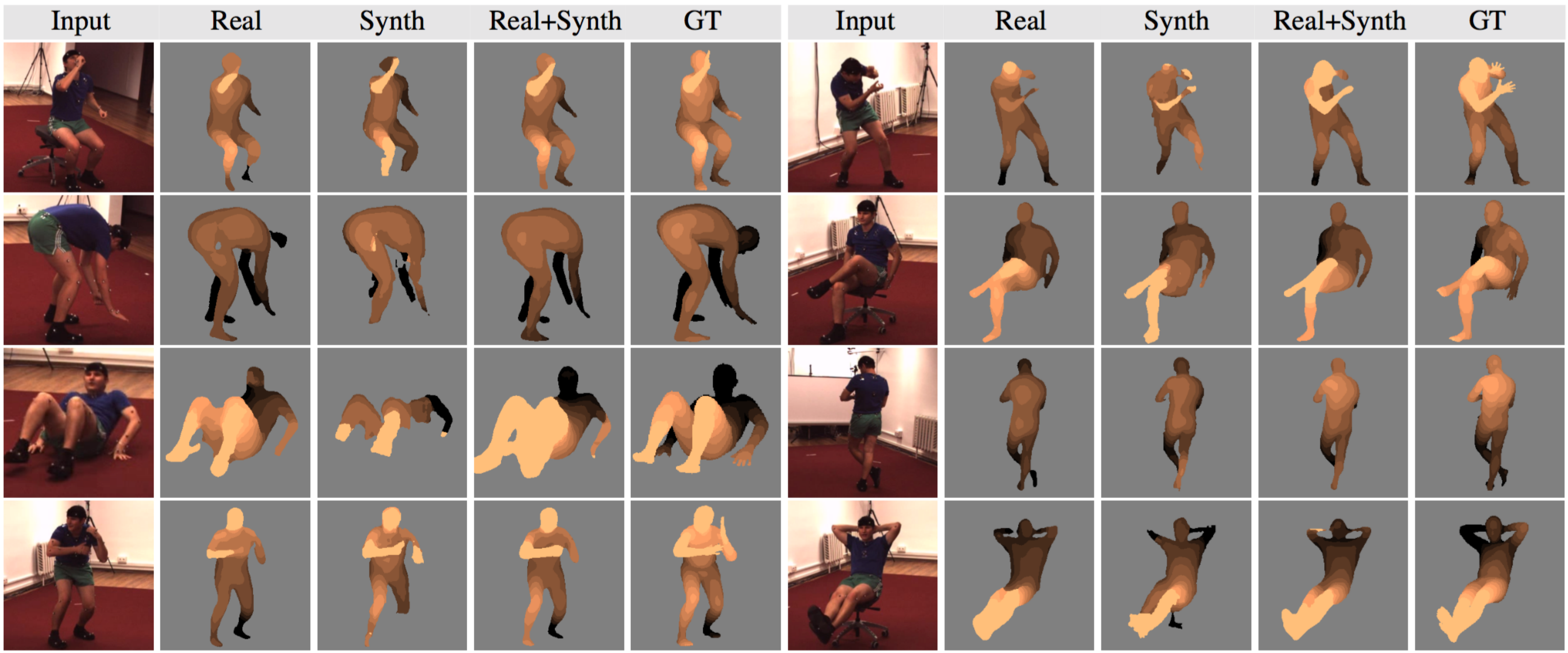
Training data	Head IOU	Torso IOU	Legs$_{up}$ IOU	mean IOU	mean Acc.
Real+Pascal[21]	-	-	-	64.10	81.78
Real	58.44	24.92	30.15	28.77	38.02
Synth	73.20	65.55	39.41	40.10	51.88
Synth+Real	72.88	80.76	65.41	59.58	78.14
Synth+Real+up	**85.09**	**87.91**	**77.00**	**68.84**	**83.37**

Experiments - Human3.6M Dataset



Training data	IOU fg+bg	IOU fg	Accuracy fg+bg	Accuracy fg
Real	49.61	46.32	58.54	55.69
Synthetic	46.35	42.91	56.51	53.55
Synthetic+Real	57.07	54.30	67.72	65.53

Experiments - Human3.6M Dataset



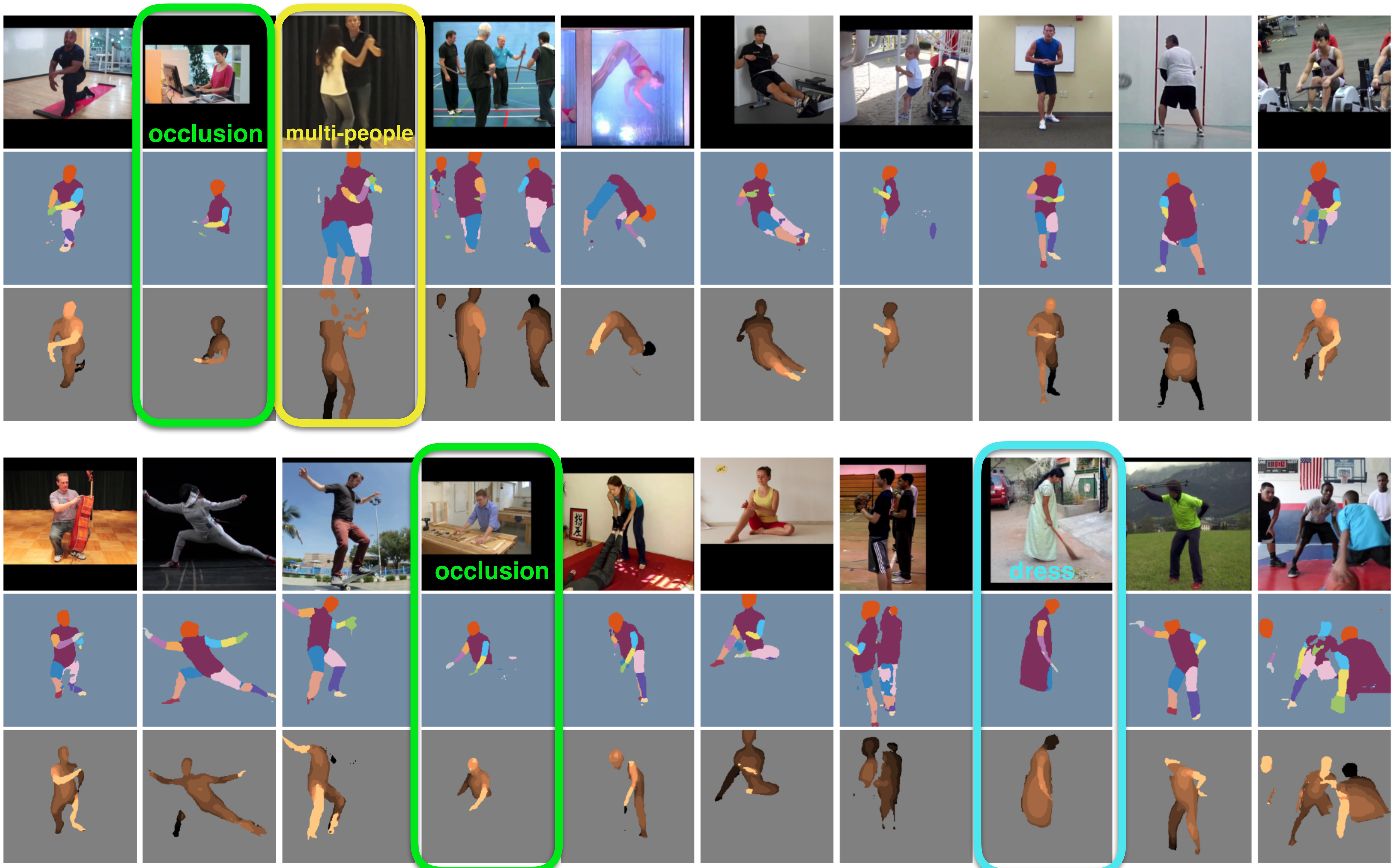
Training data	RMSE	st-RMSE	PoseRMSE	st-PoseRMSE
Real	96.3	75.2	122.6	94.5
Synthetic	111.6	98.1	152.5	131.5
Synthetic+Real	90.0	67.1	92.9	82.8

(mm)

Experiments - Human3.6M Dataset

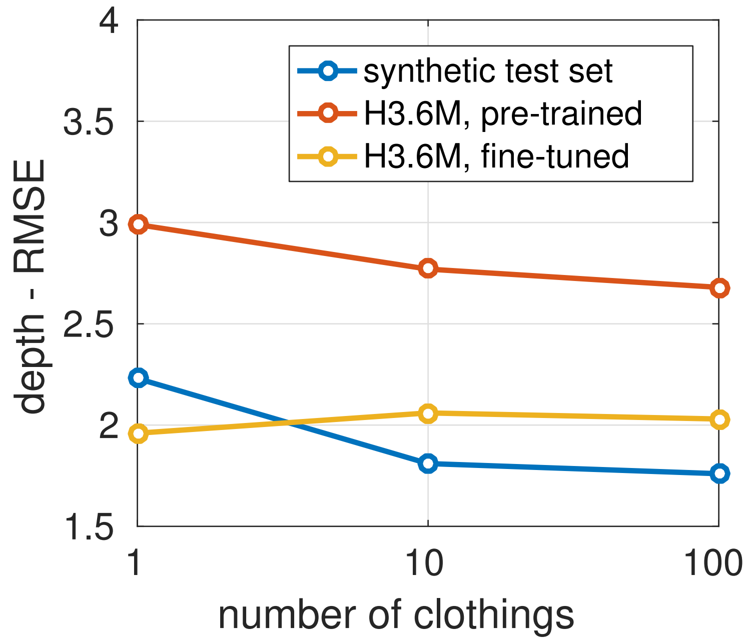
https://www.youtube.com/watch?v=bK4tAGOWayE

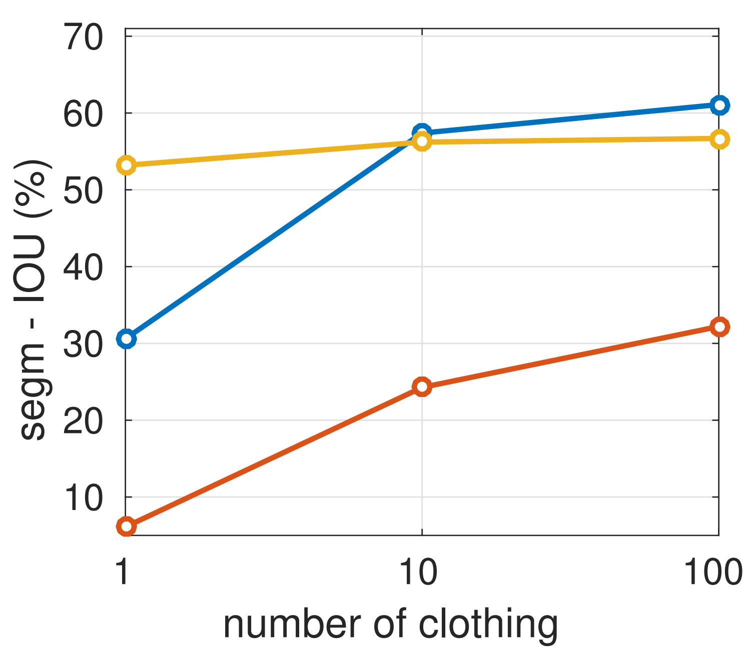
Experiments - MPII Human Pose Dataset



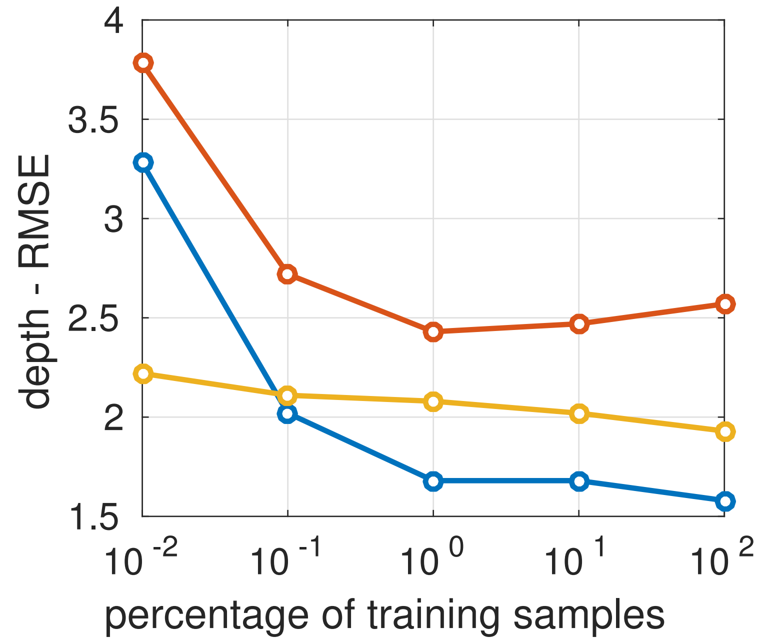
Experiments - Design choices

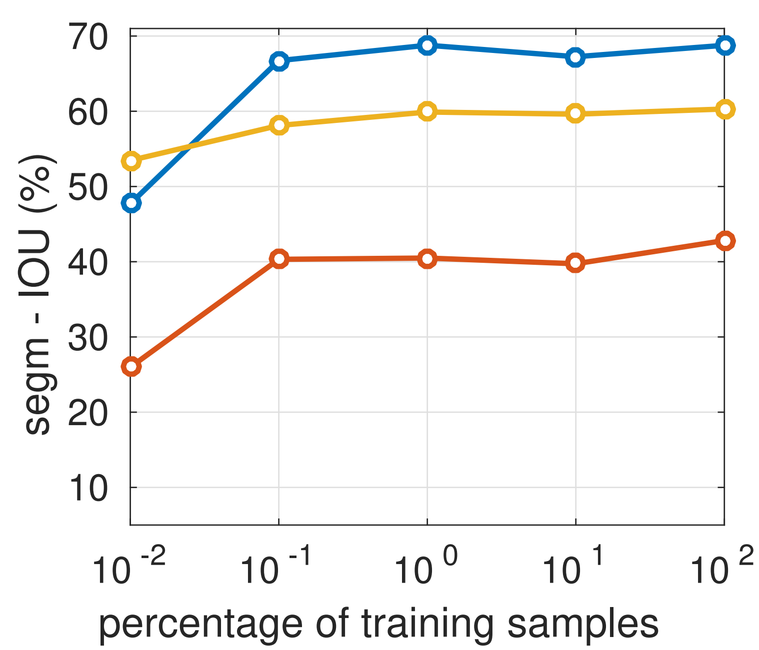
Clothing variation





Amount of data





MoCap variation

We rendered synthetic data using Human3.6M MoCap.

	MoCap source	
	s-CMU	s-H3.6M
depth - RMSE	2.57	**2.44**
segm - IOU (%)	42.82	**48.11**

Tested on real-H3.6M

Conclusions

- It is possible to learn from synthetic images of people.
- We have shown the generalization capability of CNNs trained on synthetic people on two tasks:
    - segmentation,
    - depth estimation.
- The rich ground truth can potentially be used for other tasks.

Thanks

www.di.ens.fr/willow/research/surreal

Data and code are available.