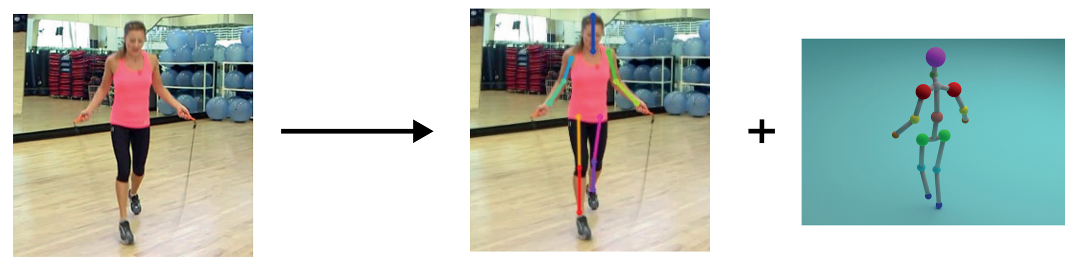
Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image

Denis Tomé[1] Chris Russell[2,3] Lourdes Agapito[1]
[1]University College London [2]University of Surrey [3]Alan Turing Institute

Problem: 3D Human Pose Estimation from a Single Image

INPUT — RGB Image; OUTPUT — 2D pose + 3D pose

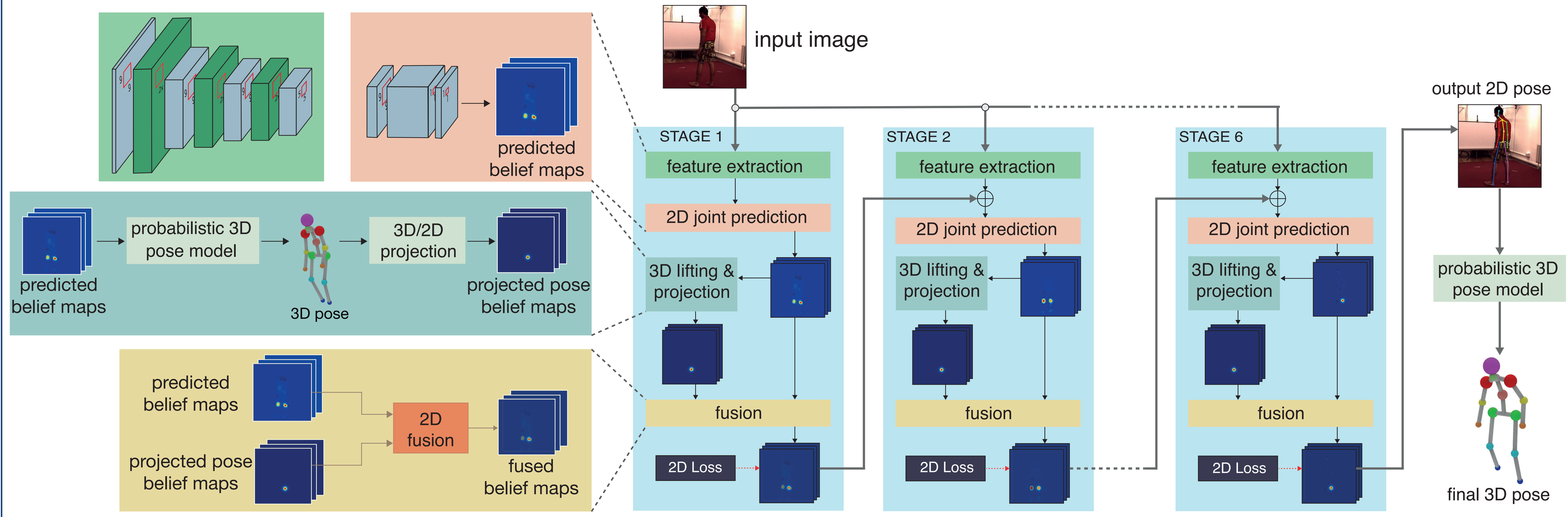


No temporal information

2 **independent** sources of training data

2D pose annotations 3D mocap

End-to-end Architecture

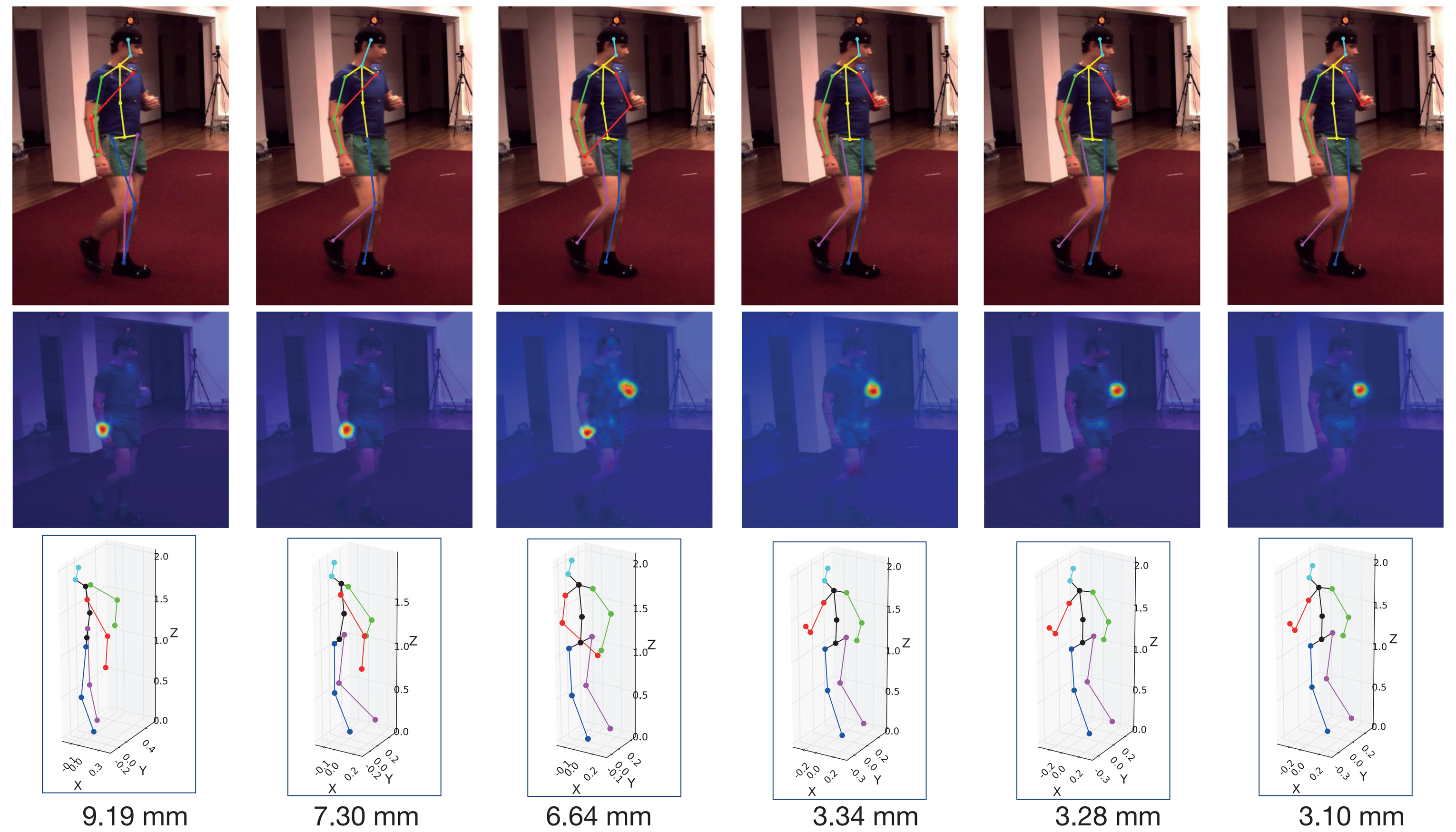


- Multi-stage deep architecture for joint 2D + 3D human pose estimation with our **3D pose layer.**

2D Pose

Belief map (left hand)

Reconstructed 3D Pose



- The accuracy of both 2D and 3D landmark locations improves progressively through the stages.

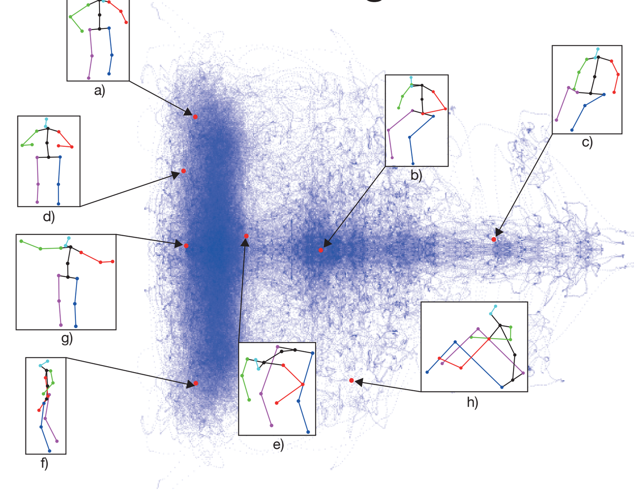
Training the Probabilistic 3D Human Pose Model

First step: aligning the data

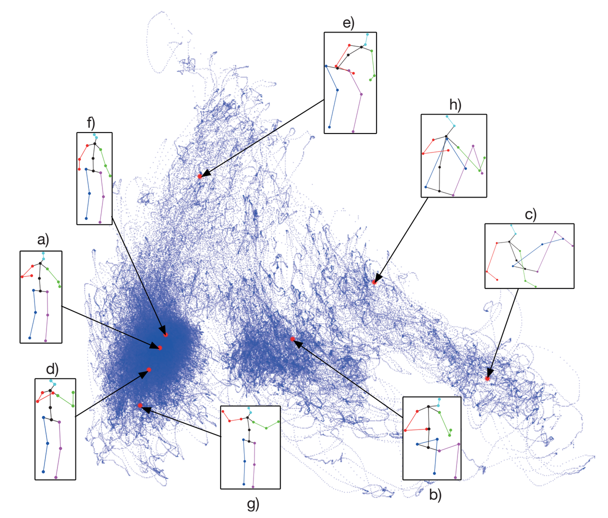
We seek the optimal rotations for each pose such that after rotating the poses they are closely approximated by a low-rank compact Gaussian distribution.

$$\underset{\mathbf{R},\mu,a,\mathbf{e},\sigma}{\arg\min} \sum_{i=1}^{N} (||\mathbf{P_i} - \mathbf{R_i}\,(\mu + a_i \cdot \mathbf{e})\,||_2^2 + \sum_{j=1}^{J} (a_{i,j} \cdot \sigma_j)^2 + \ln \sum_{j=1}^{J} \sigma_j^2)$$

Before alignment



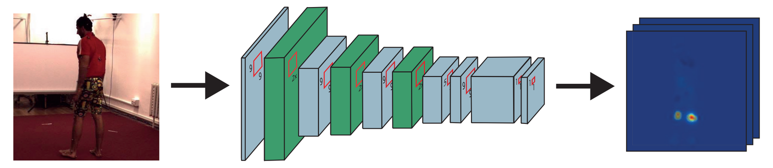
After alignment



Second step: train a mixture of PPCA models

Architecture of Each Stage

1. From image to belief maps

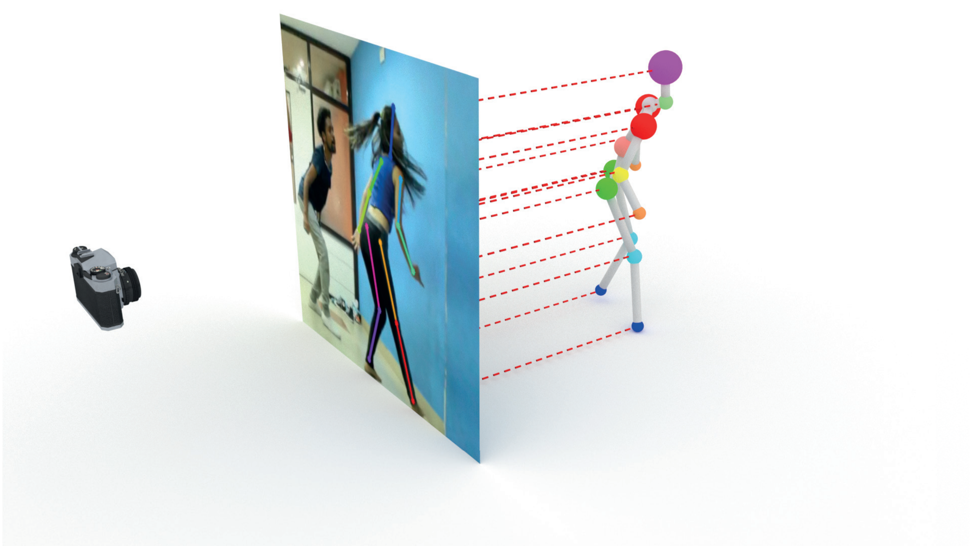


2. From belief maps to 2D pose

$$Y_p = \underset{(u,v)}{\arg\max}\, b_p[u,v]$$

3. From 2D to 3D pose

$$\underset{R,a}{\arg\min} ||Y - s\Pi ER(\mu + a \cdot \mathbf{e})||_2^2$$
$$+ ||\sigma \cdot a||_2^2$$



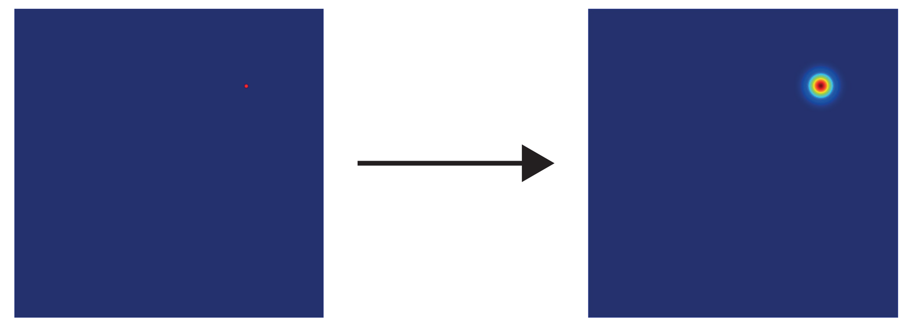
4. From 3D back to 2D

$$\hat{Y} = s\Pi ER(\mu + a \cdot \mathbf{e})$$

5. Generate projected belief maps

$$\hat{b}_{i,j}^p = \begin{cases} 1 & \text{if}(i,j) = \hat{Y}_p \\ 0 & \text{otherwise.} \end{cases}$$

followed by convolution with Gaussian filter



6. Fusion

$$f_t^p = w_t * b_t^p + (1 - w_t) * \hat{b}_t^p$$

weights learned in the end-to-end learning process

Quantitative Results on Human3.6M Dataset

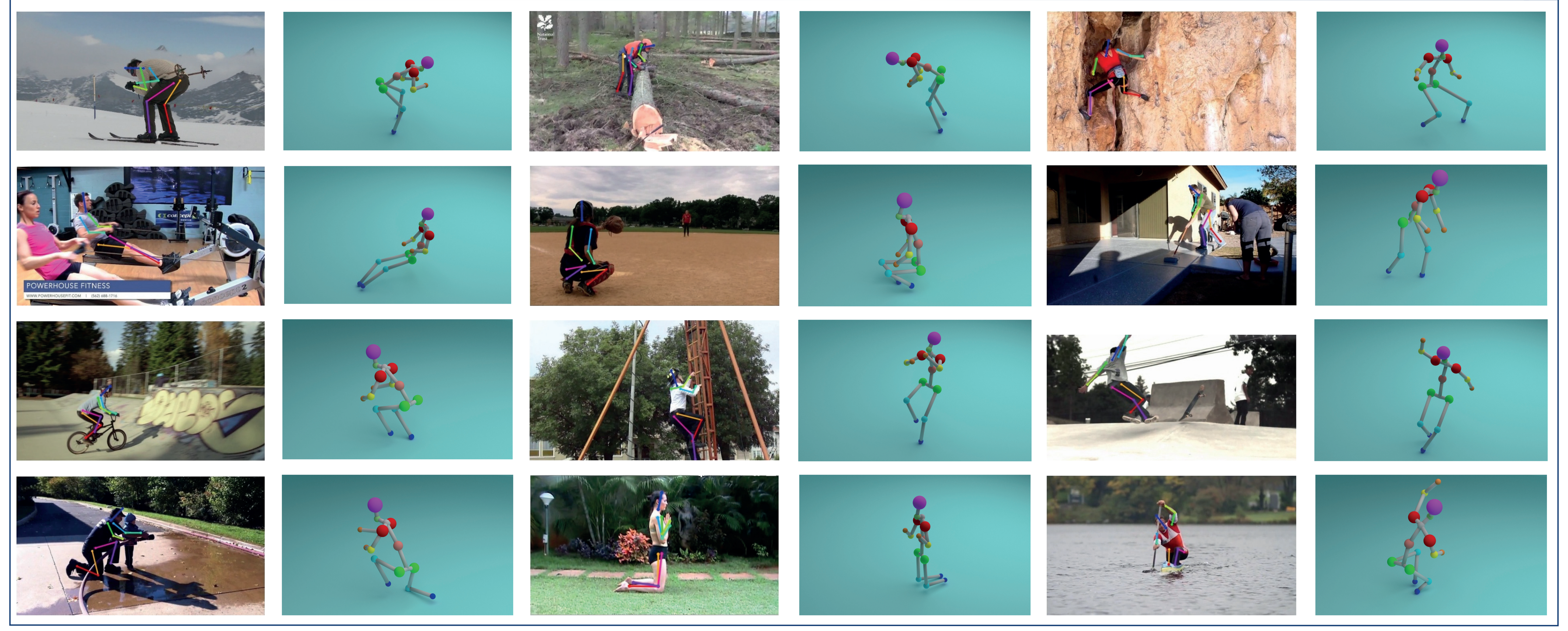
	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Tekin *et al.* [2]	85.03	108.79	84.38	98.94	119.39	**95.65**	98.49	93.77
Zhou *et al.* [3]	87.36	109.31	87.05	103.16	116.18	143.32	106.88	99.78
Sanzari *et al.* [4]	**48.82**	**56.31**	95.98	**84.78**	96.47	105.58	**66.30**	107.41
Ours - Single PPCA Model	68.55	78.27	77.22	89.05	91.63	110.05	74.92	83.71
Ours - Mixture PPCA Model	64.98	73.47	**76.82**	86.43	**86.28**	110.67	68.93	**74.79**
	Sitting	Sitting Down	Smoking	Waiting	Walk Dog	Walking	Walk Together	Average
Tekin *et al.* [2]	**73.76**	170.4	85.08	116.91	113.72	**62.08**	94.83	100.08
Zhou *et al.* [3]	124.52	199.23	107.42	118.09	114.23	79.39	97.70	113.01
Sanzari *et al.* [4]	116.89	**129.63**	97.84	**65.94**	130.46	92.58	102.21	93.15
Ours - Single PPCA Model	115.94	185.72	88.25	88.73	92.37	76.48	77.95	92.96
Ours - Mixture PPCA Model	110.19	173.91	**84.95**	85.78	**86.26**	71.36	**73.14**	**88.39**

3D error (mm)	Protocol #2
Yasin *et al.* [5]	108.3
Rogez *et al.* [6]	88.1
Ours	**70.7**

3D error (mm)	Protocol #3
Bogo *et al.* [7]	82.3
Ours	**79.6**

2D pixel error	
Zhou *et al.* [3]	10.85
Trained CPM [1] architecture	10.04
Ours using 3D refinement	**9.47**

Qualitative Results on MPII Dataset



References

[1] S. Wei, V. Ramakrishna, T. Kanade, and Y Sheikh. Convolutional pose machines. In CVPR, 2016
[2] B. Tekin, P. Marquez-Neila, M. Salzmann, and P. Fua. Fusing 2D Uncertainty and 3D Cues for Monocular Body Pose Estimation. In ArXiv, 2016
[3] X. Zhou, *et al.* Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In CVPR, 2016
[4] M. Sanzari, V. Ntouskos, and F. Pirri. Bayesian Image Based 3D Pose Estimation. In ECCV, 2016
[5] H. Yasin, U. Iqbal, B. Kruger, A. Weber and J. Gall. A dual-source approach for 3D pose estimation from a single image. In CVPR, 2016
[6] G. Rogez and C. Schmid. MoCap-guided data augmentation for 3D pose estimation in the wild. In NIPS, 2016
[7] F. Bogo *et al.* Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In ECCV, 2016
[8] C. Ionescu *et al.* Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. In PAMI, 2014

Acknowledgements

This work has been partly supported by the SecondHands project, funded from the European Union Horizon 2020 Research and Innovation programme under grant agreement No 643950. Chris Russell was funded by a UCL/BBC research fellowship and under the EPSRC grant EP/N510129/1.