3D Poses in the Wild (3DPW) Challenge
Winning Approach
ECCV 2020

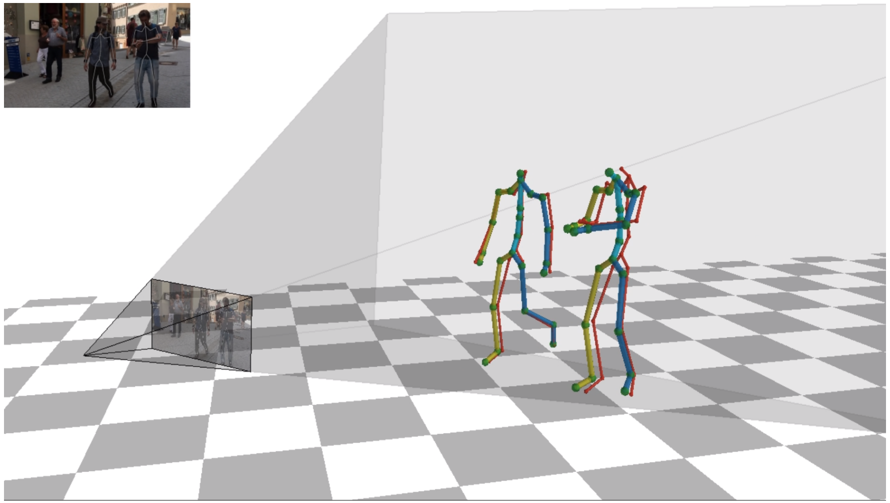
Presented by István Sárándi (RWTH Aachen University)

Based on the methods from

- I. Sárándi, T. Linder, Kai O. Arras, B. Leibe: “*Metric-Scale Truncation-Robust Heatmaps for 3D Human Pose Estimation*”, IEEE Conf. Automatic Face and Gesture Recog. (FG) 2020
- I. Sárándi, T. Linder, Kai O. Arras, B. Leibe: “*MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation*”, ArXiv preprint 2020

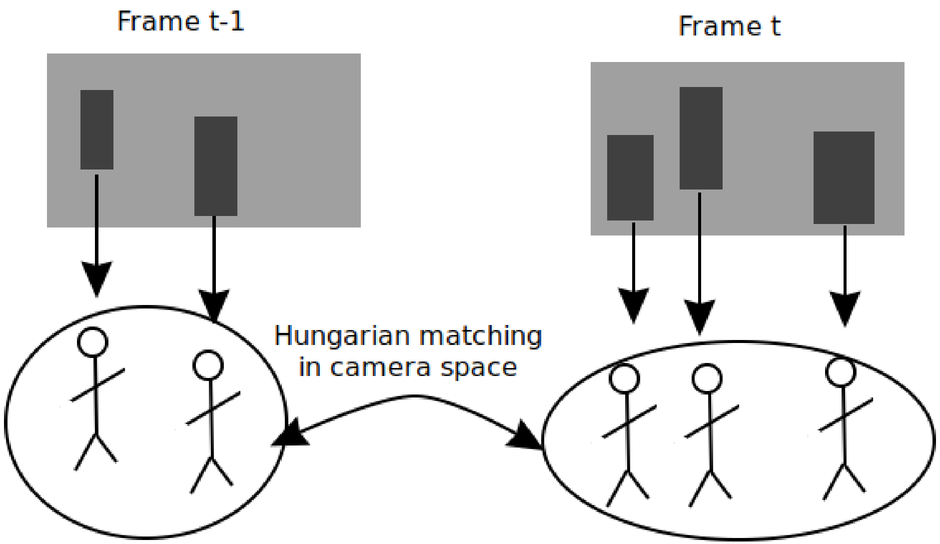
Task: 3D Human Pose Tracking

- Given an RGB video, estimate the 3D pose of each annotated person in each frame
- Two variations:
  - Known association: Estimated poses can be matched to 2D GT in each frame
  - Unknown association: IDs can be matched only in the first frame (using 2D reference), to determine who to track



Our Approach

- Detect people per frame with an off-the-shelf detector (YOLOv3)
- **Estimate absolute 3D pose for each detection (MeTRAbs)**
- Associate based on pose distance (naive frame-to-frame tracking)
  - Either to the predicted poses of the previous frame (in camera coords)
  - Or to reference 2D pose (if allowed)

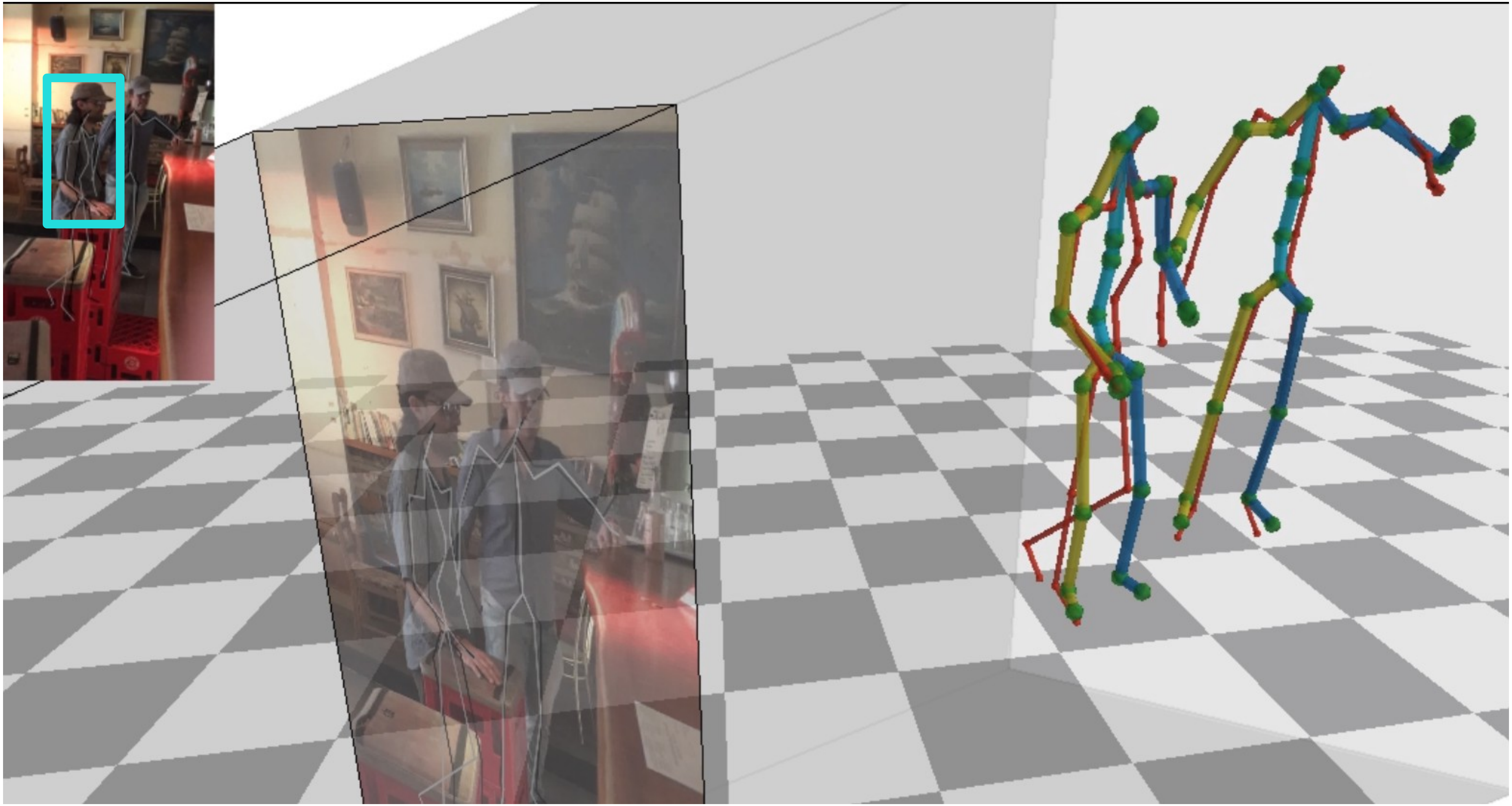


Our 3D Pose Estimator

- Volumetric heatmaps are a powerful representation
  - Introduced in [Pavlakos et al. ‘17] with a coarse-to-fine estimation scheme
  - Combined in [Sun et al. ‘18] with soft-argmax heatmap decoding
  - Our contributions
    - Further simplification: directly predict low-res (8x8) heatmap with a standard backbone (e.g. ResNet), no need for any additional learned layers
    - Define the heatmap axes in a novel way...

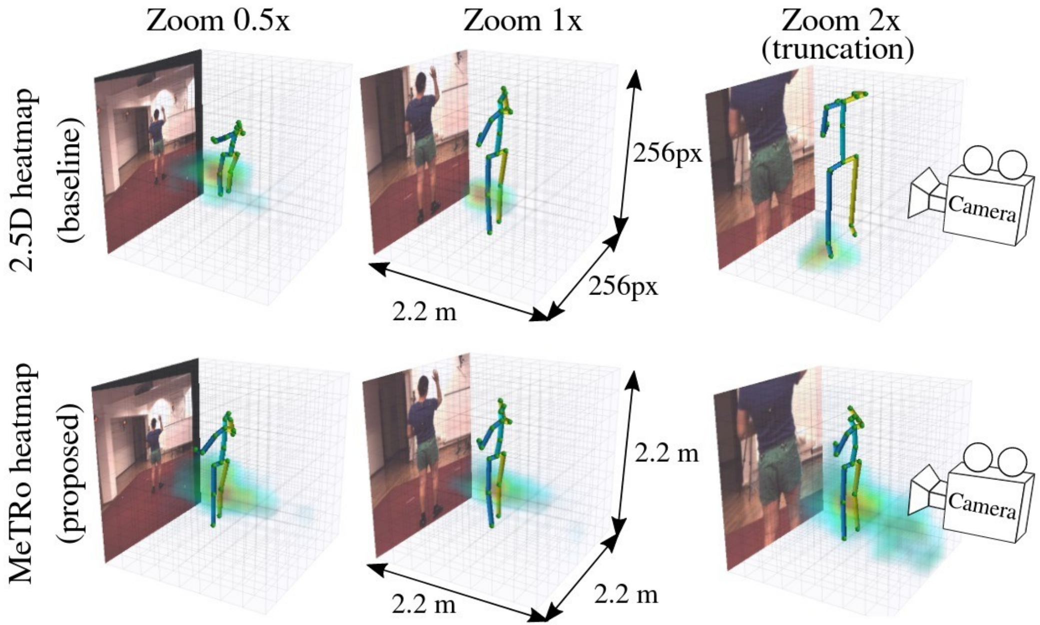
Metric-Scale Truncation-Robust Heatmaps

- Limitations of the (2.5D) heatmap approach:
  - **Scale recovery:** The X and Y coordinates are in image (pixel) space, but we want metric-scale predictions in millimeters (common: non-learned post-processing step)
  - **Truncation:** No way to estimate body joints outside the image boundaries
- Direct numerical coordinate regression would not suffer from these problems
- However, heatmap prediction fits better to the convolutional structure and leads to more accurate localization
- Question: Can we have the best of both?
  - Recover a full metric-scale pose while staying in the heatmap paradigm?



Metric-Scale Truncation-Robust (MeTRo) Heatmaps

- Define the heatmap axes directly in the 3D metric space, irrespective of image zooming



Metric-Scale Truncation-Robust Heatmaps

- Benefit:
  - No need for heuristic scale recovery, the backbone is trained to estimate people’s size implicitly
  - We always get a complete 3D pose
- Downside:
  - No 2D image space pose, only 3D root-relative
  - As heatmap peaks are not predicted at their image space position, localization might be slightly less precise

Metric-Scale Truncation-Robust Heatmaps

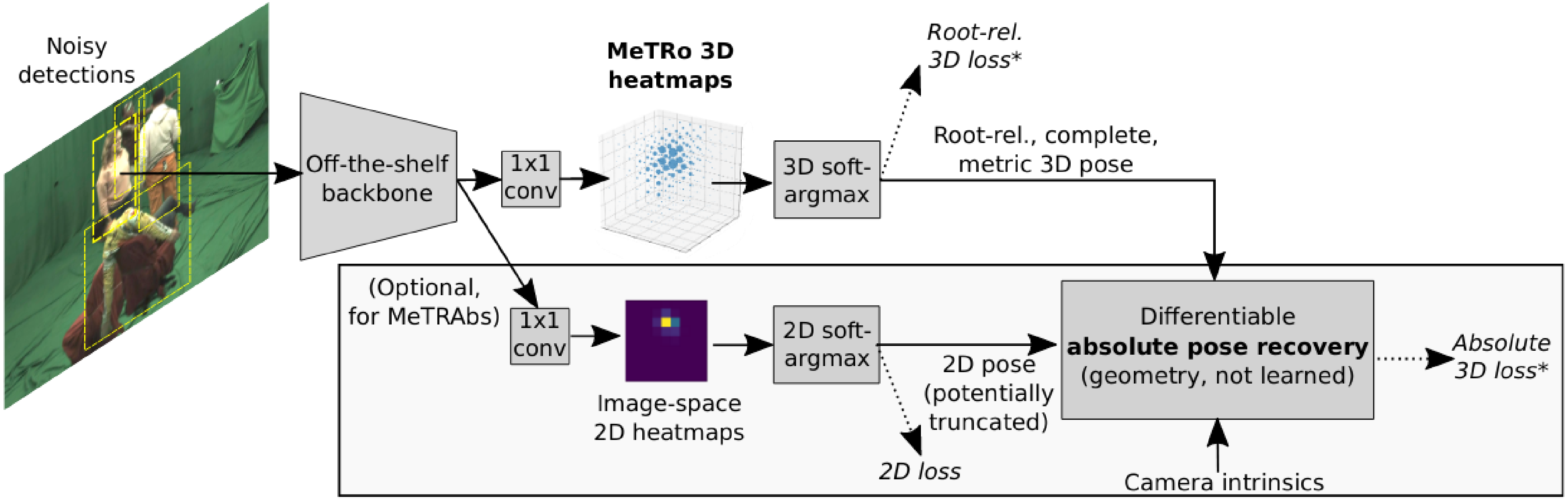
- How can this even work if input and output are not aligned?
  - Receptive fields are large enough (signal can move to new position as we go deeper in the network)
  - Backbone learns a scaling transformation
  - Truncation is sensed by the network through the zero padding at the borders (c.f. [1])

[1] Islam et al.: "*How much position information do convolutional networks encode?*" ICLR 2020

MeTRAbs

- Now we can recover **3D metric-scale complete** poses, without having to give up the heatmap idea
- But we lost the alignment to the image space
- However, we can jointly estimate **both 2D and 3D** heatmaps:
  - negligible extra computational cost: 8*J → 9*J channels
  - if the camera is calibrated, we can even recover the camera-space (<u>absolute</u>) 3D coordinates!

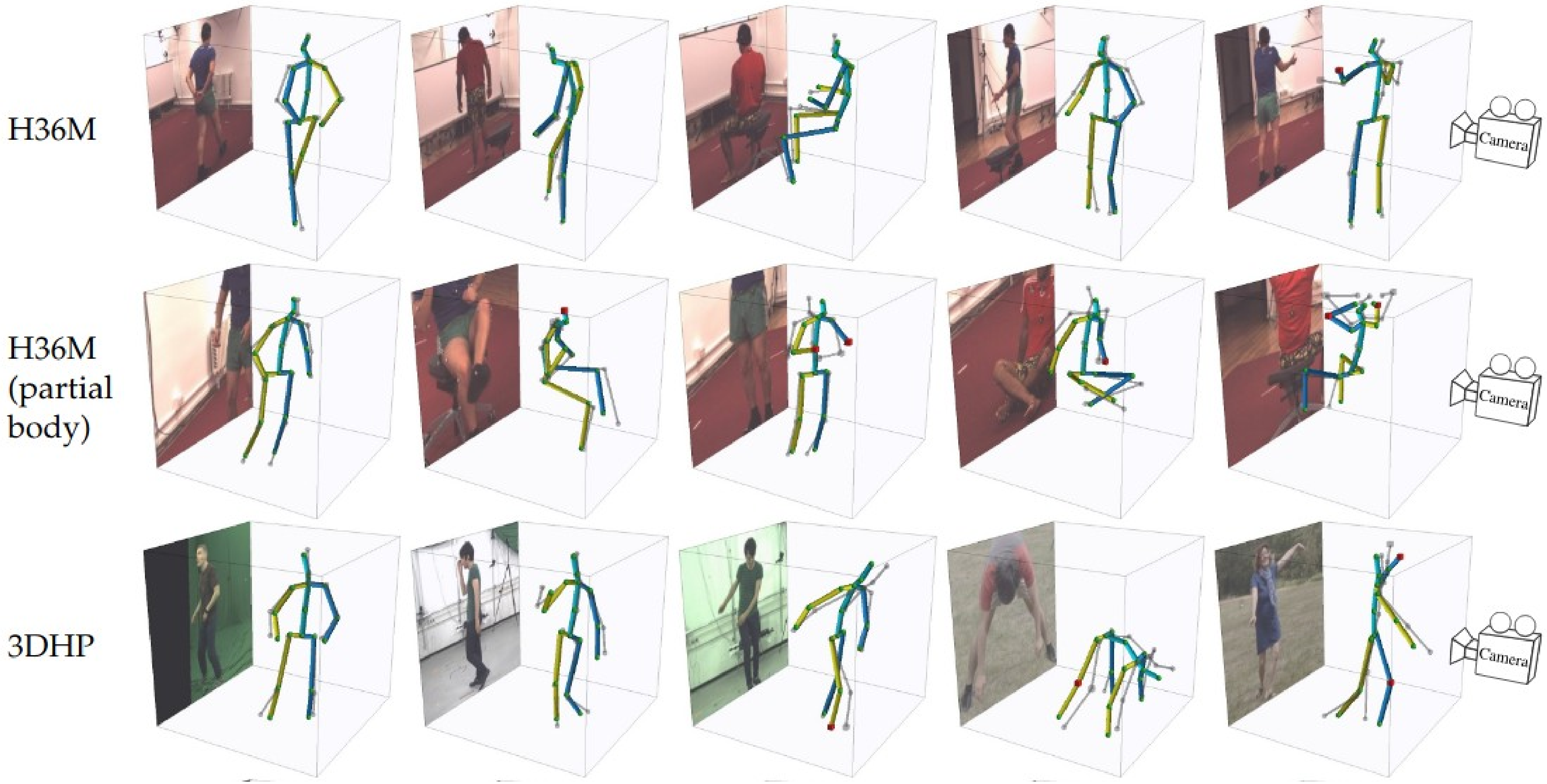
MeTRAbs Architecture



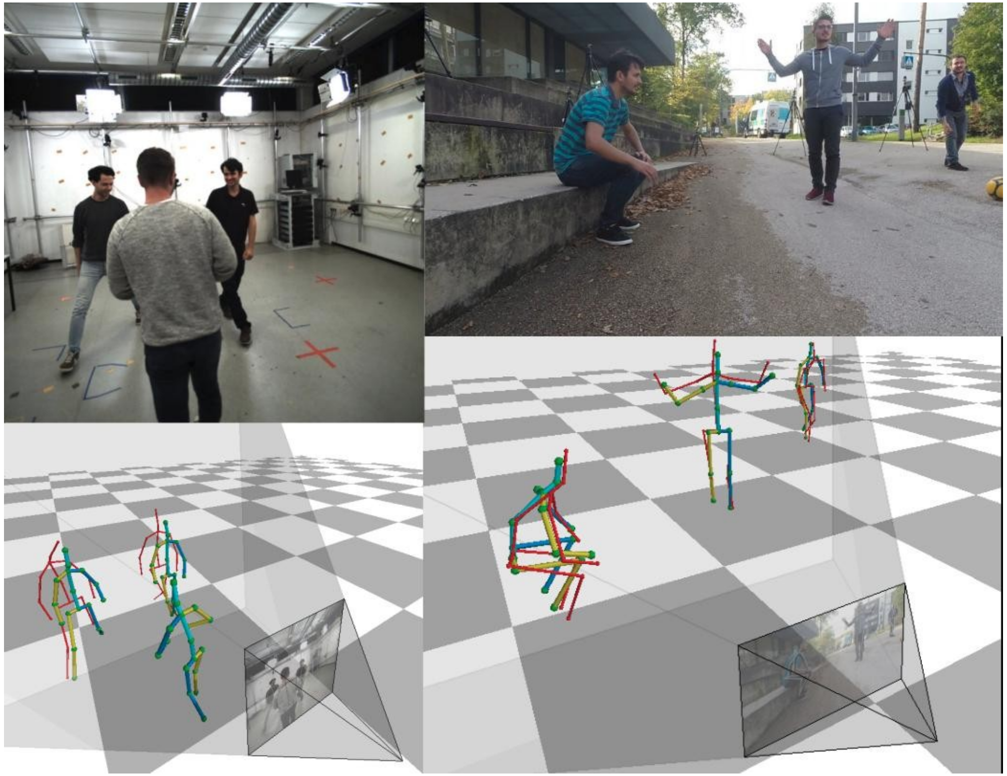
MeTRAbs

- Despite its simplicity, state-of-the-art results on
  - Human3.6M
  - MPI-INF-3DHP
  - MuPoTS-3D
- Fast execution:
  - ~500 crops per second (ResNet-50, 256x256 px, stride 32, batch size 8, 2080Ti)

Qualitative results



Qualitative Results (MuPoTS-3D)

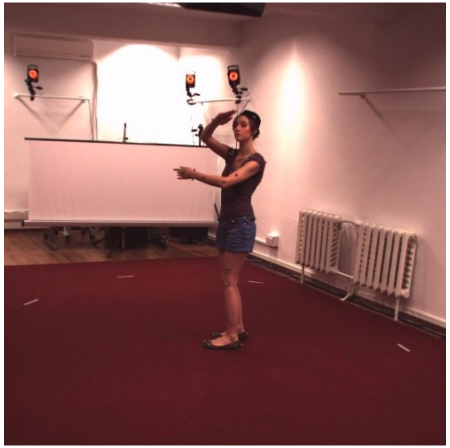


Training Data for the Challenge

- Many 3D pose datasets released in recent years
- **Large-scale supervised learning** is thus possible
  - Human3.6M – 164,528 [number of sufficiently different poses (thresh. 100 mm)]
  - MuCo-3DHP – 676,875 [includes repeated poses in different composites]
  - CMU-Panoptic – 858,390
  - SURREAL – 1,577,006
  - SAILVOS – 90,611
- For better generalization, **weak supervision** from 2D datasets is also important
  - COCO, MPII, LSP, …

Dataset Merging

- Learn jointly from all datasets with mixed batches:
  - 36 examples from the real 3D datasets (H36M, MuCo, CMU)
  - 12 examples from SAILVOS
  - 8 examples from SURREAL
  - 8 examples from COCO (2D only)
- All datasets use somewhat different joint definitions
- 3DPW benchmark requires SMPL body joints (24 keypoints)
- Goal: use all available supervision but keep final output specific to SMPL
  - Merge joints across datasets that are sufficiently similar (e.g. wrists)
  - Do not merge others (e.g. define multiple hip joints, one per dataset)
  - We use a total of 73 distinct joints in the model
  - Some SMPL joints are only supervised through SURREAL



Human3.6M

MuCo-3DHP

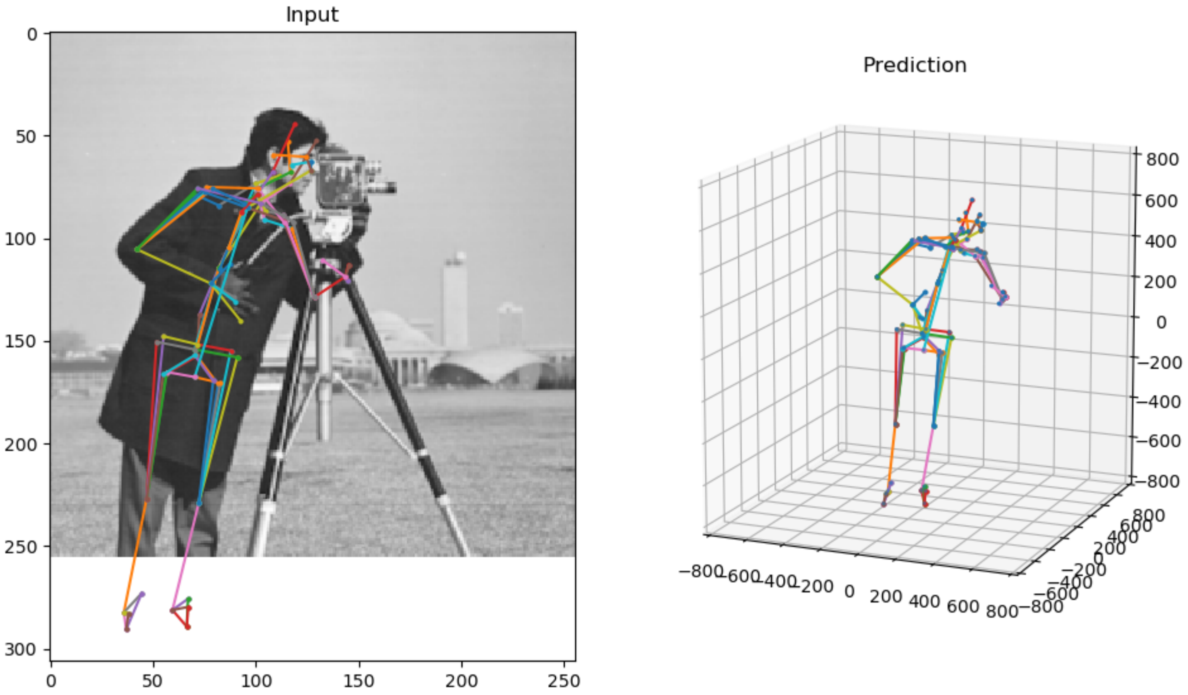
CMU-Panoptic

SURREAL

SAILVOS

Dataset Merging

- Redundant joints, but allows dataset-specific benchmarking



Strong Data Augmentation

- Crucial for generalization to in-the-wild scenes
  - Synthetic occlusions (paste object segments onto the image)
  - Background replacement (segment the training images beforehand if masks not given in dataset)
  - Color distortion
  - Geometric transformations (scale, flip, shift, rotate)
- Test-time augmentation: 5 crops

Results

Known association

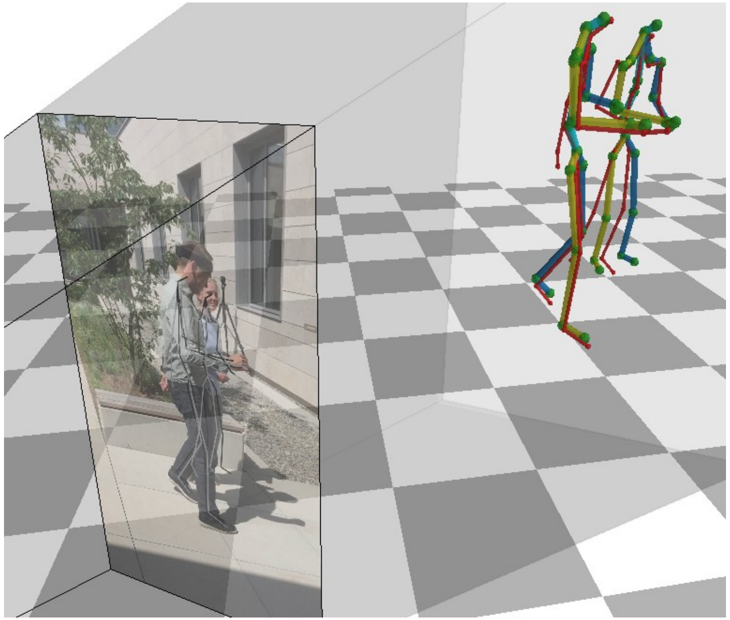
#	User	Entries	Date of Last Entry	Team Name	Rank	MPJPE	MPJPE_PA	PCK	AUC	MPJAE	MPJAE_PA
1	isarandi	2	08/01/20		1.0000	72.3768 (1)	53.0564 (1)	47.3431 (1)	0.6624 (1)	- (14)	- (14)
2	DJ_Walker	7	08/22/20	JDAI-CV	3.0000	81.7641 (2)	58.6131 (2)	37.3293 (4)	0.5991 (4)	20.8089 (3)	19.0901 (1)
3	milo	12	08/01/20	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)

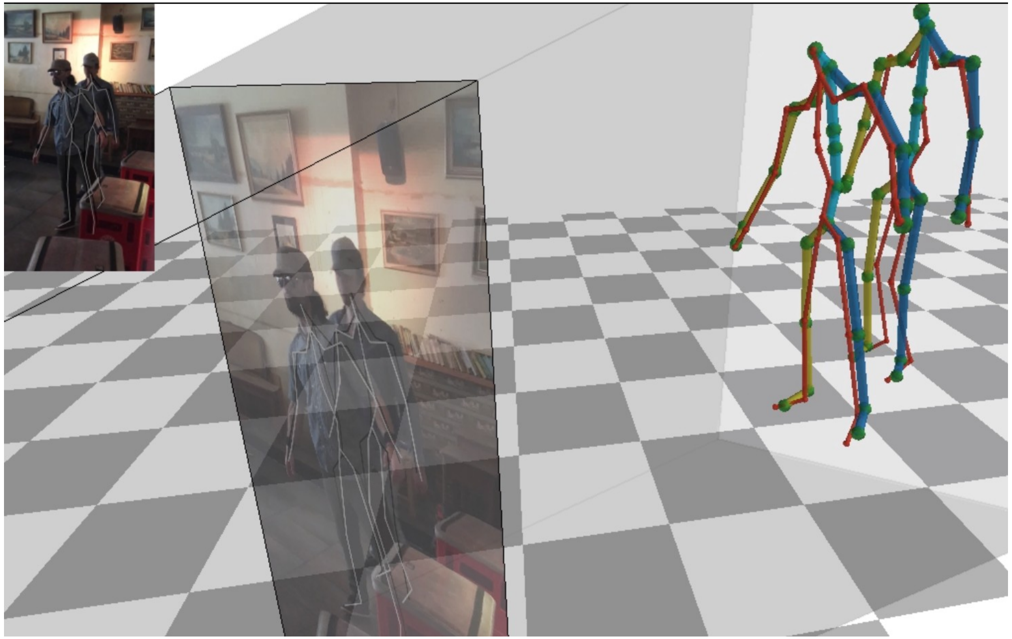
Unknown association

#	User	Entries	Date of Last Entry	Team Name	Rank	MPJPE	MPJPE_PA	PCK	AUC	MPJAE	MPJAE_PA
1	isarandi	2	08/02/20		1.0000	83.3594 (1)	58.8521 (1)	44.6241 (1)	0.6315 (1)	- (9)	- (9)
2	mks0601	12	08/01/20	SNU CVLAB	2.7500	86.3566 (2)	63.1444 (3)	36.2311 (4)	0.5908 (2)	22.2005 (1)	20.3347 (2)
3	root9527	4	08/02/20		3.7500	87.0182 (3)	61.4503 (2)	36.0264 (5)	0.5888 (5)	22.2413 (2)	19.3888 (1)
4	redarknight	16	08/01/20	SNU CVLAB	3.7500	87.8796 (4)	64.2319 (4)	36.4411 (3)	0.5890 (4)	24.2251 (4)	21.3447 (4)
5	zenluo	2	08/02/20		5.5000	91.5124 (5)	67.6933 (5)	33.9416 (6)	0.5655 (6)	23.5770 (3)	20.6931 (3)

Results

- Robust predictions even under occlusion and bad illumination





Discussion

- Overall recipe:
  - Formulate the task such that a CNN <u>can</u> predict the desired output (here: metric-space complete pose)
  - In a well-suited <u>representation</u> (here: heatmap)
  - Then <u>supervise</u> it with strongly augmented, diverse examples from many sources by carefully merging datasets
- Limitations:
  - Single-frame estimator, no temporal smoothing
  - Needs a separate person detector
  - Naive pose matching, no ReID (still not many ID switches on 3DPW)

Thank you!

- Inference code and pretrained model available (self-contained model file → just a few lines of code to run!)
  - https://github.com/isarandi/metrabs
- Also thanks to my co-authors of the underlying papers!

István Sárándi

Timm Linder

Kai O. Arras

Bastian Leibe