15th IEEE International Conference on
Automatic Face and Gesture Recognition

IEEE Trans. Biometrics, Behavior, and Identity Science
“The Best of FG 2020” Special Issue

MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation

István Sárándi[1], Timm Linder[2], Kai O. Arras[2], Bastian Leibe[1]

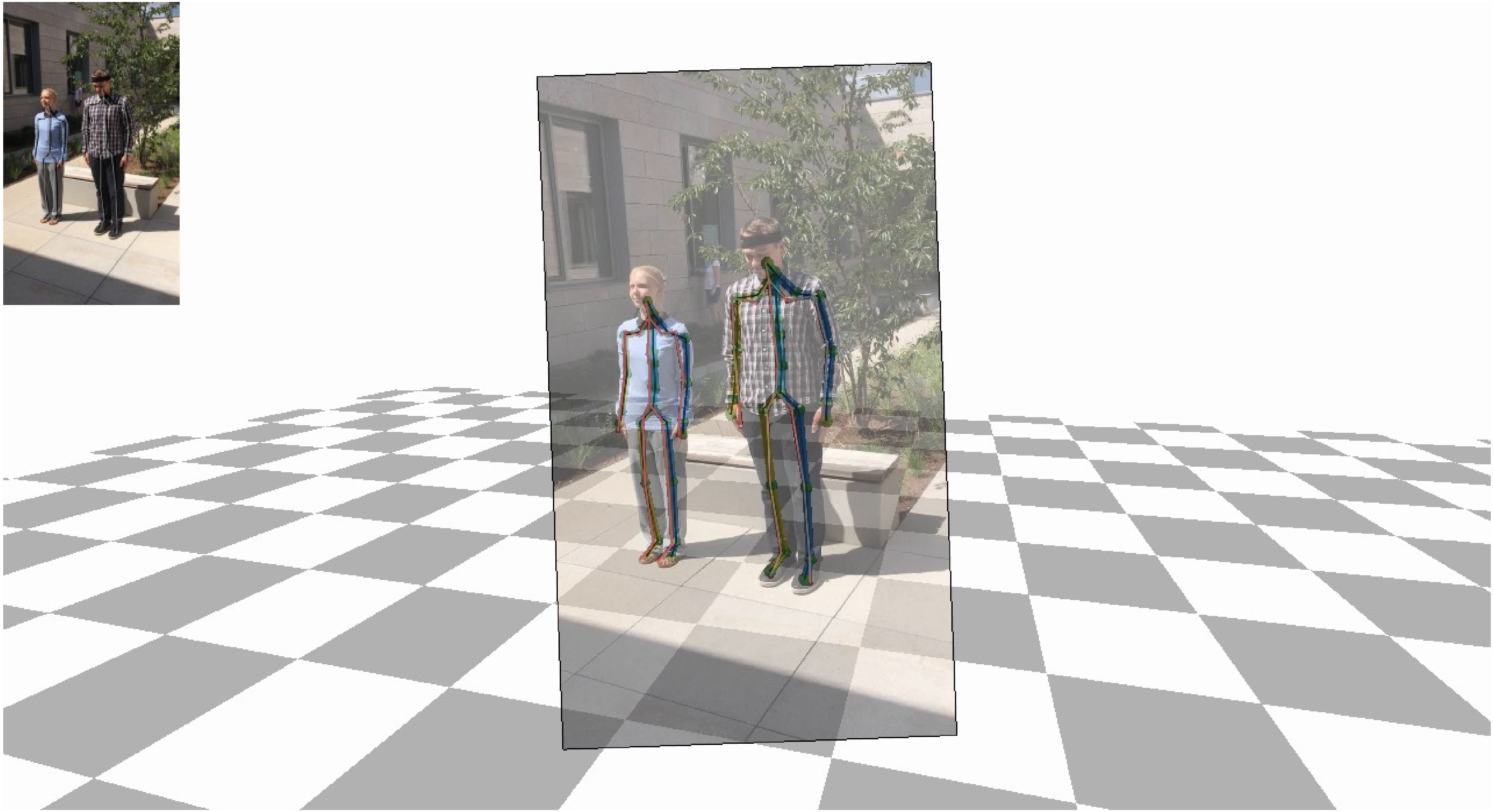
[1] Computer Vision Group, RWTH Aachen University, Germany
[2] Robert Bosch GmbH, Corporate Research, Renningen, Germany

BOSCH
Invented for life

BOSCH-FORSCHUNGSSTIFTUNG
IM STIFTERVERBAND

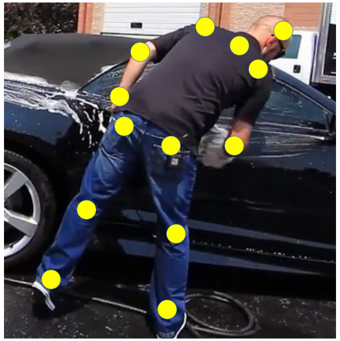
RWTH AACHEN UNIVERSITY

Reference Prediction



Background: Human Pose Estimation in 2D and 3D

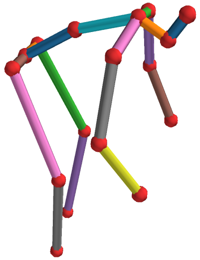
2D: pixels



$(x_1, y_1, \ldots, x_N, y_N)$

(126 px, 50 px, …)

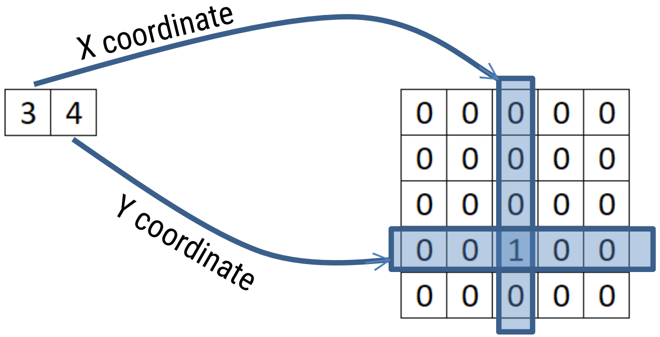
3D: meters



$(X_1, Y_1, Z_1, \ldots, X_N, Y_N, Z_N)$

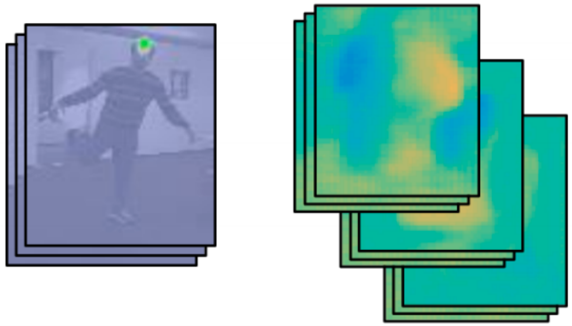
(1.5 m, 0.6 m, 3.1 m, …)

Background: How to Represent Joint Positions?

	Represent as activation *values*	Represent as activation *location*
2D: pixels	$(x_1, y_1, \ldots, x_N, y_N)$	
3D: meters	$(X_1, Y_1, Z_1, \ldots, X_N, Y_N, Z_N)$	?

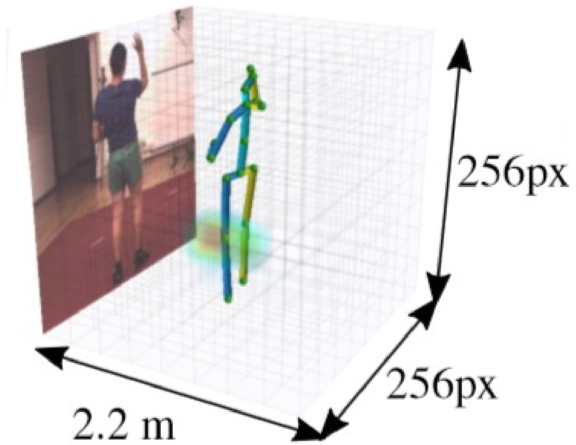
Related Work: Generalizing Heatmaps to 3D Pose

- [Mehta17TOG] “Hybrid”: Location maps



- [Pavlakos17CVPR] As activation *location*

Volumetric heatmaps (2.5D)



Key Idea: Combine the Benefits

Direct regression of coordinates

Can directly regress metric 3D

Not limited by image truncation

Continuous output

Does not exploit the conv. structure

2.5D heatmaps

Needs post-processing for metric 3D

Can only predict within FOV

Effective use of convolutional structure

High-resolution needed?

Discrete output?

Key Idea: Combine the Benefits

Direct regression of coordinates

Can directly regres

Not limited by imag

Continuous output

Does not exploit th

2.5D heatmaps

s post-processing for metric 3D

nly predict within FOV

ve use of convolutional structure

esolution needed?

te output?

Our approach

Heatmap representation

Directly regress metric 3D

Not limited by image truncation

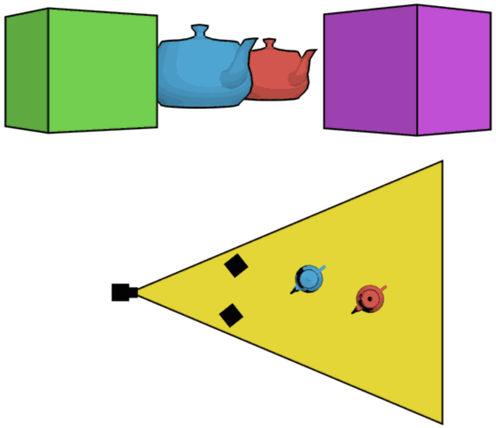
Continuous output

Low-res heatmap is enough

Simple and fast architecture

Background: Scale/Distance Ambiguity

- d: Distance of person to camera
- f: Focal length
- S: Metric size of person
- s: Projected size of person



$$s = S \cdot f / d$$

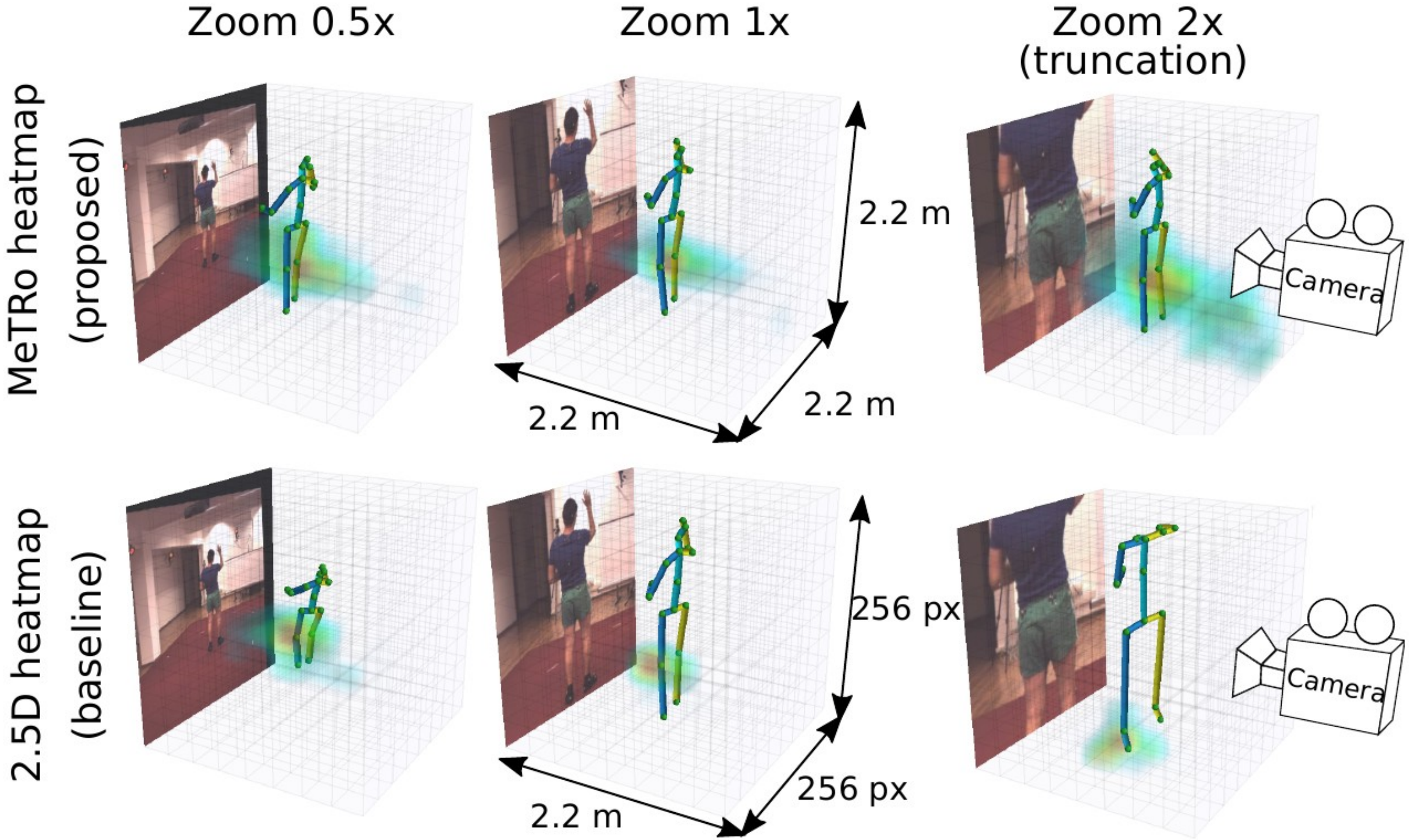
Easy, direct image measurement

Moderately hard, but plausible

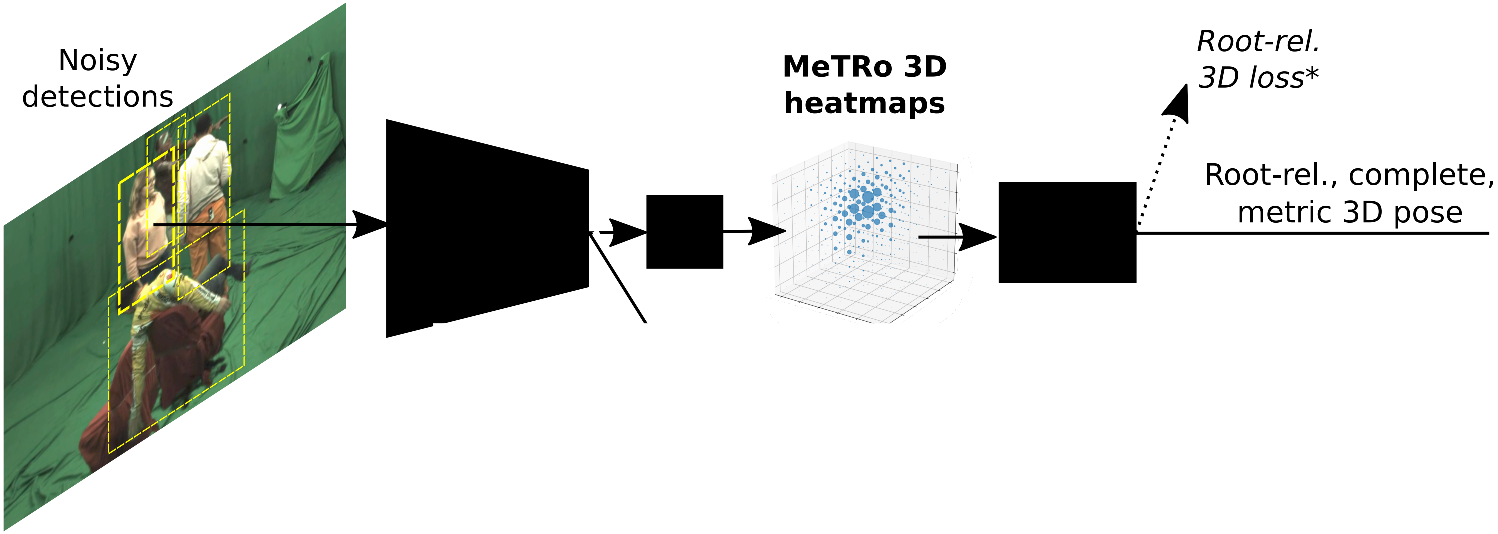
Often known and calibrated, otherwise hard to estimate

Very hard to estimate directly from a crop

MeTRo 3D Heatmap vs 2.5D Heatmap



Our Approach

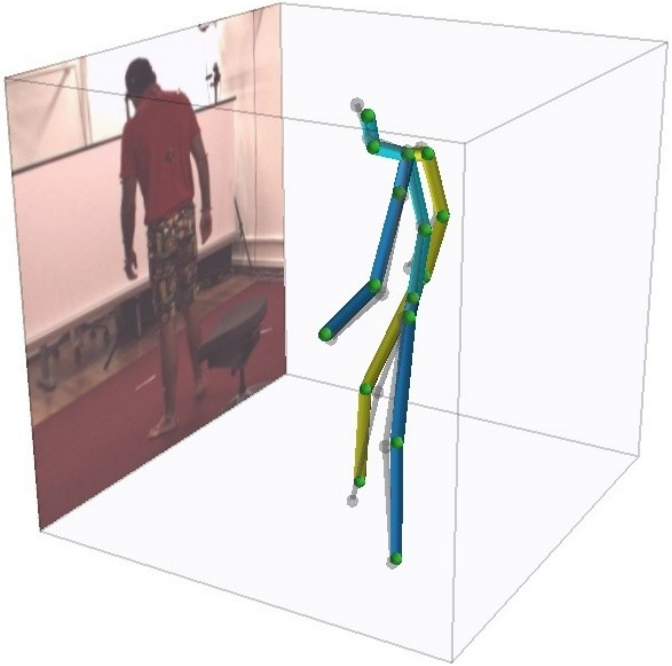


(Backbone is ResNet-50, except for 3DPW Challenge experiment)

Soft-argmax: [Levine16JMLR, Nibali18Arxiv, Sun18ECCV]

Results: Human3.6M

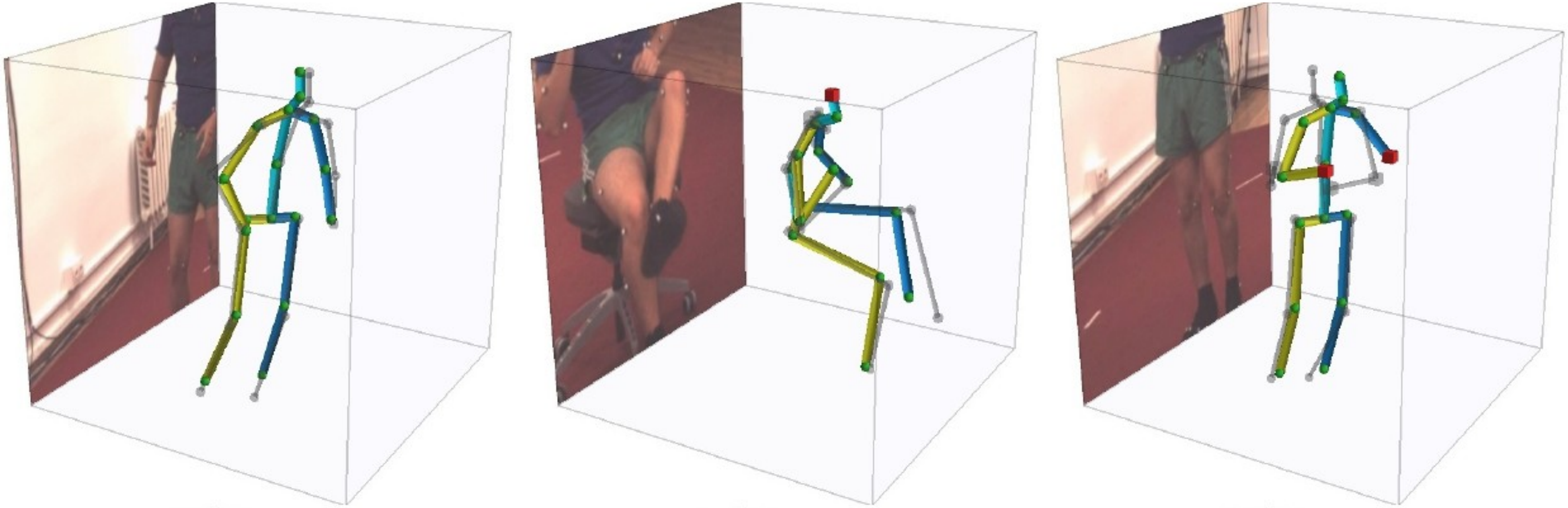
	MPJPE↓
Pavlakos *et al.* [13]	71.9
Zhou *et al.* [10]	64.9
Martinez *et al.* [8]	62.9
Fang *et al.* [61]	60.4
Yang *et al.* [62]	58.6
Pavlakos *et al.* [63]	56.2
Liu *et al.* [64]	52.4
Xu *et al.* [65]	**49.2**
Sharma *et al.* [66]	58.0
Cai *et al.* [67]	50.6
2.5D baseline	50.2±0.3
MeTRo (ours)	**49.3**±**0.7**

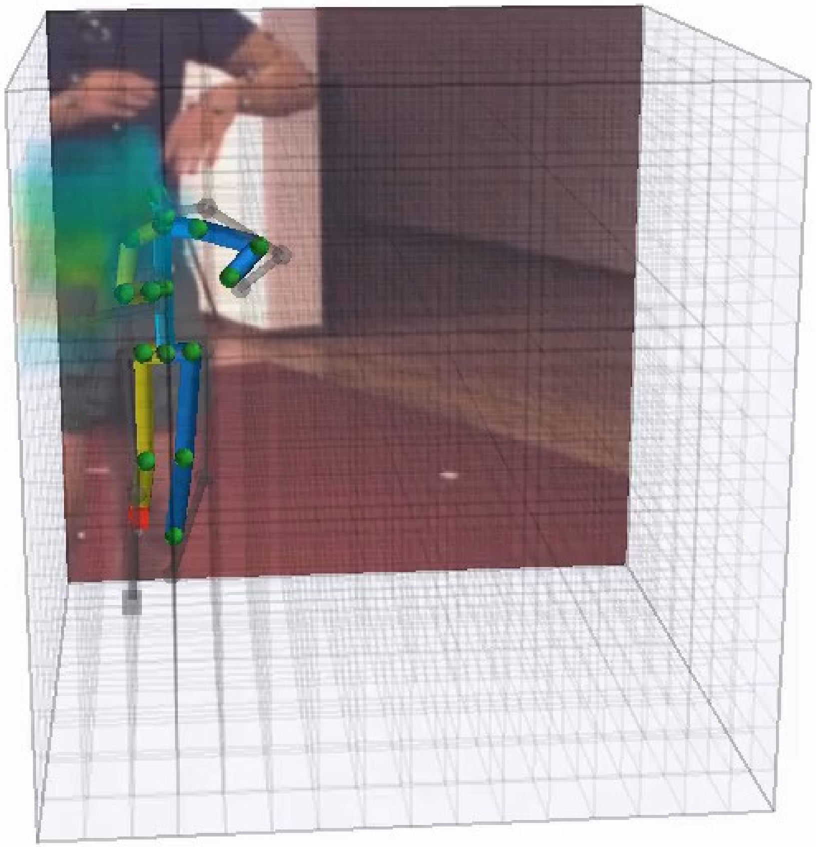


Results: Truncated Human3.6M

	Mehta* [9]	Zhou* [10]	Vosoughi [46]	**MeTRo***	**MeTRo**
All joints	396.4	400.5	185.0	124.7	**77.8**
Present joints	338.0	332.5	173.6	76.8	**59.8**

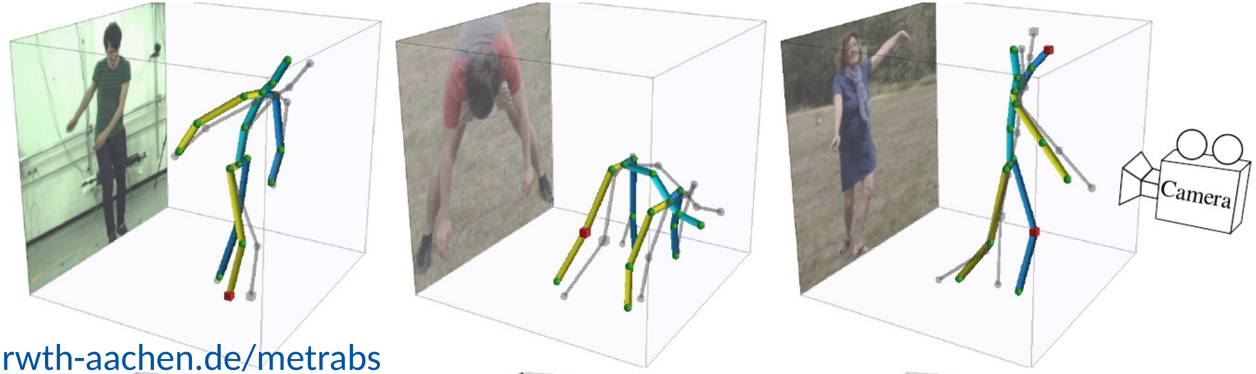
(*No strong truncations applied during training)





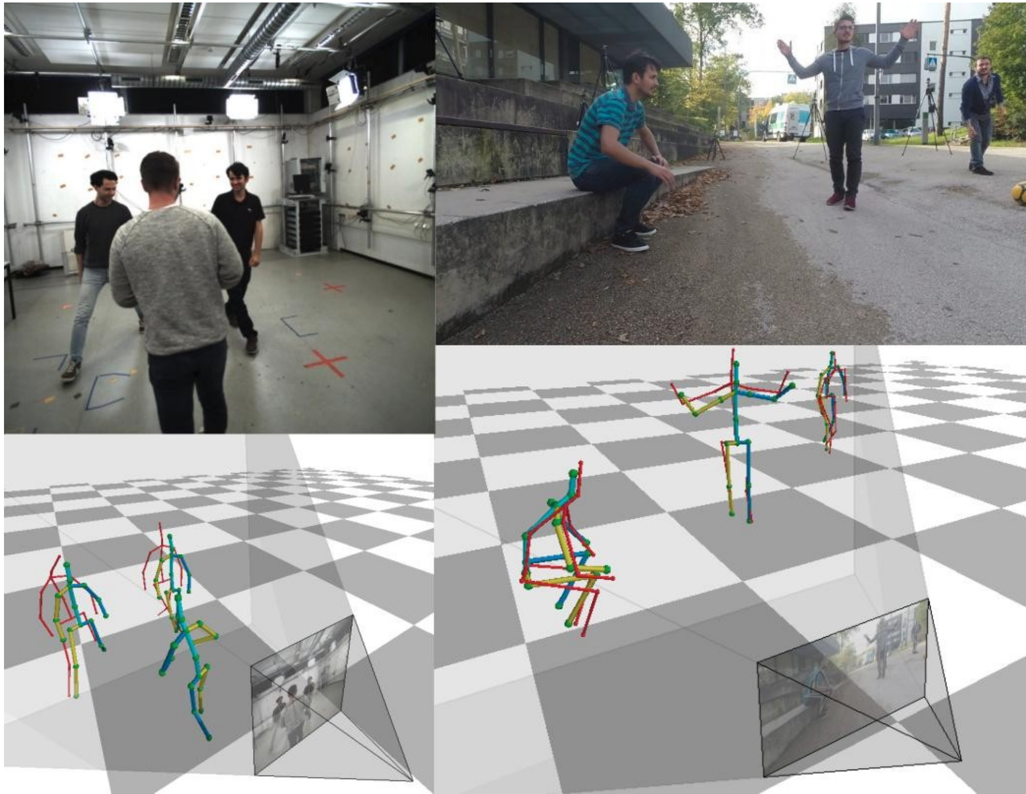
Results: MPI-INF-3DHP

		Green screen	No gr.sc.	Out-door	Total PCK↑	Total AUC↑	Total MPJPE↓
Scale normalized	Rogez *et al.* [74]*	–	–	–	59.7	27.6	158.4
	Zhou *et al.*$^{\text{H+M}}$ [10]*	71.7	64.7	72.7	69.2	32.5	137.1
	Zhou *et al.*$^{\text{H+M}}$ [76]	75.6	71.3	80.3	75.3	38.0	–
	Mehta *et al.*$^{\text{3+M+L+H}}$ [9]*	–	–	–	76.6	40.4	124.7
	Mehta *et al.*$^{\text{3+M+L+H}}$ [34]*	84.6	72.4	69.7	75.7	39.3	117.6
	Mehta *et al.*$^{\text{3+M+L+C}}$ [31]*	–	–	–	75.2	37.8	122.2
	Luo *et al.*$^{\text{3+M+H}}$ [11], [77]	–	–	–	84.3	47.5	84.5
	Nibali *et al.*$^{\text{3+M}}$ [12]	–	–	–	87.6	48.8	87.6
	2.5D baseline$^{\text{3+M}}$	92.1	89.0	**87.7**	89.9±0.2	52.8±0.4	79.7±0.6
	MeTRo (ours)$^{\text{3+M}}$	**93.4**	**90.3**	86.5	**90.6**±0.4	**56.2**±0.5	**74.9**±1.4
Unnormalized	2.5D baseline$^{\text{3+M}}$	89.0	**87.9**	**89.4**	**88.7**±0.6	48.6±1.3	**87.1**±2.2
	MeTRo (ours)$^{\text{3+M}}$	**90.1**	87.8	85.7	88.2±0.5	**48.7**±0.7	88.4±1.3



Results: MuPoTS-3D

	A-MPJPE↓	MPJPE↓	A-PCK↑
Rogez *et al.* [74]	–	$146^{‡}$	–
Mehta *et al.* [31]	–	$132^{‡}$	–
Baseline in [39]	$320^{†}$	$122^{‡}$	–
Véges *et al.* [39]	$292^{†}$	$120^{‡}$	–
Véges *et al.* [75]*	257.2 ($255^{†}$)	119.4 ($108^{‡}$)	38.1
2.5D baseline	317.6 ($313.6^{†}$)	114.0 ($110.0^{‡}$)	**40.0**$\pm$1.0
MeTRAbs	**248.2 ($246.9^{†}$)**	**108.2 ($104.3^{‡}$)**	**40.2**$\pm$1.9
w/o abs. loss	328.8 ($327.8^{†}$)	108.4 ($104.7^{‡}$)	36.7$\pm$3.2



vision.rwth-aachen.de/metrabs

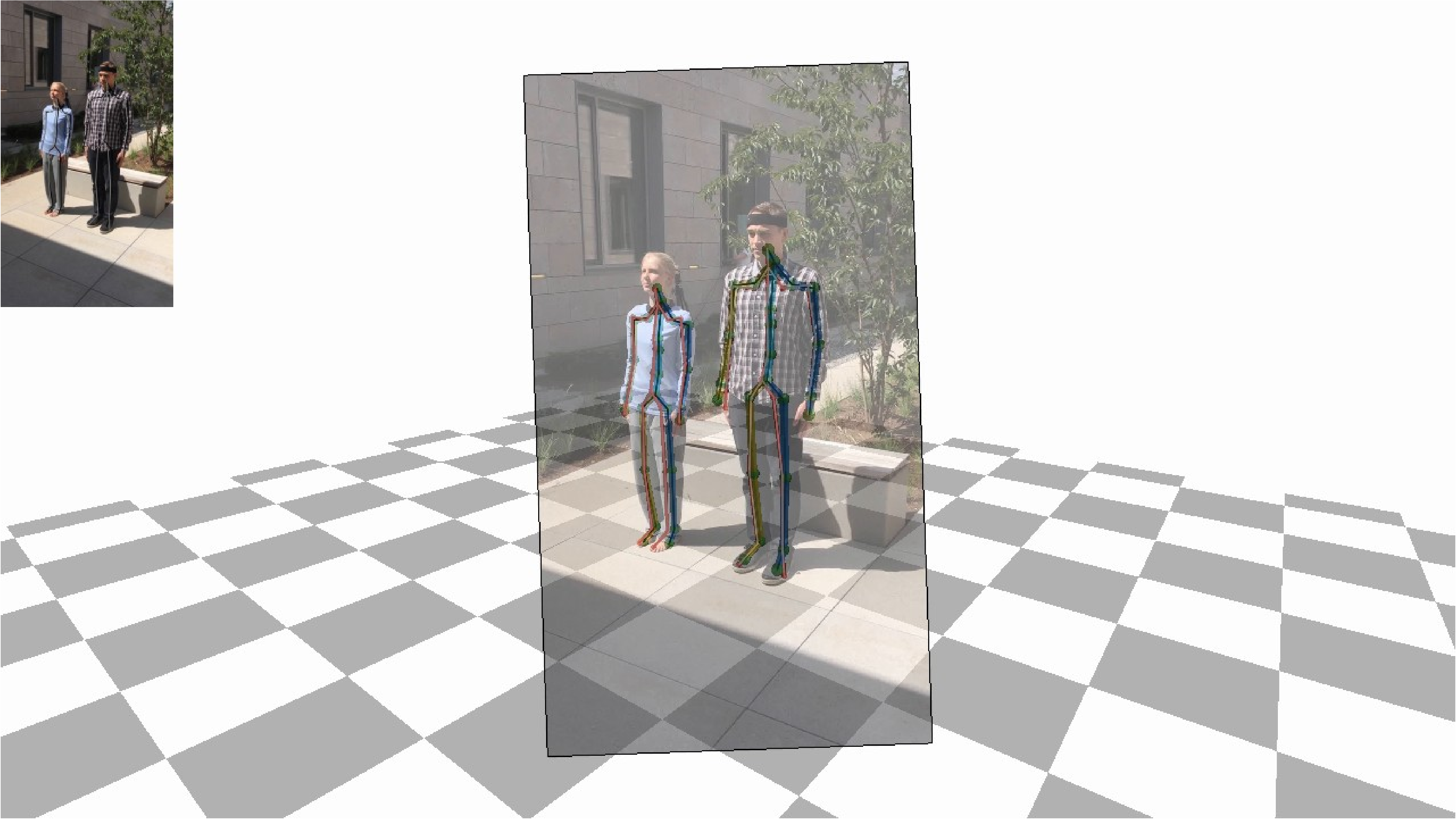
ECCV'20 3DPW Challenge Win

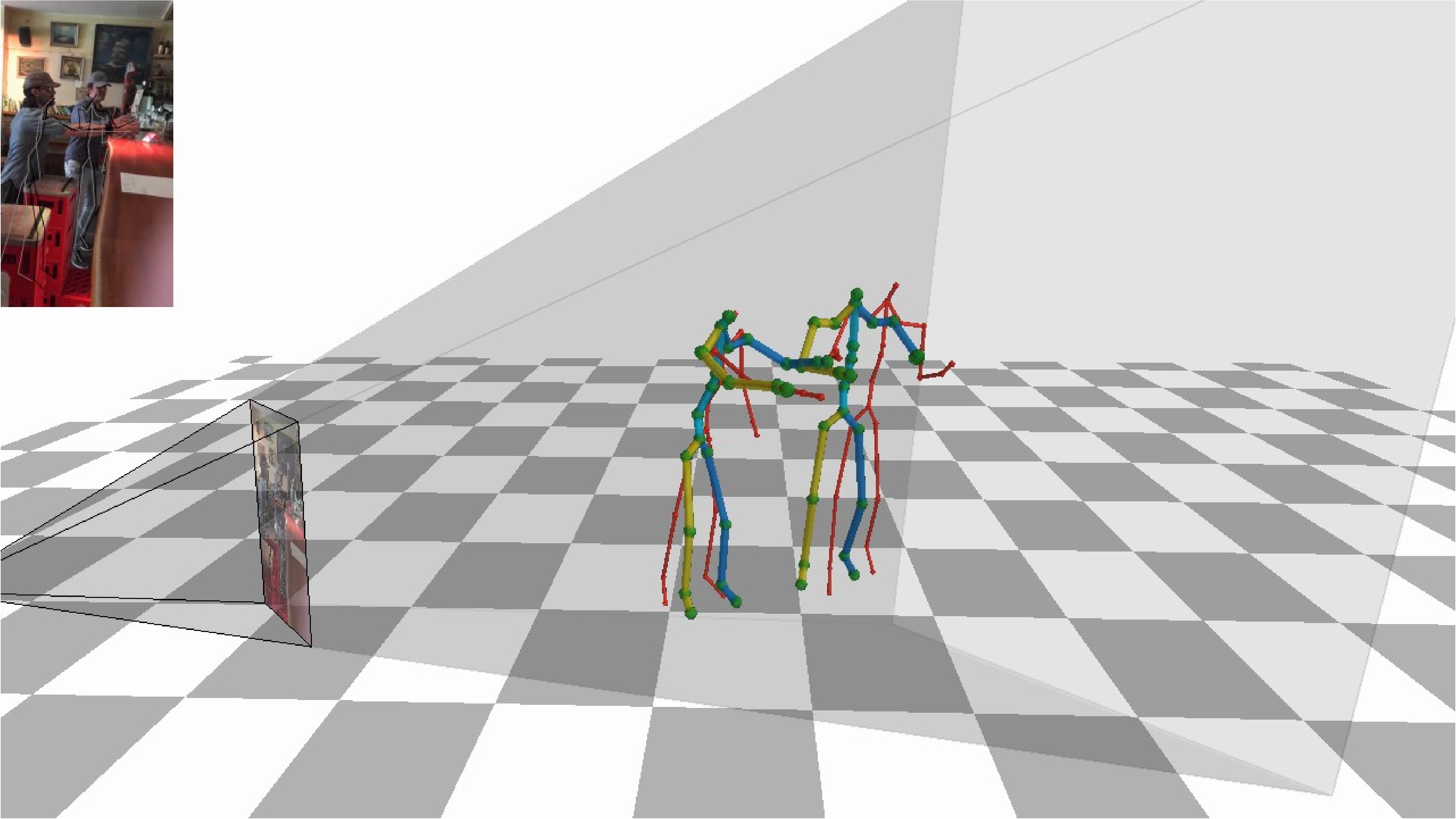
- Trained MeTRAbs on the union of many public datasets
- ResNet-101 backbone
- 5-crop test-time augmentation

competitions.codalab.org/competitions/24938#results

Results

#	User	Entries	Date of Last Entry	Team Name	Rank ▲	MPJPE ▲	MPJPE_PA ▲	PCK ▲	AUC ▲	MPJAE ▲	MPJAE_PA ▲
1	isarandi	3	09/29/20		1.0000	68.8397 (1)	49.6909 (1)	48.7720 (1)	0.6679 (1)	- (14)	- (14)
2	DJ_Walker	7	08/22/20	JDAI-CV	3.0000	81.7641 (2)	58.6131 (2)	37.3293 (4)	0.5991 (4)	20.8089 (3)	19.0901 (1)
3	milo	12	08/01/20	milo	3.2500	83.1544 (3)	59.7027 (4)	42.4194 (3)	0.6231 (3)	19.6965 (1)	19.1486 (2)
4	rbr	12	08/20/20		4.2500	83.1845 (4)	64.1717 (9)	46.9092 (2)	0.6323 (2)	20.1264 (2)	19.9578 (5)
5	mks0601	16	08/01/20	SNU CVLAB	6.0000	84.2889 (5)	61.7517 (6)	36.6064 (7)	0.5966 (6)	21.2543 (4)	19.7324 (4)
6	xuchen	8	08/01/20		4.7500	85.0523 (6)	59.3378 (3)	37.1122 (5)	0.5985 (5)	- (14)	- (14)
7	root9527	6	08/01/20		7.2500	85.7423 (7)	61.1041 (5)	36.1977 (9)	0.5915 (8)	21.5570 (5)	19.2689 (3)
8	Arthursy	22	07/30/20		8.0000	86.0644 (8)	63.1549 (7)	36.3558 (8)	0.5868 (9)	22.2771 (6)	20.5152 (6)
9	redarknight	15	08/02/20	SNU CVLAB	7.5000	86.3765 (9)	63.5519 (8)	36.7535 (6)	0.5932 (7)	23.5012 (7)	21.1888 (9)





Inference Speed

		Test stride			
		32	16	8	4
Speed (crop per sec.)	no batching	160	150	105	38
	batch size 8	**511**	475	292	92

Using a single Nvidia RTX 2080Ti consumer GPU
Person detection not included

Publicly Available for TensorFlow 2!

vision.rwth-aachen.de/metrabs

In [1]:

```python
import tensorflow as tf

image = tf.image.decode_jpeg(tf.io.read_file('./test_image.jpg'))
intrinsic_matrix = tf.convert_to_tensor([[1030, 0, 980], [0, 1030, 550], [0, 0, 1]], tf.float32)
person_detections = tf.convert_to_tensor([[621, 238, 204, 658], [932, 207, 250, 783]], tf.float32)

metrabs = tf.saved_model.load('./metrabs_fullimage_smpl_model')
metrabs(image, intrinsic_matrix, person_detections)
```

Out[1]:

```
<tf.Tensor: shape=(2, 24, 3), dtype=float32, numpy=
array([[[-762.3343  ,   65.86958 , 2772.7231  ],
        [-654.5931  , -140.52255 , 2762.523   ],
        [-552.06177 ,  433.45905 , 2763.0042  ],
        [-780.02216 ,  447.97833 , 2869.2644  ],
        [-654.70984 , -286.88297 , 2760.9946  ],
        [-552.76086 ,  817.2549  , 2812.7212  ],
        [-769.5841  ,  807.46014 , 2946.1006  ],
        [-657.1698  , -352.63895 , 2745.4973  ],
        [-574.6483  ,  868.4689  , 2681.4917  ],
        [-867.92957 ,  856.29315 , 2861.1292  ],
        [-640.55084 , -559.65436 , 2708.677   ],
        [-565.3983  , -457.15265 , 2677.2026  ],
        [-725.0487  , -470.16095 , 2770.3647  ],
        [-668.6167  , -628.4125  , 2660.2937  ],
        [-480.72577 , -431.4521  , 2620.2185  ],
        [-821.40784 , -455.7679  , 2817.3267  ],
        [-458.43857 , -206.46533 , 2603.1274  ],
        [-868.421   , -225.19409 , 2893.397   ],
        [-483.9546  ,   15.832471, 2525.8289  ],
        [-914.83575 ,   13.147165, 2871.1206  ],
        [-477.6687  ,   84.54141 , 2482.779   ],
        [-932.82355 ,   85.472694, 2861.5317  ],
```

Summary

- End-to-end learned scale-recovery (metric output)
- Express everything as heatmaps
- Guess joints outside the input crop (truncation-robustness)
- No focal length needed for (root-relative) metric output
- Fast and simple architecture (up to 511 crops per second)
- Extension to absolute pose with differentiable root joint reconstruction
- State-of-the-art results on Human3.6M, MPI-INF-3DHP, MuPoTS-3D
- 1st place at the ECCV2020 3D Poses in the Wild Challenge

Extended Journal Version

I. Sárándi, T. Linder, K. O. Arras, B. Leibe:

MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation

In: “The Best of FG” Special Issue, IEEE Trans. Biometrics, Behavior, and Identity Science (2020)

Conference Version

I. Sárándi, T. Linder, K. O. Arras, B. Leibe:

Metric-Scale Truncation-Robust Heatmaps for 3D Human Pose Estimation

In: IEEE Int. Conf. Automatic Face and Gesture Recognition (FG) (2020)

Thank you!

sarandi@vision.rwth-aachen.de

István Sárándi

Timm Linder

Kai O. Arras

Bastian Leibe