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

1. Read-out Process

An algorithmic description of the read-out process is provided in Alg. 1.

Algorithm 1 3D Pose Inference

Given: $\mathcal{P}^{2D}$, $\mathcal{C}^{2D}$, $\mathcal{M}$
for all $i \in (1..m)$ **do**
  if $\mathbf{C}_i^{2D}[k] > thresh$, $k \in \{pelvis, neck\}$ **then**
    Person i is detected
    for all joints $j \in (1..n)$ **do**
      $rloc = \mathbf{P}_i^{2D}[k]$
      $\mathbf{P}_i[:, j]$ = ReadLocMap(j, rloc)
    for all limbs $l \in \{arm_l, arm_r, leg_l, leg_r, head\}$ **do**
      for j = getExtremity(l); $j \notin \{pelvis, neck\}$; $j = parent(j)$ **do**
        if isValidReadoutLoc(i, j) **then**
          refineLimb(l, $\mathbf{P}_i^{2D}[j]$)
          break
  else
    No person detected
function getExtremity(limb l)
  if $l = leg_s$ **then return** $ankle_s$
  else
    if $l = arm_s$ **then return** $wrist_s$
    else return $head$
function ReadLocMap(joint j, 2DLocation rloc)
  $rloc = rloc/locMap\_scale\_factor$
  return $\mathbf{M}_j[rloc]$
function refineLimb(limb l, 2DLocation rloc)
  for all joints $b \in$ limb l **do**
    $\mathbf{P}_i[:, b]$=ReadLocMap(b, rloc)
function isValidReadoutLoc(person i, joint j)
  if $(\mathbf{C}_i^{2D}[j] > 0)$ **then**
    return isIsolated(i,j)
  else
    return 0
function isIsolated(person i, joint j)
  $isol = 1$
  for all persons $\bar{i} \in (1..m), \bar{i} \neq i$ **do**
    for all 2DLocations $a \in \rho_{\bar{i}}(j)$ **do**
      if $||a - \mathbf{P}_i^{2D}[j]||_2 < isoThresh$ **then**
        $isol = 0$
        break
  return $isol$

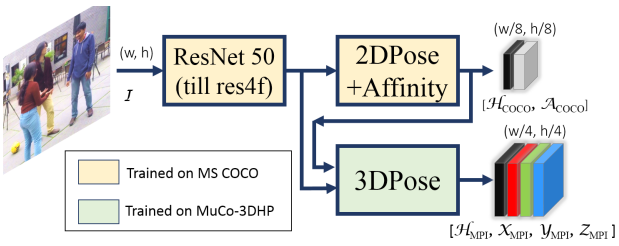

Figure 1. The network architecture with *2DPose+Affinity* branch predicting the 2D *heatmaps* $\mathcal{H}_{COCO}$ and *part affinity maps* $\mathcal{A}_{COCO}$ with a spatial resolution of $(W/8, H/8)$, and *3DPose* branch predicting 2D *heatmaps* $\mathcal{H}_{MPI}$ and ORPMs $\mathcal{M}_{MPI}$ with a spatial resolution of $(W/4, H/4)$, for an input image with resolution (W, H).

2. Network Details

2.1. Architecture

A visualization if our network architecture using the web-based visualization tool *Netscope* can be found at: `http://ethereon.github.io/netscope/#/gist/069a592125c78fbdd6eb11fd45306fa0`.

2.2. Data

We use 12 out of the 14 available camera viewpoints (using only 1 of the 3 available top down views) in MPI-INF-3DHP [9] training set, and create 400k composite frames of MuCo-3DHP, of which half are without appearance augmentation. For training, we crop around the subject closest to the camera, and apply rotation, scale, and bounding-box jitter augmentation. Since the data was originally captured in a relatively restricted space, the likelihood of there being multiple people visible in the crop around the main person is high. The combination of scale augmentation, bounding-box jitter, and cropping around the subject closest to the camera results in many examples with truncation from the frame boundary, in addition to the inter-person occlusions occurring naturally due to the compositing.

2.3. Training

We train our network using the Caffe [5] framework. The core network's weights were initialized with those trained for 2D body pose estimation on MPI [1] and LSP [6, 7] datasets as done in [9]. The core network and the *2DPose + Affinity* branch are trained for multi-person 2D pose estimation using the framework provided by Cao et al. [2]. We use the AdaDelta solver, with a momentum of 0.9 and weight decay multiplier of 0.005, and a batch size of 8. We train for 640k iterations with a cyclical learning rate ranging from 0.1 to 0.000005. The *3DPose* branch is trained with the

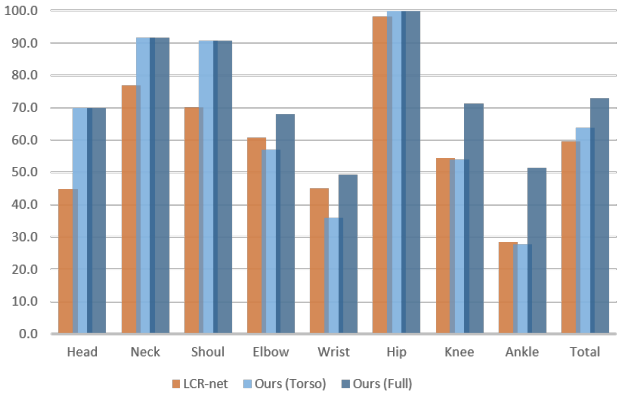

Figure 2. Joint-wise accuracy comparison of our method and LCR-net [14] on the single person MPI-INF-3DHP test set. 3D Percentage of Correct Keypoints (@150mm) as the vertical axis. LCR-net predictions were mapped to the ground truth bone lengths for fairness of comparison.

core network and *2DPose + Affinity* branch weights frozen. We use a batch size of 6 and train for 360k iterations with a cyclical learning rate ranging from 0.1 to 0.000001. We empirically found that training the part affinity fields and occlusion-robust pose-maps at lower resolution (see Fig. 1) leads to better results.

3. Joint-wise Analysis

Figure 2 shows joint-wise accuracy comparison of our approach with LCR-net [14] on the single person MPI-INF-3DHP test set. For limb joints (elbow, wrist, knee, ankle) LCR-net performs comparably or better than our torso-only readout, but our full readout performs significantly better. See Figure 3.

Figure 5 shows joint-wise accuracy comparison of our approach with LCR-net on our proposed multi-person 3D pose test set. We see that our approach obtains a better accuracy for all joint types for most sequences, only performing worse than LCR-net for a select few joint types on certain sequences (TestSeq18,19,20).

4. Evaluation on Single-person Test Sets

Here we provide a detailed comparison against other methods for single-person 3D pose estimation. Evaluation on Human3.6m is in Table 3, and on MPI-INF-3DHP test set in Table 4. We additionally provide comparisons with the VNect location-maps trained on our training setup, which includes the 2D pretraining, and the 3D pose samples.

Table 4 provides a sequencewise breakdown for the synthetic occlusion experiment on MPI-INF-3DHP test set wherein through randomly placed occlusions ≈14% of the joints are occluded. This doesn't account for self-occlusions.

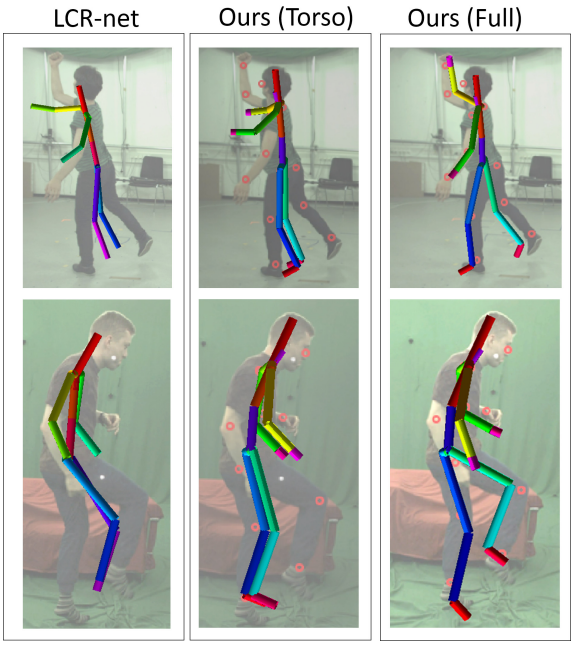

Figure 3. Qualitative comparison of LCR-net [14] and our method. LCR-net predictions are limited in terms of the extent of articulation of limbs, tending towards neutral poses. For our method, the base pose read out at the torso is similarly limited in terms of degree of articulation of limbs, and our full read-out addresses the issue.

Figure 4. Examples from our MuPoTS-3D evaluation set. Ground truth 3D pose reference and joint occlusion annotations are available for up to 3 subjects in the scene (shown here for the frame on the top right). The set covers a variety of scene settings, activities and clothing.

Table 1. Comparison of results on Human3.6m [4], for single un-occluded person. Human3.6m, subjects 1,5,6,7,8 used for training. Subjects 9 and 11, all cameras used for testing. Mean Per Joint Postion Error reported in mm

	Direct	**Disc.**	**Eat**	**Greet**	**Phone**	**Pose**	**Purch.**	**Sit.**
Pavlakos et al [13]	60.9	67.1	61.8	62.8	67.5	58.8	64.4	79.8
Mehta et al [9]	52.5	63.8	55.4	62.3	71.8	52.6	72.2	86.2
Tome et al [16]	65.0	73.5	76.8	86.4	86.3	69.0	74.8	110.2
Chen et al [3]	89.9	97.6	90.0	107.9	107.3	93.6	136.1	133.1
Moreno et al [11]	67.5	79.0	76.5	83.1	97.4	74.6	72.0	102.4
Zhou et al [17]	54.8	60.7	58.2	71.4	62.0	53.8	55.6	75.2
Martinez et al [8]	51.8	56.2	58.1	59.0	69.5	55.2	58.1	74.0
Tekin et al [15]	53.9	62.2	61.5	66.2	80.1	64.6	83.2	70.9
Nie et al [12]	62.8	69.2	79.6	78.8	80.8	72.5	73.9	96.1
VNect [10]	62.6	78.1	63.4	72.5	88.3	63.1	74.8	106.6
LCR-net [14]	76.2	80.2	75.8	83.3	92.2	79.9	71.7	105.9
VNect (with our setup)	65.52	78.8	64.8	75.0	85.2	66.4	88.1	110.2
Our Single-Person	58.2	67.3	61.2	65.7	75.82	62.2	64.6	82.0

	Sit Down	**Smk.**	**Photo**	**Wait**	**Walk**	**Walk Dog**	**Walk Pair**	**Avg.**
Pavlakos et al [13]	92.9	67.0	72.3	70.0	54.0	71.0	57.6	67.1
Mehta et al [9]	120.6	66.0	79.8	64.0	48.9	76.8	53.7	68.6
Tome et al [16]	173.9	84.9	110.7	85.8	71.4	86.3	73.1	88.4
Chen et al [3]	240.1	106.6	139.2	106.2	87.0	114.0	90.5	114.2
Moreno et al [11]	116.7	87.7	100.4	94.6	75.2	87.8	74.9	85.6
Zhou et al [17]	111.6	64.1	65.5	66.0	63.2	51.4	55.3	64.9
Martinez et al [8]	94.6	62.3	78.4	59.1	49.5	65.1	52.4	62.9
Tekin et al [15]	107.9	70.4	79.4	68.0	52.8	77.8	63.1	70.8
Nie et al [12]	106.9	88.0	86.9	70.7	71.9	76.5	73.2	79.5
VNect [10]	138.7	78.8	93.8	73.9	55.8	82.0	59.6	80.5
LCR-net [14]	127.1	88.0	105.7	83.7	64.9	86.6	84.0	87.7
VNect (with our setup)	155.9	82.0	95.2	76.8	59.7	94.1	64.3	84.3
Our Single-Person	93.0	68.8	84.5	65.1	57.6	72.0	63.6	69.9

Ours

	Head	Neck	Shoul	Elbow	Wrist	Hip	Knee	Ankle	Total
TestSeq1	96.8	100.0	96.6	78.0	50.9	99.8	81.0	62.3	81.0
TestSeq2	63.9	85.9	68.3	54.9	47.6	75.6	55.3	42.8	59.9
TestSeq3	79.9	91.9	90.5	56.5	46.8	98.9	42.0	30.2	64.4
TestSeq4	73.1	82.7	76.5	56.6	49.3	97.0	50.7	31.4	62.8
TestSeq5	56.5	82.0	79.7	66.2	65.7	85.5	67.3	42.0	68.0
TestSeq6	6.5	32.1	35.5	12.7	10.9	99.4	23.3	10.8	30.3
TestSeq7	66.6	97.8	81.5	47.0	21.0	98.7	63.0	61.9	65.0
TestSeq8	55.2	71.6	65.8	57.2	45.2	96.1	44.4	42.8	59.2
TestSeq9	67.2	84.1	81.5	30.1	25.7	100.0	62.7	73.1	64.1
TestSeq10	98.2	100.0	100.0	63.2	52.1	100.0	95.4	77.8	83.9
TestSeq11	66.0	92.7	84.1	56.0	44.1	89.2	73.8	43.9	67.2
TestSeq12	46.1	73.1	76.2	73.4	66.8	97.1	64.4	40.8	68.3
TestSeq13	58.5	77.9	74.5	48.6	38.5	84.0	68.9	41.3	60.6
TestSeq14	47.5	73.3	69.7	43.8	38.4	79.8	62.1	41.0	56.5
TestSeq15	62.3	91.2	84.7	58.3	42.6	97.1	77.4	52.3	69.9
TestSeq16	72.9	87.8	86.1	82.3	80.1	92.9	81.9	51.9	79.4
TestSeq17	74.4	73.8	78.0	78.1	61.3	96.5	91.0	78.6	79.6
TestSeq18	54.8	73.8	77.1	73.1	44.2	87.6	74.6	41.5	66.1
TestSeq19	44.9	78.4	79.4	55.2	54.1	84.4	77.7	37.9	64.3
TestSeq20	50.8	73.5	71.7	62.5	58.6	73.0	69.0	47.5	63.5

LCR-net

	Head	Neck	Shoul	Elbow	Wrist	Hip	Knee	Ankle	Total
TestSeq1	73.1	81.8	74.4	61.3	41.9	97.0	74.9	47.1	67.7
TestSeq2	54.6	69.3	53.9	43.5	31.7	69.3	48.6	39.3	49.8
TestSeq3	71.2	81.0	56.4	35.8	29.6	95.1	49.1	31.7	53.4
TestSeq4	57.6	80.7	63.2	55.1	48.3	94.7	52.0	31.3	59.1
TestSeq5	68.0	84.9	72.1	70.5	60.5	84.7	65.1	43.2	67.5
TestSeq6	6.3	26.8	18.6	17.1	12.8	84.8	7.6	2.3	22.8
TestSeq7	34.4	66.8	38.8	31.5	22.9	87.6	49.1	25.6	43.7
TestSeq8	56.3	70.0	52.1	44.6	35.4	75.7	46.8	31.7	49.9
TestSeq9	34.9	41.9	34.1	20.0	15.9	45.1	32.8	31.4	31.1
TestSeq10	80.5	97.6	98.0	53.1	31.9	100.0	87.0	87.9	78.1
TestSeq11	23.3	43.0	39.4	61.5	44.1	88.2	57.4	27.4	50.2
TestSeq12	24.2	41.5	39.2	61.1	65.1	97.1	41.2	20.6	51.0
TestSeq13	42.0	62.2	51.5	48.2	37.4	78.5	57.1	36.5	51.6
TestSeq14	36.6	63.2	50.7	39.9	29.2	80.7	57.5	37.4	49.3
TestSeq15	39.4	72.8	59.1	44.6	34.4	91.4	73.6	34.6	56.2
TestSeq16	48.1	67.8	65.0	78.6	68.3	93.1	67.0	35.4	66.5
TestSeq17	44.1	77.7	75.7	68.8	58.9	85.4	59.7	46.8	65.2
TestSeq18	53.7	89.9	83.5	63.6	42.5	89.9	60.8	28.4	62.9
TestSeq19	69.6	80.0	67.2	62.4	53.0	81.1	73.9	50.2	66.1
TestSeq20	70.5	48.8	49.5	67.3	61.9	72.6	64.4	38.0	59.1

Difference

	Head	Neck	Shoul	Elbow	Wrist	Hip	Knee	Ankle	Total
TestSeq1	23.6	18.2	22.3	16.7	9.0	2.7	6.1	15.2	13.3
TestSeq2	9.4	16.5	14.4	11.4	15.9	6.3	6.7	3.5	10.2
TestSeq3	8.7	10.8	34.2	20.6	17.2	3.7	-7.0	-1.5	11.0
TestSeq4	15.4	2.0	13.3	1.5	0.9	2.3	-1.3	0.1	3.6
TestSeq5	-11.5	-2.9	7.6	-4.3	5.2	0.9	2.2	-1.1	0.5
TestSeq6	0.2	5.3	16.9	-4.4	-1.8	14.6	15.7	8.5	7.5
TestSeq7	32.2	30.9	42.7	15.5	-1.8	11.1	13.8	36.3	21.3
TestSeq8	-1.1	1.6	13.7	12.6	9.8	20.4	-2.4	11.1	9.3
TestSeq9	32.2	42.2	47.4	10.2	9.8	54.9	29.8	41.6	33.0
TestSeq10	17.7	2.4	2.0	10.2	20.2	0.0	8.5	-10.2	5.8
TestSeq11	42.7	49.7	44.7	-5.5	0.0	1.0	16.4	16.5	17.0
TestSeq12	21.9	31.5	37.0	12.3	1.7	0.0	23.2	20.3	17.3
TestSeq13	16.4	15.7	23.0	0.3	1.1	5.5	11.8	4.8	8.9
TestSeq14	10.9	10.2	19.0	3.8	9.3	-0.9	4.6	3.6	7.1
TestSeq15	22.9	18.3	25.6	13.8	8.2	5.7	3.7	17.8	13.6
TestSeq16	24.8	20.0	21.1	3.7	11.8	-0.2	14.9	16.5	12.9
TestSeq17	30.3	-3.9	2.2	9.3	2.3	11.1	31.3	31.7	14.5
TestSeq18	1.1	-16.1	-6.3	9.5	1.7	-2.4	13.8	13.1	3.1
TestSeq19	-24.7	-1.5	12.2	-7.2	1.0	3.2	3.8	-12.4	-1.8
TestSeq20	-19.6	24.7	22.1	-4.9	-3.3	0.4	4.6	9.5	4.4

Figure 5. Comparison of our method and LCR-net [14] on our proposed multi-person test set, here visualized as joint-wise breakdown of PCK for all 20 sequences, as well as the difference in accuracy between our method and LCR-net. LCR-net predictions were mapped to the ground truth bone lengths for fairness of comparison.

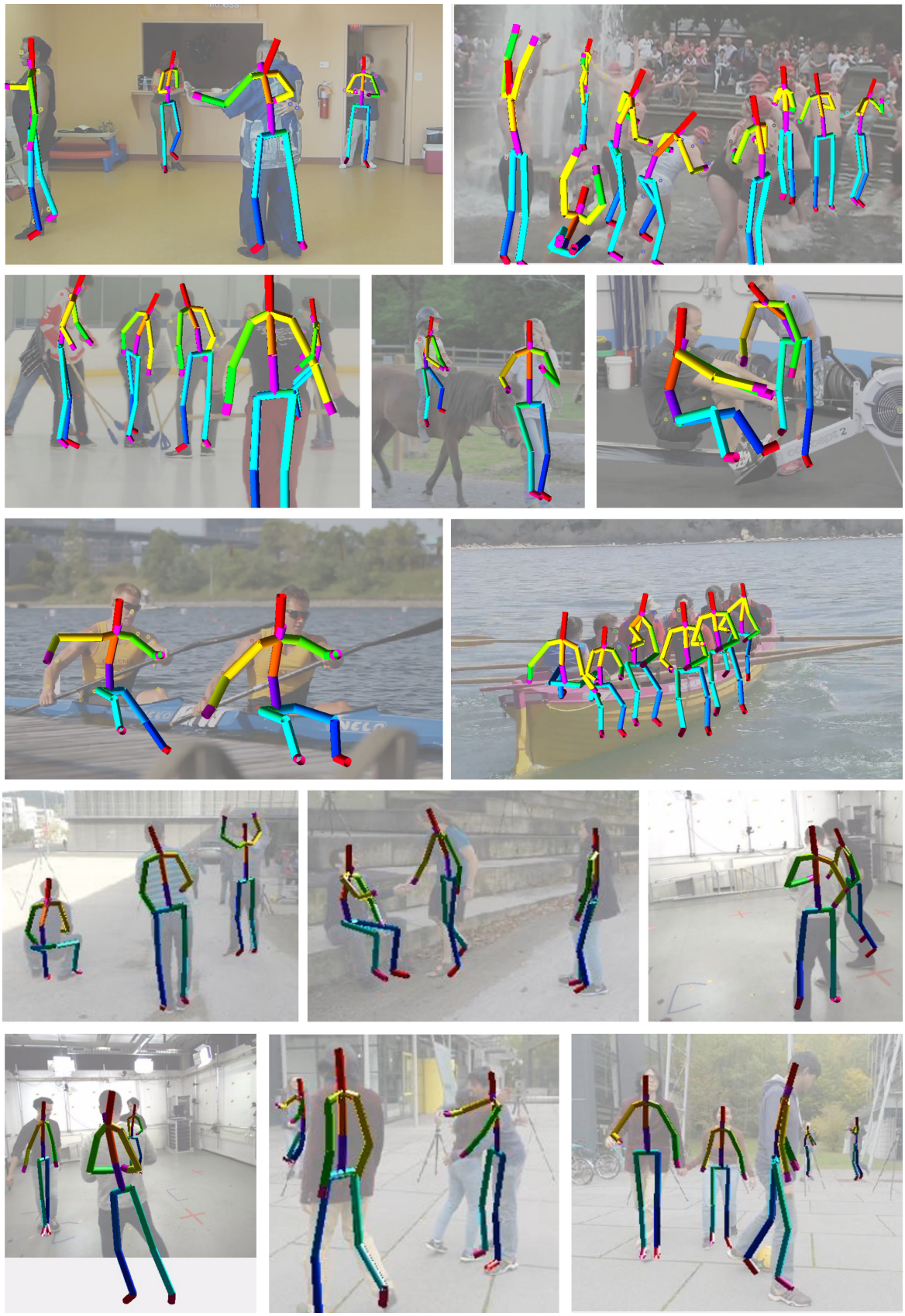

Figure 6. More qualitative results of our approach on MPI 2D pose dataset [1] and our proposed MuPoTS-3D test set.

Table 2. Comparison of our method against the state of the art on single person MPI-INF-3DHP test set. All evaluations use ground-truth bounding box crops around the subject. We report the *Percentage of Correct Keypoints measure in 3D* (@150mm), and the Area Under the Curve for the same, as proposed by MPI-INF-3DHP. We additionally report the Mean Per Joint Position Error in mm. Higher PCK and AUC is better, and lower MPJPE is better.

Network	Stand/ Walk	Exercise	Sit On Chair	Crouch/ Reach	On the Floor	Sports	Misc.	Total		
	PCK	PCK	PCK	PCK	PCK	PCK	PCK	PCK	AUC	MPJPE(mm)
VNect [10]	**87.7**	**77.4**	74.7	72.9	51.3	**83.3**	**80.1**	76.6	40.4	124.7
LCR-net [14]	70.5	56.3	58.5	69.4	39.6	57.7	57.6	59.7	27.6	158.4
Zhou et al.[17]	85.4	71.0	60.7	71.4	37.8	70.9	74.4	69.2	32.5	137.1
Mehta et al.[9]	86.6	75.3	74.8	73.7	52.2	82.1	77.5	75.7	39.3	117.6
Ours Single-Person (Torso)	75.0	64.8	69.1	68.7	48.6	70.0	60.6	65.6	32.6	142.8
Ours Single-Person (Full)	83.8	75.0	<u>77.8</u>	**77.5**	<u>55.1</u>	80.4	72.5	75.2	37.8	122.2
Ours Multi-Person (Torso)	73.7	63.7	64.6	65.8	44.7	69.5	60.2	63.6	31.1	146.8
Ours Multi-Person (Full)	82.0	74.5	75.9	<u>73.9</u>	51.6	79.0	71.8	73.4	36.2	126.3
VNect (our train. setup)	85.7	75.4	**78.6**	72.3	**60.2**	81.8	73.4	75.8	38.9	120.1

Table 3. Testing occlusion robustness of our method through synthetic occlusions on MPI-INF-3DHP single person test set. The synthetic occlusions cover about 14% of the evaluated joints overall. We report the *Percentage of Correct Keypoints measure in 3D* (@150mm) overall, as well as split by occlusion. Higher PCK.

	Seq1	Seq2	Seq3	Seq4	Seq5	Seq6	Total
	PCK	PCK	PCK	PCK	PCK	PCK	PCK
Overall							
Ours Multi-Person	78.7	70.0	71.9	65.2	61.4	60.7	69.0
Ours Single-Person	80.9	72.8	72.6	65.7	62.5	65.8	71.1
VNect [10]	80.1	72.4	72.4	61.5	50.2	69.8	69.4
VNect (our train. setup)	79.3	74.4	72.2	67.2	55.7	64.6	70.4
Occluded Subset of Joints							
Ours Multi-Person	73.3	66.5	55.0	56.5	45.1	64.9	62.8
Ours Single-Person	74.9	63.2	59.0	54.2	48.0	68.4	64.0
VNect [10]	61.4	54.5	47.6	36.4	30.5	66.2	53.2
VNect (our train. setup)	69.6	61.9	49.0	50.8	43.5	63.4	59.2
Un-occluded Subset of Joints							
Ours Multi-Person	79.9	70.5	73.7	66.2	64.6	59.5	70.0
Ours Single-Person	82.1	74.0	74.1	67.0	65.3	65.1	72.2
VNect [10]	83.9	74.6	75.0	64.4	54.0	70.9	72.1
VNect (our train. setup)	81.3	76.0	74.6	69.0	58.1	64.8	72.2

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] C.-H. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[4] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(7):1325–1339, 2014.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678, 2014.

[6] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*, 2010. doi:10.5244/C.24.12.

[7] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[8] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[9] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*, 2017.

[10] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36, 2017.

[11] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[12] B. X. Nie, P. Wei, and S.-C. Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *IEEE International Conference on Computer Vision*, 2017.

[13] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[14] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[15] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 398–407, 2017.