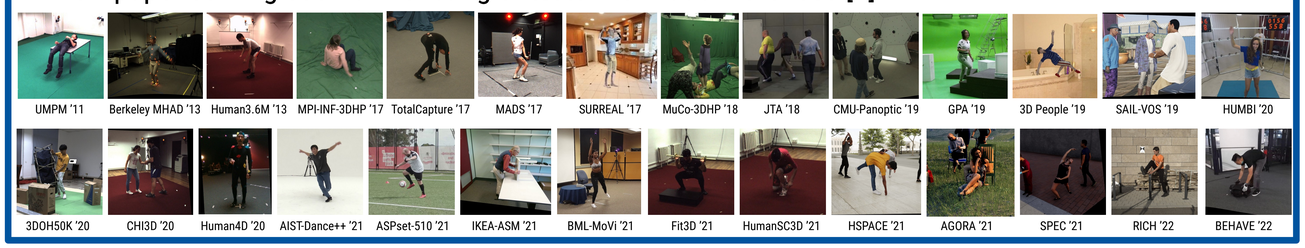
Demo: Accurate and Efficient Absolute 3D Human Pose Estimation Trained on Dozens of Datasets

István Sárándi, Alexander Hermans, Bastian Leibe (Computer Vision Group, RWTH Aachen University, Germany)

Overview

- Task: **monocular, single-image, RGB-based 3D human pose estimation**
  - Goal: build accurate and efficient off-the-shelf model
  - *...which is also easy-to-use* for non-specialist downstream researchers!
- More data: Deep learning has shown excellent scaling ability
  - 3D pose data is expensive to collect
  - But a lack of data is less of an issue today than often imagined!
  - Lots of labeled data available — *just in separate datasets.* Merge them!
- Base method: MeTRAbs [1], our ECCV'20 3DPW Challenge winner, but:
  - Improved handling of skeleton annotation format discrepancy
  - EfficientNetV2 backbone instead of ResNet
  - YOLOv4 detector instead of YOLOv3
  - 28 training datasets combined instead of 5 (+4 with 2D annot.)
  - Crop resolution 384×384 px instead of 256×256
- New paper coming soon on handling **skeleton format differences** [2]

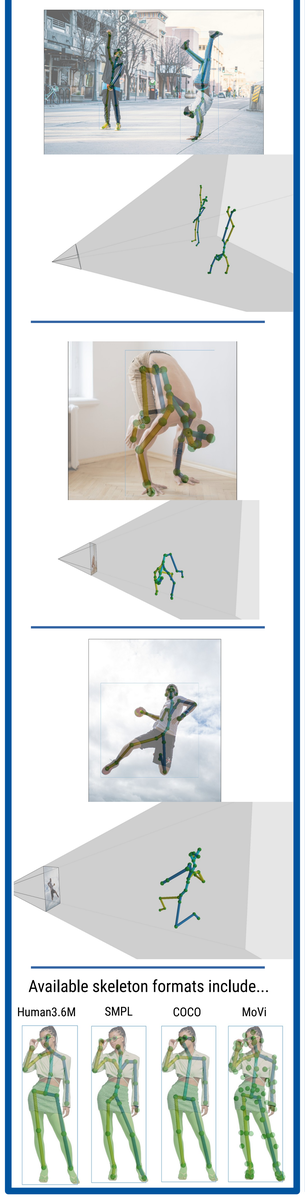
Dataset name	#Examples	#Real Subj.	#Keypoints
Real images with markerless MoCap			
MuCo-3DHP (Mehta *et al.*, 2018)	677k	8	28
CMU-Panoptic (Joo *et al.*, 2019)	2.81M	>60	19
AIST-Dance++ (Tsuchida *et al.*, 2019; Li *et al.*, 2021)	1.86M	30	19
HUMBI (Yu *et al.*, 2020)	1.26M	772	19
MPI-INF-3DHP (Mehta *et al.*, 2017)	627k	8	28
RICH (Huang *et al.*, 2022)	96k	15	42
BEHAVE (Bhatnagar *et al.*, 2022)	42k	7	43
ASPset (Nibali *et al.*, 2021)	124k	15	17
3DOH50K (Zhang *et al.*, 2020)	50k	<10	14
IKEA ASM (Ben-Shabat *et al.*, 2021)	23k	48	17
Real images with marker-based MoCap			
Human3.6M (Ionescu *et al.*, 2013)	165k	5	25
TotalCapture (Trumble *et al.*, 2017)	130k	5	21
BML-MoVi (Ghorbani *et al.*, 2021)	553k	13	87
Berkeley-MHAD (Ofli *et al.*, 2013)	526k	12	43
UMPM (Aa *et al.*, 2011)	164k	30	15
Fit3D (Fieraru *et al.*, 2021a)	147k	8	25
GPA (Wang *et al.*, 2019)	109k	13	34
HumanSC3D (Fieraru *et al.*, 2021b)	72k	4	25
CHI3D (Fieraru *et al.*, 2020)	46k	6	25
Human4D (Chatzitofis *et al.*, 2020)	40k	4	32
MADS (Zhang *et al.*, 2017)	33k	5	15
Synthetic images			
SURREAL (Varol *et al.*, 2017)	1.9M		24
3DPeople (Pumarola *et al.*, 2019)	946k		29
JTA (Fabbri *et al.*, 2018)	562k		22
HSPACE (Bazavan *et al.*, 2021)	195k		35
SAIL-VOS (Hu *et al.*, 2019)	101k		26
AGORA (Patel *et al.*, 2021)	79k		66
SPEC (Kocabas *et al.*, 2021)	59k		24
Real images with 2D annotations (weak supervision)			
COCO (Lin *et al.*, 2014)	47k		17
MPII (Andriluka *et al.*, 2014)	27k		16
PoseTrack (Andriluka *et al.*, 2018)	40k		15
JRDB (Martin-Martin *et al.*, 2021)	59k		17
Totals (for 3D-labeled data)			
GRAND TOTAL (28 ds.)	13.4M	>1k	555



Quantitative Evaluation

	MuPoTS	3DPW			MPI-INF-3DHP		Human3.6M	FPS on MuPoTS (≤ 3 people/frame)			
								Desktop (3090)		Laptop (2080)	
	PCK_{150}↑	MPJPE↓	PMPJPE↓	PCK_{50}↑	MPJPE↓	PCK_{150}↑	MPJPE↓	Batched	Non-B.	Batched	Non-B.
ROMP (Sun *et al.*, 2021)	–	80.1	56.8	36.5	–	–	–				
Lin *et al.* (2021b)	–	74.7	45.6	–	–	–	51.2				
PoseAug (Gong *et al.*, 2021)	–	–	–	–	71.1	89.2	50.2				
Cheng *et al.* (2022)	89.6	–	–	–	–	–	49.3				
Ours EffV2S	94.9	59.5	41.0	53.1	58.7	96.2	41.4	63	28	25	11
Ours EffV2S 5-crop	95.2	58.9	39.9	53.6	57.5	96.7	40.1	28	18	9	6
Ours EffV2L	95.4	58.9	39.5	53.9	55.4	97.1	36.5	42	19	12	6
Ours EffV2L 5-crop	95.7	57.0	38.1	55.4	53.6	97.6	35.5	14	10	2	2

Qualitative Results



You can run it yourself!

bit.ly/metrabs_demo

Open in Colab

Get an RGB image...

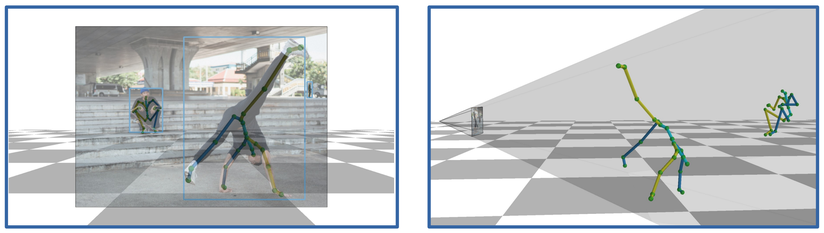
... run a few lines of code

```
import tensorflow as tf
import tensorflow_hub as hub

model = hub.load('https://bit.ly/metrabs_l')
img = tf.image.decode_image(tf.io.read_file('test.jpg'))
pred = model.detect_poses(img, skeleton='smpl_24')
pred['poses3d'].shape
```

```
TensorShape([2, 24, 3])
```

3D poses in camera space!



- **Batteries included!** One TensorFlow SavedModel file has:
  - Built-in person detector (YOLOv4)
  - Perspective undistortion, lens undistortion (if calibration known)
  - Cropping, test-time augmentation, batching
  - Output sanity checking, 3D-pose-based non-max suppression
  - Choice among 23 different skeleton/landmark formats
  - All behind a simple API, no dependency besides TF itself
- *Note: trained models are only for non-commercial research use!*

[1] Sárándi, I., Linder, T., Arras, K. O., Leibe, B. (2021). MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation. In *IEEE T-BIOM Special Issue Selected Best Works from FG 2020.*
[2] Sárándi, I., Hermans, A., Leibe, B. (2023). [Title not yet available] In *WACV* (to appear)

Visualizer tool: github.com/isarandi/poseviz

Acknowledgments. This work was supported by the ERC Consolidator Grant project "DeeViSe" (ERC-CoG-2017-773161) and by Robert Bosch GmbH under the project "Context Understanding for Autonomous Systems".