Tsinghua University

i-VisionGroup

文献分享

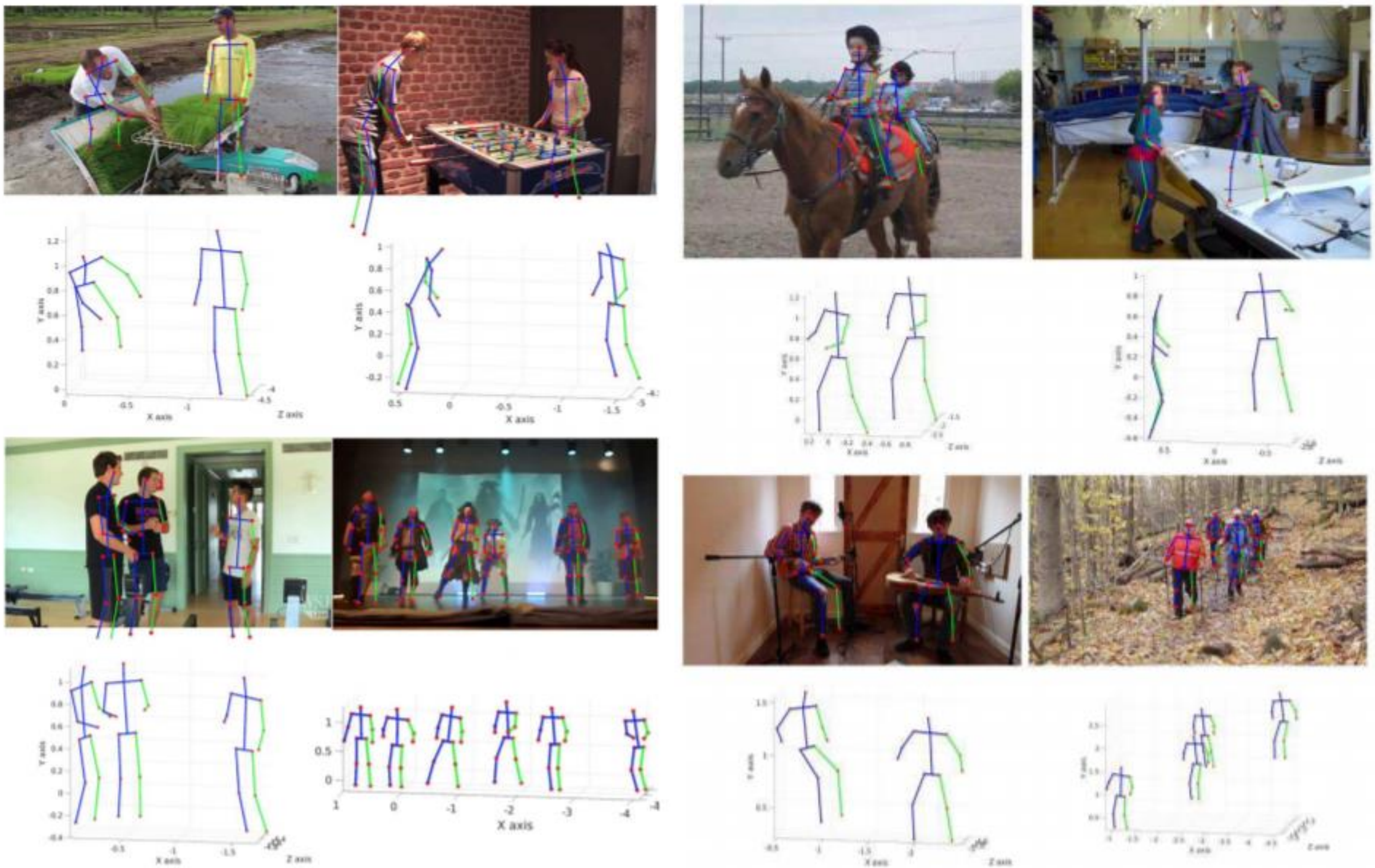
范博昊

单目多人3Dpose检测

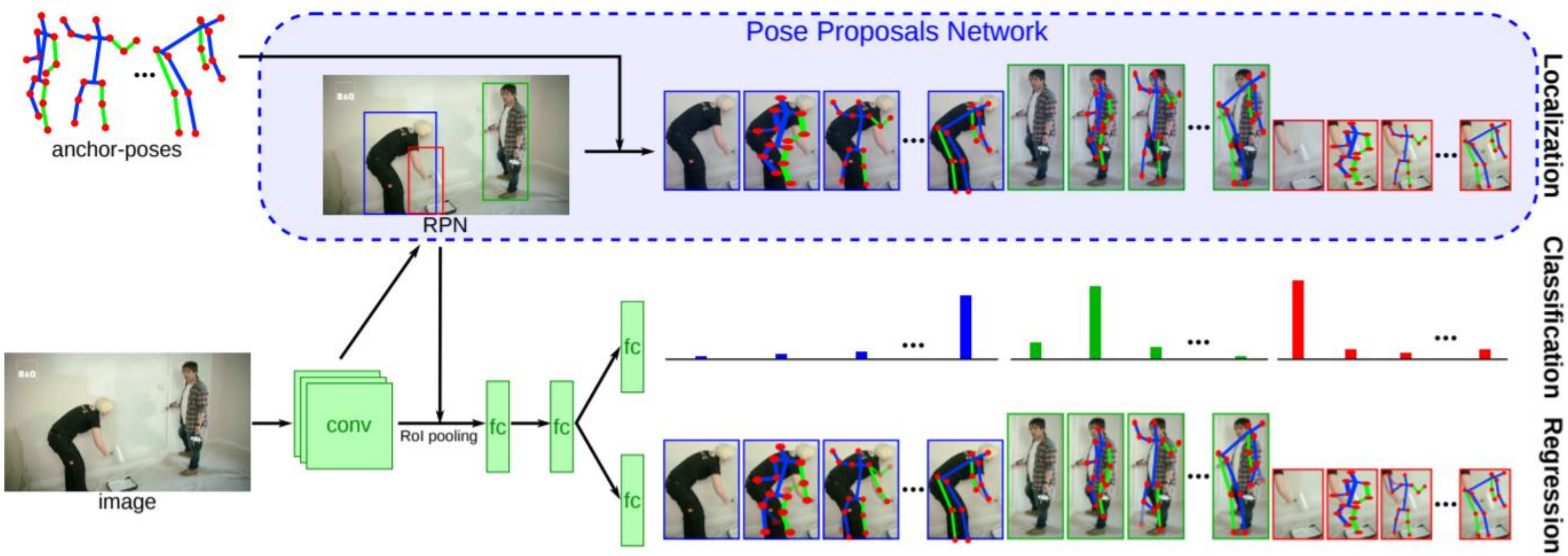
- Top-down
  - LCRNet(CVPR17)
  - LCRNet++(PAMI19)
  - HG-RCNN(3DV19)

- Bottom-up
  - ORPM(3DV18)
  - Xnect(arxiv)

LCR-net:CVPR2017



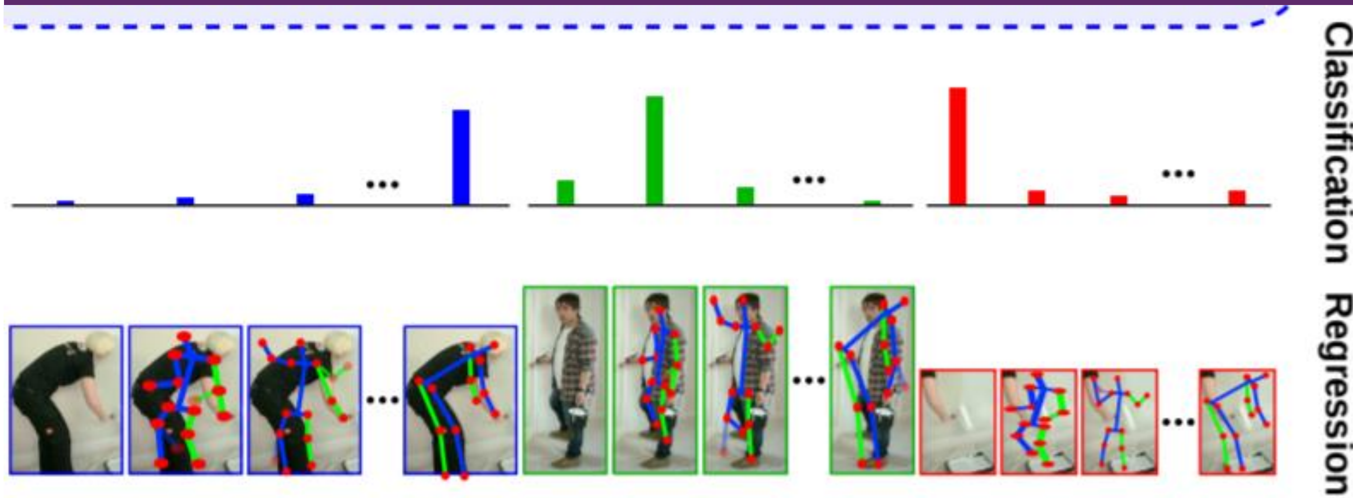
结构



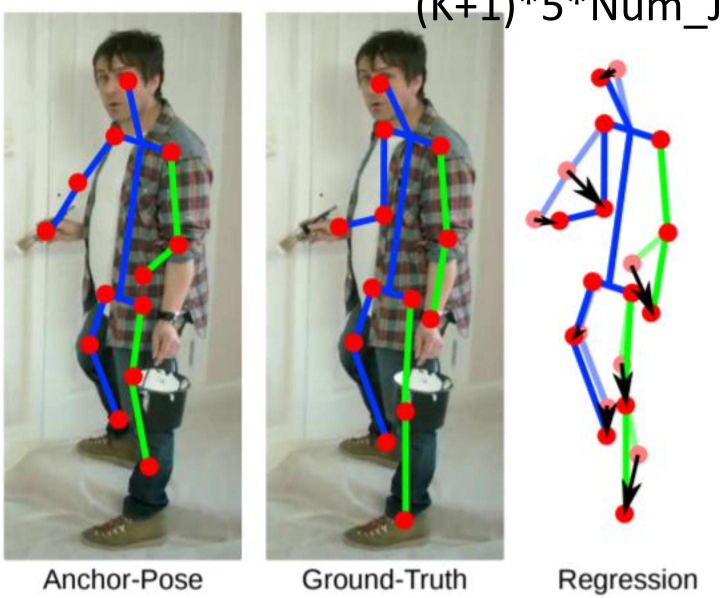
$$\mathcal{L} = \mathcal{L}_{Loc} + \mathcal{L}_{Classif} + \mathcal{L}_{Reg} \qquad \mathcal{L}_{Classif}(u, c_B) = -log\, u(c_B)$$

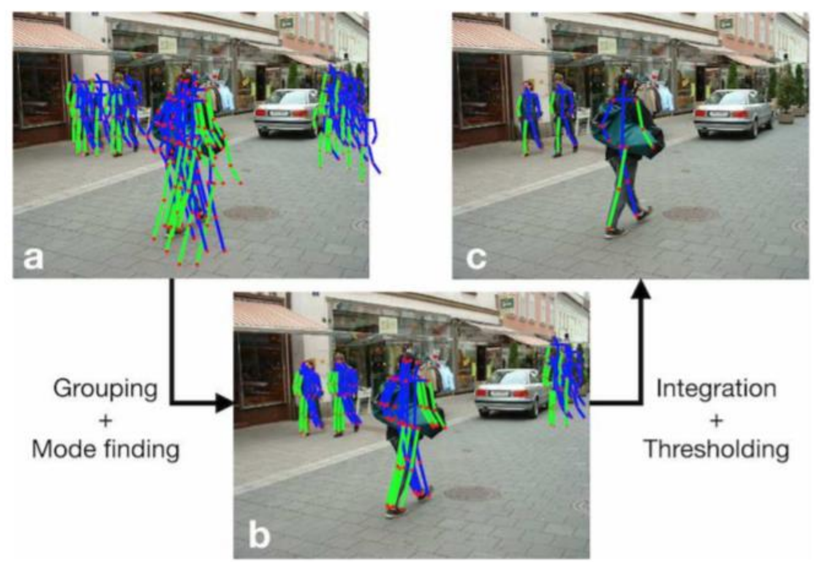
$$\mathcal{L}_{Loc} = \mathcal{L}_{RPN} \qquad \mathcal{L}_{Reg}(v, t_{c_B}) = [c_B \geqslant 1] \ \ \|t_{c_B} - v_{c_B}\|_S$$

$$\|x\|_S = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x| - 0.5 & \text{otherwise.} \end{cases}$$


Classification
Regression

(K+1)*5*Num_J


Anchor-Pose
Ground-Truth
Regression

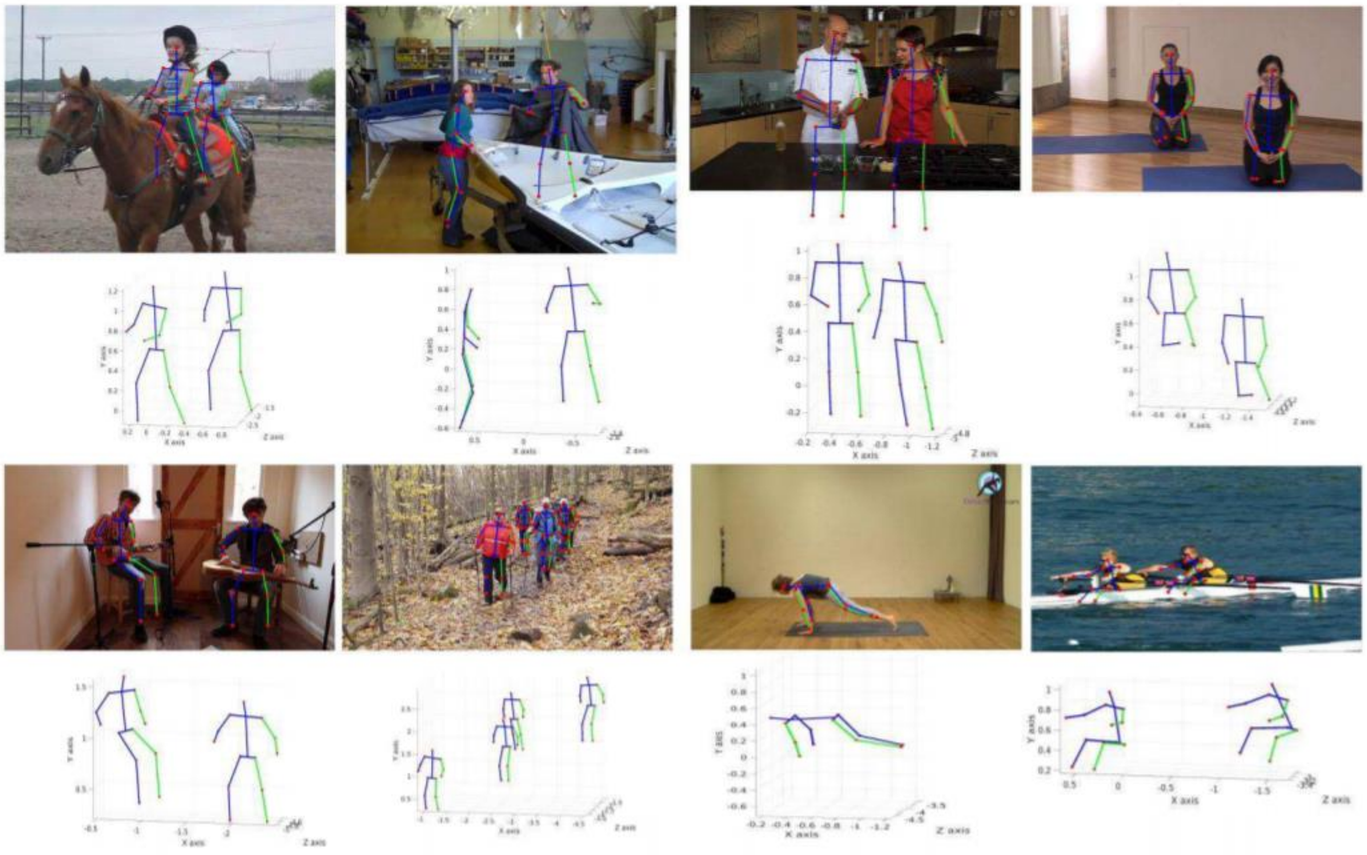

a
b
c
Grouping
+
Mode finding
Integration
+
Thresholding

results

Method	Im	Loc	Directions	Discussion	Eat	Greet	Phone	Pose	Purchase	Sit	SitDown
Tekin *et al.* [29]		✓	102.4	147.7	88.8	125.3	118.0	112.3	129.2	138.9	224.9
Zhou *et al.* [37]			87.4	109.3	87.0	103.2	116.2	106.9	99.8	124.5	199.2
Du *et al.* [7]		✓	85.1	112.7	104.9	122.1	139.1	105.9	166.2	117.5	226.9
Li *et al.* [14]	✓		-	148.8	104.0	127.2	-	-	-	-	-
Li *et al.* [15]	✓		-	134.1	97.4	122.3	-	-	-	-	-
Li *et al.* [16]	✓		-	133.5	97.6	120.4	-	-	-	-	-
Tekin *et al.* [28]	✓		-	129.1	91.4	121.7	-	-	-	-	-
Rogez & Schmid [23]	✓		94.5	110.4	109.3	143.9	125.9	95.5	89.8	134.2	179.2
Sanzari *et al.* [24]	✓	✓	**48.8**	**56.3**	96.0	84.8	96.5	**66.3**	107.4	116.9	129.6
LCR-Net + NMS	✓	✓	79.8	84.5	76.4	86.6	94.2	81.6	74.2	106.3	129.4
LCR-Net + PPI	✓	✓	76.2	80.2	**75.8**	**83.3**	**92.2**	79.0	**71.7**	**105.9**	**127.1**

Method	Im	Loc	Smoke	Photo	Wait	Walk	WalkDog	WalkTogether	Avg. (All)	Avg. (6)
Tekin *et al.* [29]		✓	118.4	182.7	138.7	**55.1**	126.3	**65.8**	125.0	121.0
Zhou *et al.* [37]			107.4	143.3	118.1	79.4	114.2	97.7	113.0	106.1
Du *et al.* [7]		✓	120.0	135.9	117.6	99.3	137.4	106.5	126.5	118.7
Li *et al.* [14]	✓		-	189.1	-	77.6	146.6	-	-	132.2
Li *et al.* [15]	✓		-	166.2	-	68.5	132.5	-	-	121.3
Li *et al.* [16]	✓		-	163.3	-	73.7	135.2	-	-	121.6
Tekin *et al.* [28]	✓		-	162.2	-	65.7	130.5	-	-	116.8
Rogez & Schmid [23]	✓		123.8	160.3	133.0	77.4	129.5	91.3	121.2	119.5
Sanzari *et al.* [24]	✓		97.8	**105.6**	65.9	92.6	130.5	102.2	93.1	-
LCR-Net + NMS	✓	✓	90.5	106.5	86.5	64.8	92.5	84.2	89.8	85.2
LCR-Net + PPI	✓	✓	**88.0**	105.7	**83.7**	64.9	**86.6**	84.0	**87.7**	**83.0**

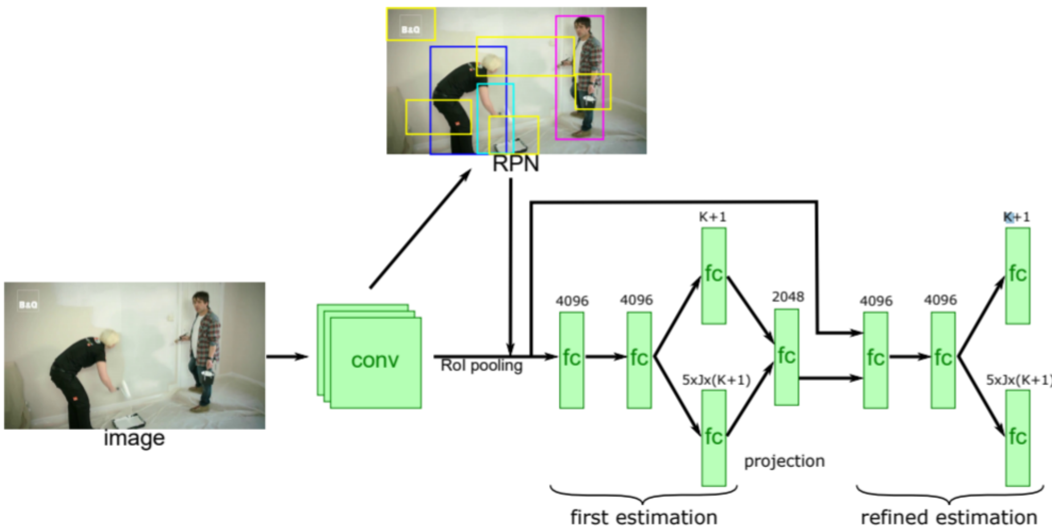
results



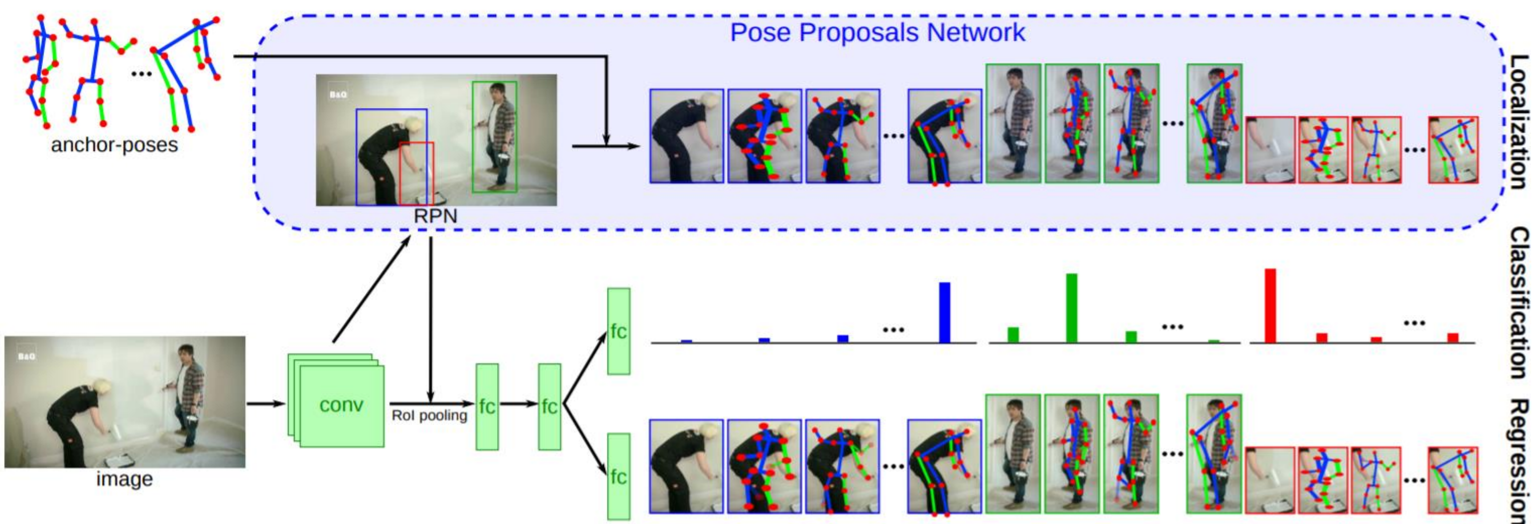
Tsinghua University

i-VisionGroup

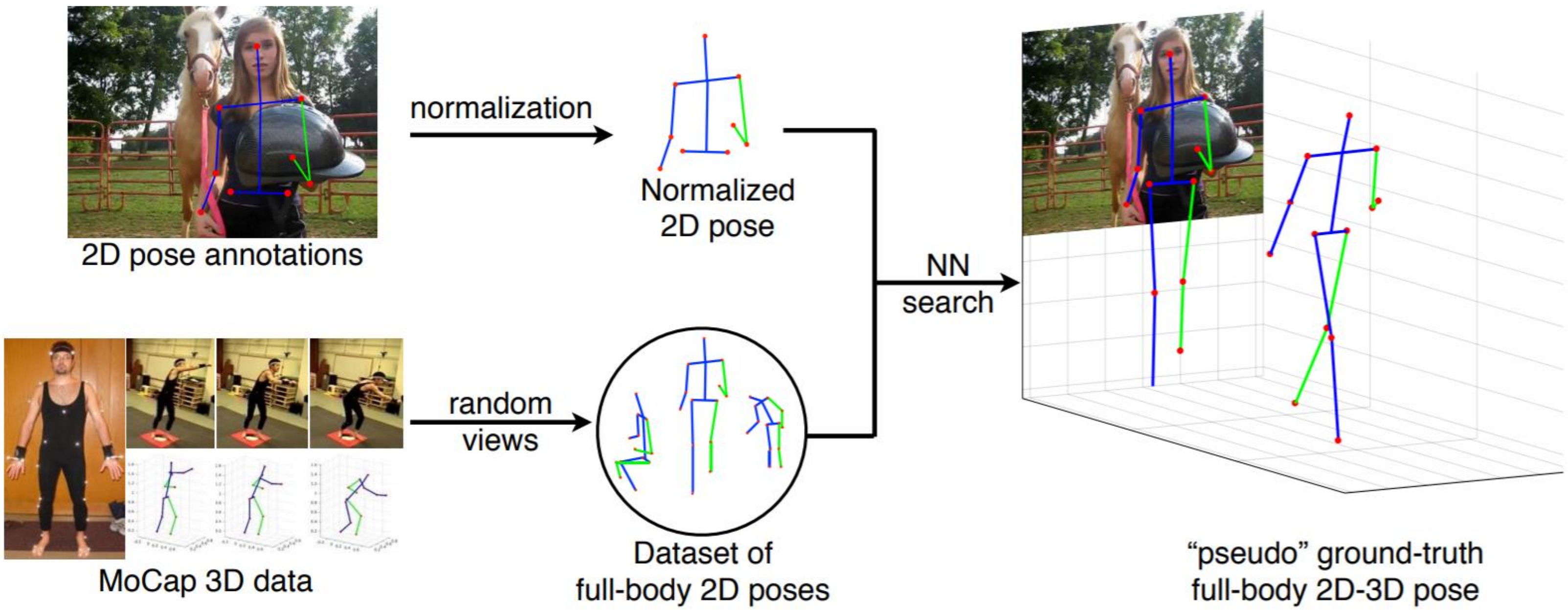
LCRNet++(PAMI2019)



Iterative Estimation



3D GT


2D pose annotations
normalization
Normalized
2D pose
NN
search
MoCap 3D data
random
views
Dataset of
full-body 2D poses
"pseudo" ground-truth
full-body 2D-3D pose

3DV2019

Multi-Person 3D Human Pose Estimation from Monocular Images

Rishabh Dabral
IIT Bombay
rdabral@cse.iitb.ac.in

Nitesh B Gundavarapu
IIT Bombay
ntesh93@gmail.com

Rahul Mitra
IIT Bombay
rmitter@cse.iitb.ac.in

Abhishek Sharma
Axogyan AI
abhisharayiya@gmail.com

Ganesh Ramakrishnan
IIT Bombay
ganesh@cse.iitb.ac.in

Arjun Jain
Axogyan AI
arjunjain@gmail.com

结构

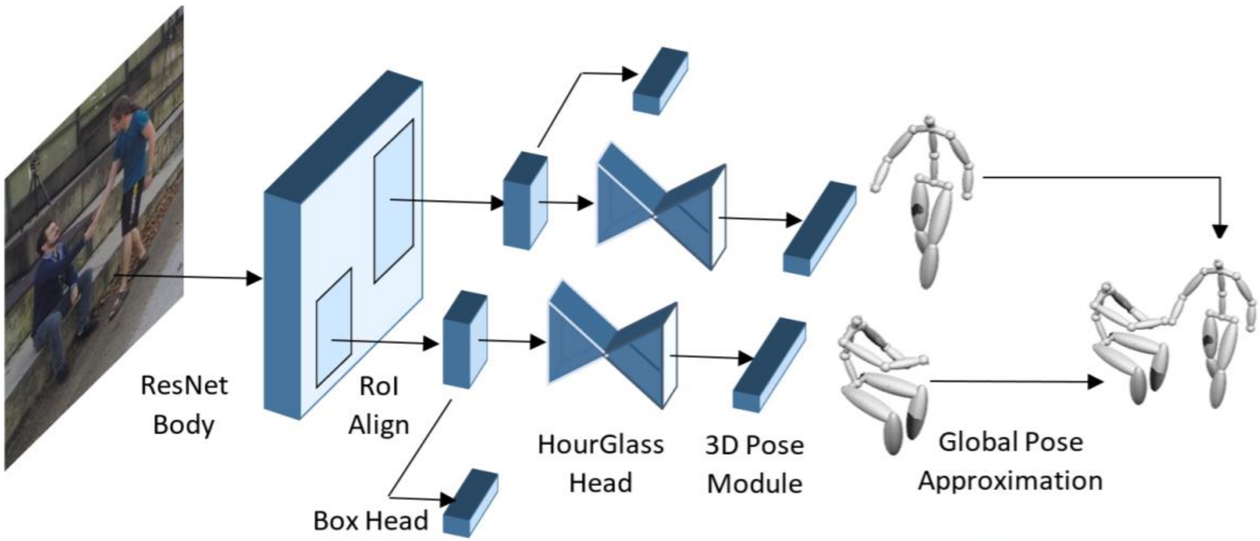


Figure 2. Schematic of our Multi-Person 3D Pose Estimation approach. We augment the Faster-RCNN [26] architecture with a shallow HourGlass Network [23]. The heatmaps generated by the hourglass are then input to a 3D Pose Module which regresses the root-relative 3D joint coordinates. The estimated 3D poses of all the Regions of Interests (RoI's) are then collected and their global root positions are approximated to ensure that relative spatial ordering is preserved.

3D Pose Module: Our 2D-to-3D pose module converts the heatmap activations to 3D pose using a residual architecture and is in line with the 2D-3D lifting pipelines proposed in [17, 32, 22]. We input the 2D poses in heatmap space after passing the heatmaps through a *softargmax* layer. This has two benefits: a) it makes learning possible from images of any given size and scale, and b) it facilitates end-to-end training of the network architecture. The network is trained using RMSProp optimizer and a learning rate of $2.5 \exp -4$ which is reduced by 10 times after 40 epochs.

$$Z = f * \frac{S_{3D}}{S_{2D}}$$

$$f^*, t^* = \arg\min_{f,t} \sum_{i=1}^{N} ||K_i - \Pi_{f,t_i} P_i||_2 \quad (3)$$

where $t = \{t_1, t_2, \ldots t_N\}$ with t_i being the translation vector of i^{th} subject's root joint and Π being the projection operator. This, finally, leads to the global pose, $P_i^G = P_i + t_i^*$.

results

10^{th} Percentile 30^{th} Percentile 60^{th} Percentile 90^{th} Percentile

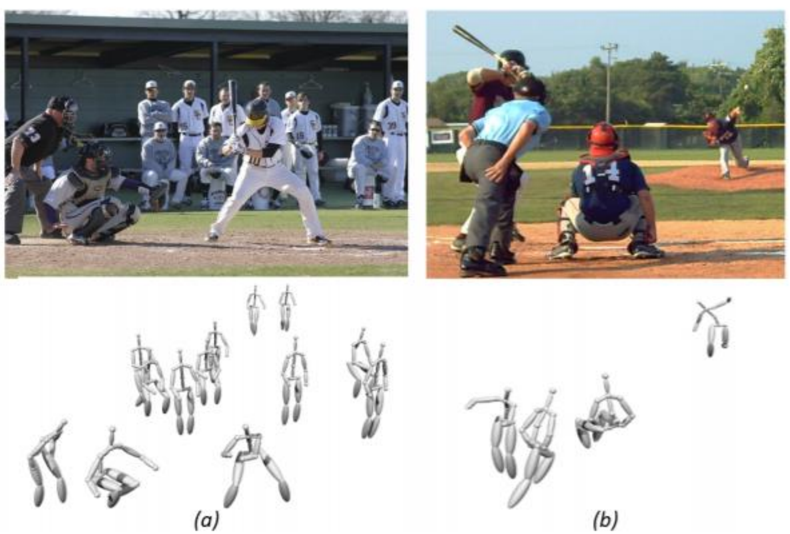


Table 5. Comparison of HG-RCNN and Mask-RCNN based models on MuPoTS 3D. The evaluation metric is 3DPCK.

	Mask-RCNN	HG-RCNN
all annotated joints	70.1	72.4
all occluded joints	61.0	64.1

results

Table 1. Comparison of our method with prior work on MuPoTS-3D on *Setting 1*. The **top half** shows results on *all annotated poses* in the test set. The **bottom half** shows results when only the detected poses are considered. The evaluation metric is 3D PCK and higher is better. *Note, that the average PCK provided in LCRNet++ [28] is not weighed by the number of persons in each test sequence unlike [27, 20] and ours.

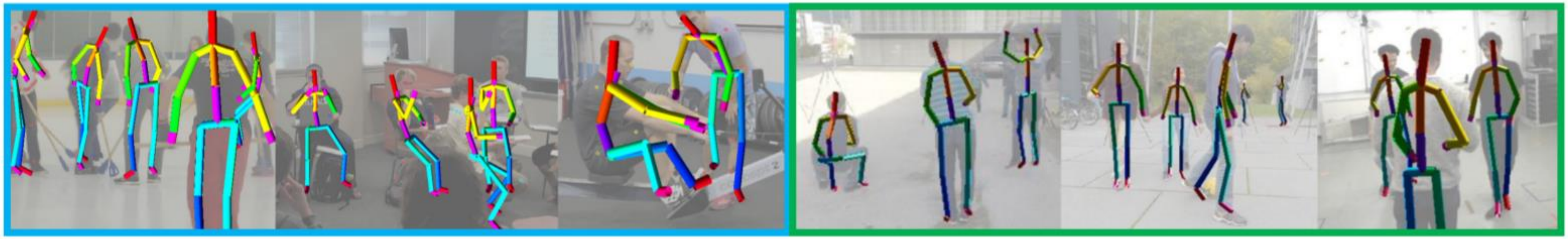
Method	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS18	TS19	TS20	Avg
[27]	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1	50.2	51.0	51.6	49.3	56.2	66.5	65.2	62.9	66.1	59.1	53.8
[20]	81.0	59.9	64.4	62.8	68.0	30.3	65.0	59.2	64.1	83.9	67.2	68.3	60.6	56.5	69.9	79.4	79.6	66.1	66.3	63.5	65.0
[28]*	87.3	61.9	67.9	74.6	**78.8**	48.9	58.3	59.7	**78.1**	**89.5**	69.2	**73.8**	**66.2**	56.0	**74.1**	82.1	78.1	72.6	73.1	61.0	70.6
[19]	**88.4**	65.1	68.2	72.5	76.2	46.2	**65.8**	**64.1**	75.1	82.4	74.1	72.4	64.4	**58.8**	73.7	80.4	**84.3**	67.2	74.3	67.8	70.4
Ours	85.1	**67.9**	**73.5**	**76.2**	74.9	**52.5**	65.7	63.6	56.3	77.8	**76.4**	70.1	65.3	51.7	69.5	**87.0**	82.1	**80.3**	**78.5**	**70.7**	**71.3**
[27]	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
[20]	81.0	64.3	64.6	63.7	73.8	30.3	65.1	60.7	64.1	83.9	71.5	69.6	69.0	69.6	71.1	82.9	79.6	72.2	76.2	85.9	69.8
[28]*	88.0	73.3	67.9	**74.6**	81.8	50.1	60.6	60.8	**78.2**	**89.5**	70.8	**74.4**	**72.8**	64.5	**74.2**	84.9	85.2	78.4	75.8	74.4	74.0
[19]	**88.4**	70.4	**68.3**	73.6	**82.4**	46.4	66.1	**83.4**	75.1	82.4	76.5	73.0	72.4	**73.8**	74.0	83.6	84.3	73.9	**85.7**	90.6	**75.8**
Ours	85.8	**73.6**	61.1	55.7	77.9	**53.3**	**75.1**	65.5	54.2	81.3	**82.2**	71.0	70.1	67.7	69.9	**90.5**	**85.7**	**86.3**	85.0	**91.4**	74.2

3DV2018

Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB

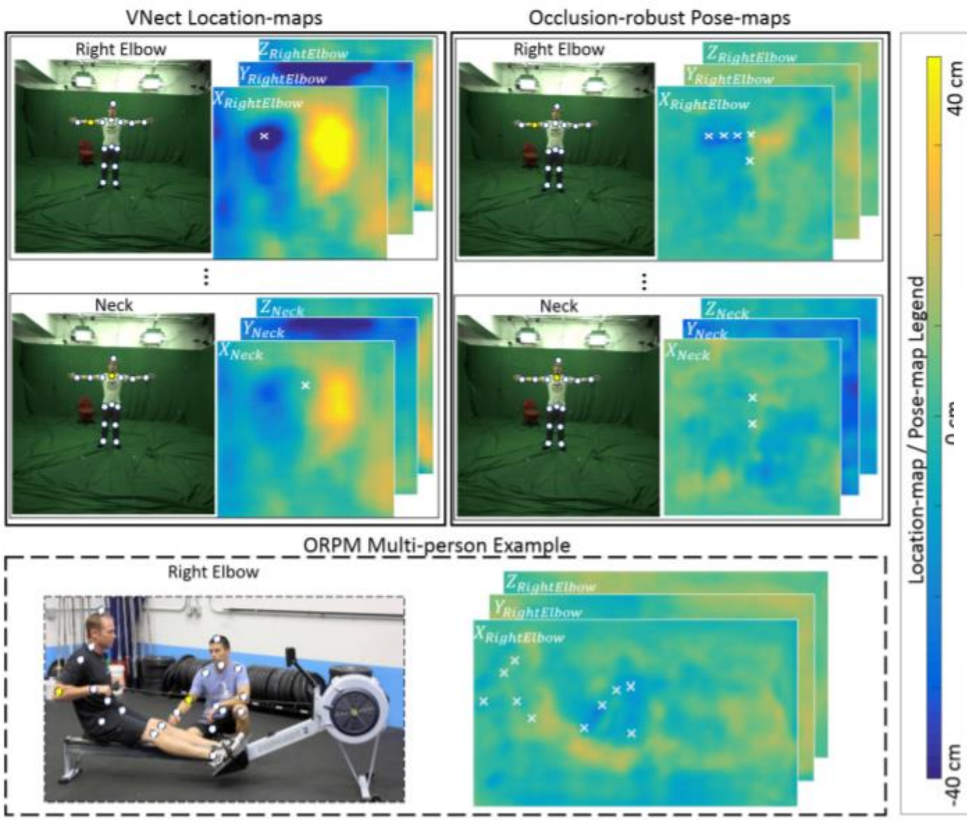
Dushyant Mehta[1,2], Oleksandr Sotnychenko[1,2], Franziska Mueller[1,2],
Weipeng Xu[1,2], Srinath Sridhar[3], Gerard Pons-Moll[1,2], Christian Theobalt[1,2]

[1] MPI For Informatics [2] Saarland Informatics Campus [3] Stanford University



Location Map VS ORPM

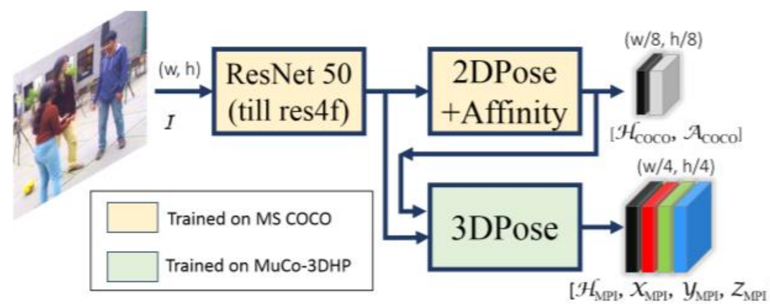
- 3n LM of size W/k*H/k



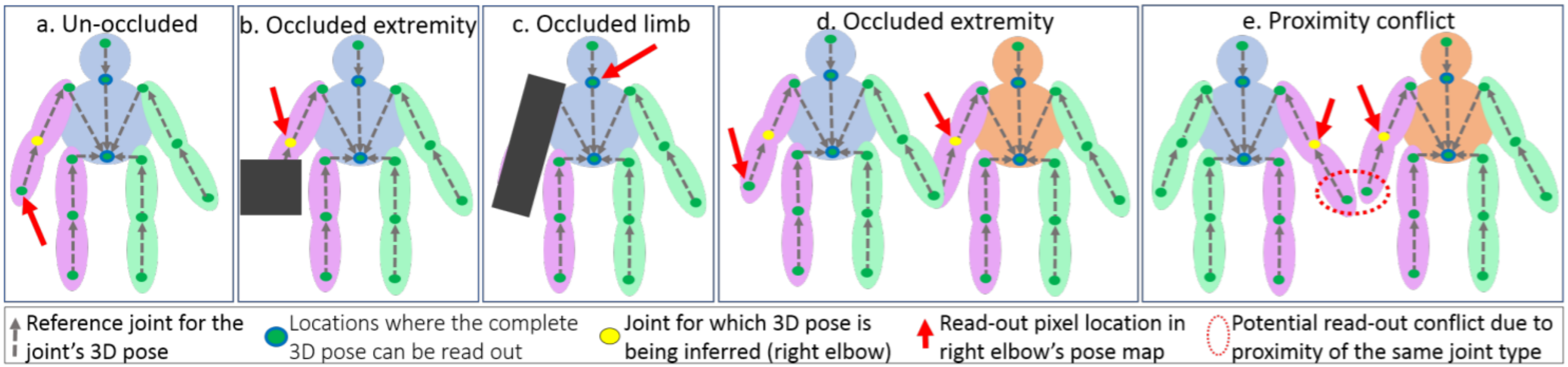
ORPM:

$$M = \{M_j\}_{j=1}^{n}$$

$$M_j \in R^{M*N*3}$$



ORPM+



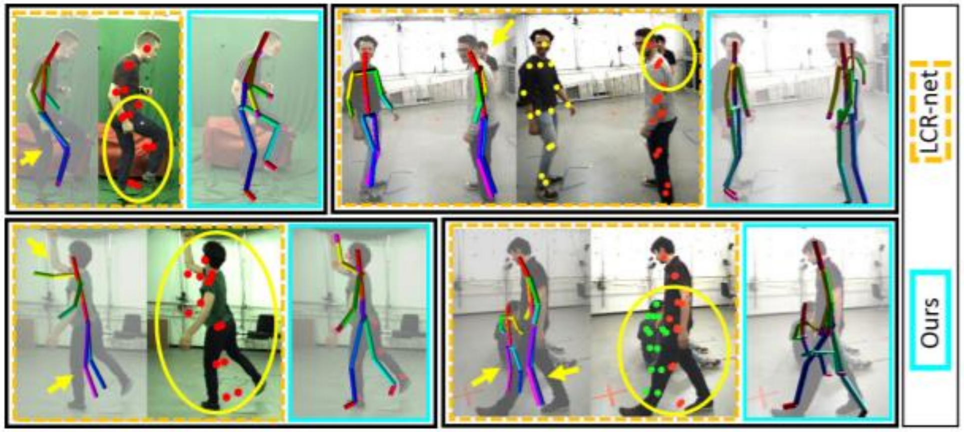
Inference: Openpose Heatmap+part affinity +confidence map+ORPM

2D Joint Validation: We declare a 2D joint location $\mathbf{P^{2D}}_i^j = (u, v)_j^i$ of person i as *valid read-out location* iff (1) it is un-occluded, *i.e.*, has confidence value higher than a threshold t_C, and (2) it is sufficiently far ($\geq t_D$) away from all read-out locations of joint j of other individuals:

$$valid(\mathbf{P^{2D}}_i^j) \Leftrightarrow \mathbf{C^{2D}}_i^j > t_C \ \wedge \ ||a - \mathbf{P^{2D}}_i^j||_2 \geq t_D \quad \forall \bar{i} = [1\!:\!m],\ \bar{i} \neq i.\ \forall a \in \rho_{\bar{i}}(j). \tag{2}$$

Loss: The 2D heatmaps $\mathcal{H}_{COCO}$ and $\mathcal{H}_{MPI}$ are trained with per-pixel $L2$ loss comparing the predictions to the reference which has unit peak Gaussians with a limited support at the ground truth 2D joint locations, as is common. The part affinity fields $\mathcal{A}_{COCO}$ are similarly trained with a per-pixel $L2$ loss, using the framework made available by Cao *et al.* [8]. While training ORPMs with our *MuCo-3DHP*, per joint type j, for all subjects i in the scene, a per-pixel $L2$ loss is enforced in the neighborhood of all possible read-out locations $\rho_i(j)$. The loss is weighted by a limited support Gaussian centered at the read-out location.

results



Method	Sit	Crouch	Total	
	PCK	PCK	PCK	AUC
VNect [34]	74.7	72.9	**76.6**	**40.4**
LCR-net [49]	58.5	69.4	59.7	27.6
Zhou et al.[68]	60.7	71.4	69.2	32.5
Mehta et al.[33]	74.8	73.7	75.7	39.3
Our Single-Person (Torso)	69.1	68.7	65.6	32.6
Our Single-Person (Full)	**77.8**	**77.5**	75.2	37.8
Our Multi-Person (Torso)	64.6	65.8	63.6	31.1
Our Multi-Person (Full)	75.9	73.9	73.4	36.2

results