H3WB: Human3.6M 3D WholeBody Dataset and Benchmark (*Supplementary Material*)

Yue Zhu Nermin Samet David Picard
LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France
{yue.zhu, nermin.samet, david.picard}@enpc.fr

1. Supplementary material – Overview

In this supplementary material:

- We share the link to download the H3WB dataset annotation files (Section 2);
- We provide the H3WB 3D whole-body dataset keypoint layout with 133 keypoints. H3WB dataset follows exactly the same layout as COCO WholeBody dataset [4] (Section 3);
- We provide the statics regarding the diversity of H3WB dataset (Section 4);
- We present web interface of the quality assessment (Section 5);
- We provide more qualitative results for all tasks, as well as qualitative results in the wild evaluated on the COCO dataset (Section 6);
- We study failure cases from SMPL-X extracted from the literature (Section 7);
- We report the results of our 5-fold cross-validation experiments (Section 8);
- We clarify long-term support planning and the license issue (Section 9).

2. H3WB annotations

To download the H3WB dataset annotations click here. The zip file contains following:

- 2Dto3D_train.json has the training annotations for 2D→3D and I2D→3D tasks. Since this file is too big, we split it into 4-parts to ease the training and data loading pipeline. We provide the splitted files as well.
- RGBto3D_train.json has the training annotations for RGB→3D task.
- 2Dto3D_test_2d.json and I2Dto3D_test_2d.json include test instances for 2D→3D and I2D→3D tasks, respectively.
- RGBto3D_test_img.json includes test samples for RGB→3D task.

3. H3WB dataset keypoint layout

We use the COCO WholeBody dataset layout with 133 keypoints illustrated in Figure 1. H3WB dataset has the same keypoints for the whole-body layout.

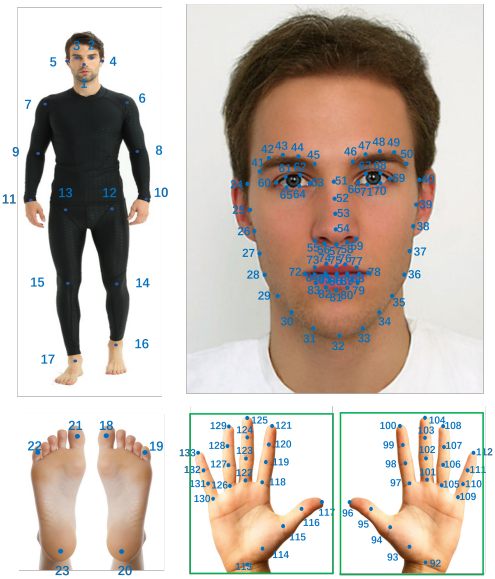

Figure 1. Whole-body keypoint layout defined in the COCO WholeBody dataset [4]. H3WB dataset follows exactly the same layout. H3WB dataset has total of 133 keypoints annotations for each human: 17 human body keypoints (top-left), 68 face (top-right), 42 hand (21 keypoints for each) (bottom-right) and 6 foot (3 for each) (bottom-left). Image source: https://github.com/jin-s13/COCO-WholeBody

H36M	**602.7**	540.0	576.3	569.3	578.7	512.8	513.3	527.7	545.3	551.3	552.8	556.5	525.7	518.9	534.8	584.5	624.1	637.4
H3WB	**518.8**	437.9	433.5	444.3	428.9	453.6	422.3	473.1	427.5	505.0	440.2	519.1	419.9	456.1	430.5	462.6	440.3	473.1

Table 1. Standard deviation in mm on average (1st column) and for each of the original 17 body joints.

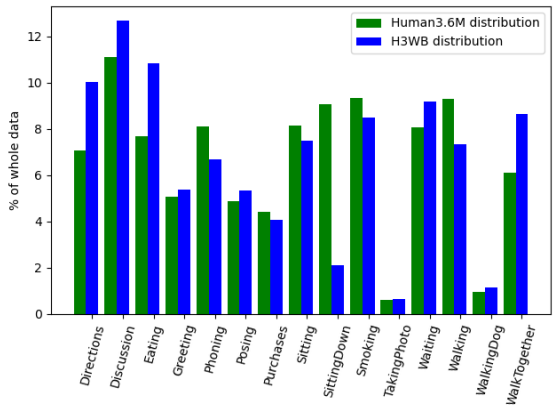

Figure 2. Distributions of Human3.6 and H3WB datasets per action class

4. Dataset diversity

The distribution of pose per action for H36M and H3WB using the original action labels is shown in Figure 2. Apart from *SittingDown*, they are about the same. Quantitatively, we show the standard deviation in mm on average (bold) and for each of the original 17 body joints in Table 1 which shows H3WB has slightly lower diversity than H36M, but no collapse.

5. Quality assessment study

We assessed the quality of the H3WB dataset by manually annotating 80K keypoints from 600 randomly selected images from the dataset. We presented a web interface to annotators and ask them to zoom-in on the body parts and correct mis-aligned keypoints by drag and drop. Sample screenshots from our web interface are presented in Figure 3.

6. More Qualitative Results

We provide more qualitative outputs obtained by Large SimpleBaseline [6] and Jointformer [5] models in Figure 4. Despite slight mis-alignments, the predicted skeletons are realistic.

We also show some examples in Figure 5 of a model trained on our H3WB benchmark for the task I2D→3D and evaluated on COCO dataset[4]. We can see that even when there are missing points in the 2D input, the model still can predict the 3D wholebody pose accurately. This validates the usefulness of the I2D→3D in real world scenario.

7. SMPL-X failure cases

Parametric body models like SMPL-X have many seminal advantages such as always producing biologically plausible poses or taking into account the shape of the person. This enables powerful applications, for example in augmented reality or animation. However, because very accurate pose is not a requirement in these application, a model like SMPL-X is not yet able to reach satisfactory accuracy, especially on extremities like the hands and the feet. This is what we show in Figure 6, where we extracted images from several articles [8, 9, 2] and zoom on the extremities to visually assess that it is well below the accuracy provided in H3WB. We also ran SMPL-X on Human3.6M to see if it can be used to generate pseudo-labels and show selected zooms on the extremities on Figure 7. Here also, the accuracy is well below what our label generation process managed to get. As such, datasets relying on SMPL-X for their groundtruth are thus by design less accurate and thus not usable for accurate pose estimation, especially on the extremities. Furthermore, assessing quantitatively the accuracy is almost impossible to do with these methods, whereas we provide an estimate for H3WB showing our benchmark is rigorous.

8. Cross validation experiments

We do not provide a validation set for the H3WB dataset. We recommend 5-fold cross-validation for model selection and hyper-parameters tuning. We split the training set into 5 sets. We take the set cv$i$ as a hold out (test set), use remaining sets to train the models, and report the results on cv$i$. We present the cross-validation results together with the test set results in Tables 2, 3, 4 for all tasks. We observe that cross-validation results are consistent and compatible with the test results which are listed in the main paper.

9. Other issues

We plan to setup a server for test set evaluation. We will also release the test data after 3-5 years once they are well studied to allow long term use without relying on our evaluation server.

Concerning the license, we only release entirely new labels, which fits the license agreement allowing research output.

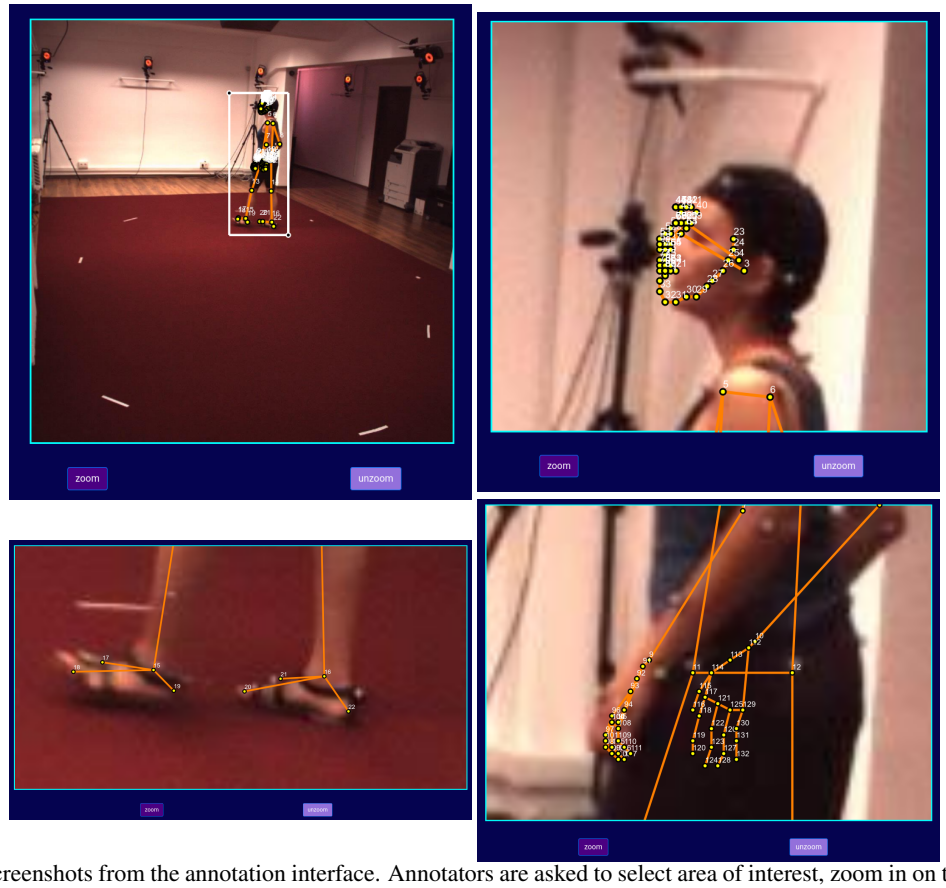

Figure 3. Sample screenshots from the annotation interface. Annotators are asked to select area of interest, zoom in on that area and correct the mis-aligned keypoints by drag-drop.

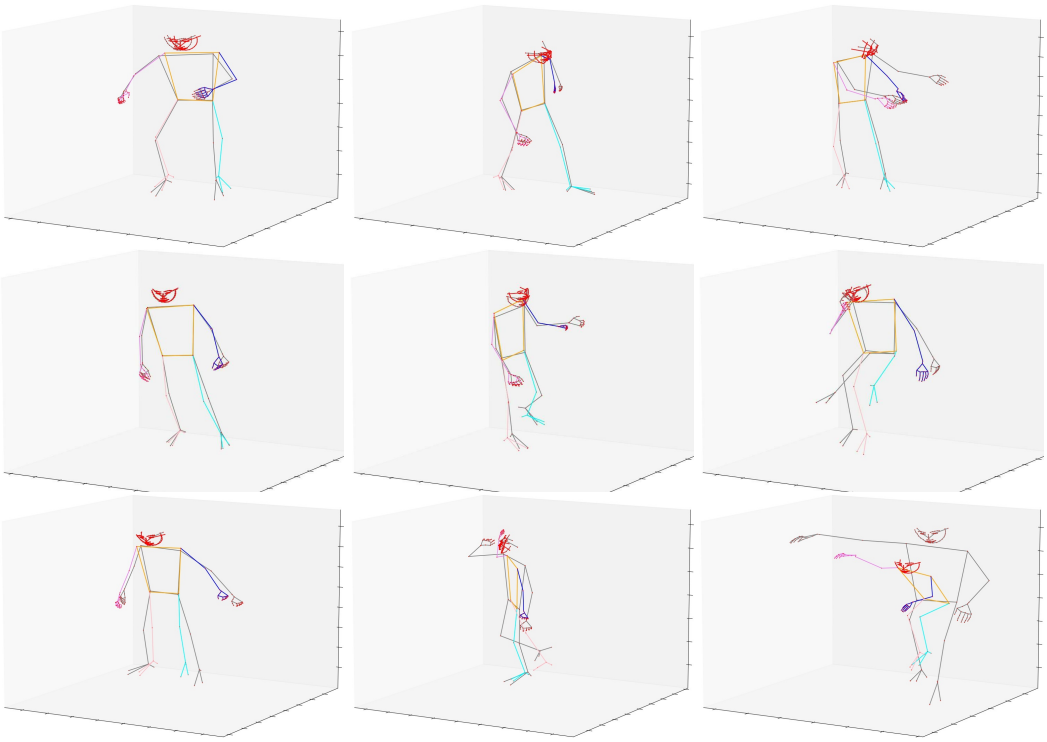

Figure 4. Example predictions from Large SimpleBaseline model for 2D→3D (1st row) and I2D→3D (2nd row) tasks. 3rd row shows predictions from Jointformer for RGB→3D task. Colored skeletons correspond to predictions and gray skeletons correspond to groundtruths. First two columns show almost-aligned successful front/side predictions, and the last column shows slightly mis-aligned predictions.

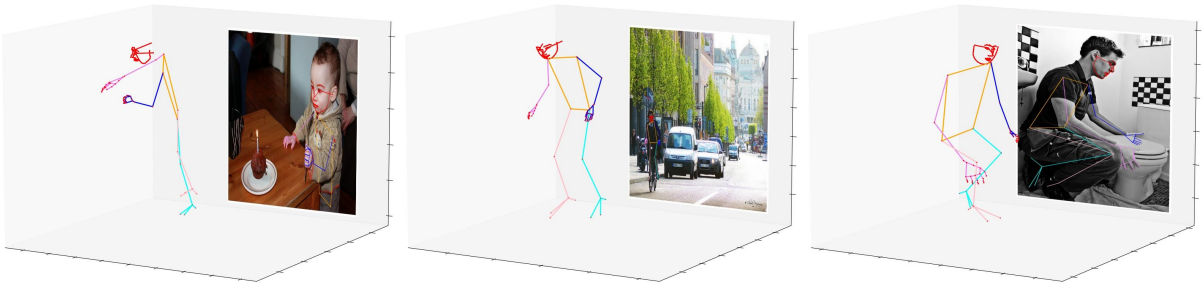

Figure 5. Visual examples of lifting on COCO. The labels on the images are the incomplete inputs.

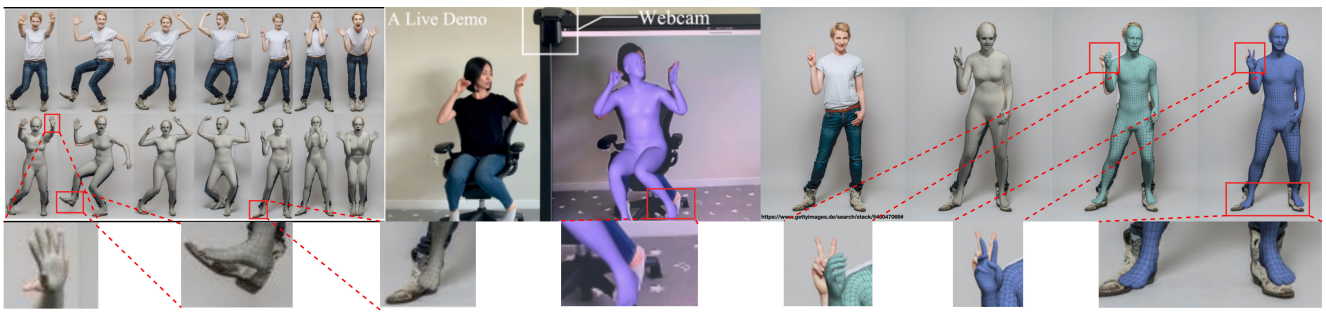

Figure 6. Several examples of failures on hands and feet with SMPL-X model copied from [8, 9, 2]

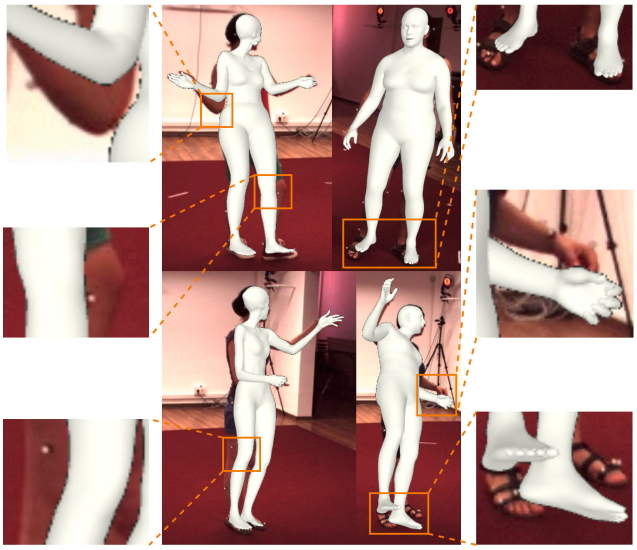

Figure 7. Our runs with SMPL-X models on the annotation, it is not as visually accurate as we require as annotations.

method	all	body	face	face †	hand	hand ‡
SimpleBaseline [6]						
cv1	134.0	128.9	126.8	34.8	148.4	46.2
cv2	128.9	126.4	120.5	29.0	136.7	42.8
cv3	136.0	130.6	135.8	23.6	139.1	44.0
cv4	132.8	131.4	126.9	29.4	143.3	46.6
cv5	139.9	139.9	133.6	33.3	150.0	46.6
Cv std	4.1	5.1	6.1	4.4	5.7	1.7
Cv mean	134.3	131.4	128.7	30.0	143.5	45.2
Test	125.4	125.7	115.9	24.6	140.7	42.5
Large SimpleBaseline [6]						
cv1	106.8	105.1	105.8	22.7	109.2	33.6
cv2	103.9	104.3	107.7	16.7	97.6	31.9
cv3	101.8	102.4	105.5	14.2	95.6	30.6
cv4	108.7	107.0	111.6	14.3	105.0	32.3
cv5	111.8	109.0	112.0	15.6	113.1	35.5
Cv std	3.9	2.5	3.1	3.5	7.5	1.9
Cv mean	106.6	105.6	108.5	16.7	104.1	32.8
Test	112.3	112.6	110.6	14.6	114.8	31.7
CanonPose [10]						
cv1	173.4	177.8	180.0	30.3	160.4	46.4
cv2	152.9	160.7	162.0	23.2	133.9	44.2
cv3	163.9	167.4	176.5	21.1	141.5	44.9
cv4	185.0	187.4	199.2	23.0	160.5	48.1
cv5	172.6	177.9	182.2	22.1	154.1	46.8
Cv std	12.0	10.4	13.3	3.7	11.9	1.6
Cv mean	169.6	174.2	180.0	23.9	150.1	46.1
Test	186.7	193.7	188.4	24.6	180.2	48.9
CanonPose [10] + 3D sv.						
cv1	121.1	121.9	116.8	27.6	127.5	42.1
cv2	115.4	118.6	116.4	20.6	111.9	40.6
cv3	112.4	113.2	113.7	16.5	110.0	38.9
cv4	116.2	117.9	115.5	17.5	116.2	40.2
cv5	168.7	170.5	180.3	22.1	149.0	49.2
Cv std	23.7	23.7	29.0	4.4	16.1	4.1
Cv mean	126.8	128.4	128.5	20.9	122.9	42.2
Test	117.7	117.5	112.0	17.9	126.9	38.3
Jointformer [5]						
cv1	94.3	85.0	76.0	29.8	129.0	48.1
cv2	87.4	80.0	71.2	21.6	117.8	47.0
cv3	94.5	86.3	84.5	16.5	115.3	49.2
cv4	91.4	88.1	74.6	16.5	123.7	48.8
cv5	104.3	96.3	82.6	19.0	143.9	53.5
Cv std	6.2	5.9	5.6	5.5	11.4	2.5
Cv mean	94.4	87.1	77.8	20.7	125.9	49.3
Test	88.3	84.9	66.5	17.8	125.3	43.7

Table 2. Results for 2D→3D task on each 5-fold and test sets. Results are shown for MPJPE metric. All results are pelvis aligned, except † and ‡ show nose and wrist aligned results for face and hands, respectively. Sv. is supervision.

method	all	body	face		hand	
				†		‡
SimpleBaseline [6]						
cv1	259.9	242.9	220.1	40.9	333.9	84.0
cv2	271.0	244.9	229.3	34.9	352.7	86.7
cv3	268.6	251.3	237.9	33.2	327.8	83.7
cv4	259.1	246.8	225.5	33.8	320.4	81.9
cv5	269.7	251.0	226.4	33.1	350.0	87.9
Cv std	5.7	3.7	6.5	3.3	14.0	2.4
Cv mean	265.7	247.4	227.8	35.2	337.0	84.8
Test	268.8	252.0	227.9	34.0	344.3	83.4
Large SimpleBaseline [6]						
cv1	137.7	130.8	134.9	33.3	146.2	47.5
cv2	125.5	124.9	123.6	23.1	129.1	46.0
cv3	126.3	124.6	125.7	19.6	128.0	44.7
cv4	136.1	129.9	134.7	19.9	141.7	47.3
cv5	139.0	135.7	133.4	21.4	149.8	51.2
Cv std	6.5	4.6	5.4	5.7	9.9	2.4
Cv mean	132.9	129.2	130.5	23.5	139.0	47.3
Test	131.4	131.6	120.6	19.8	148.8	44.8
CanonPose [10]						
cv1	256.7	237.1	278.9	39.1	231.4	55.1
cv2	255.5	244.2	284.1	35.6	215.4	56.1
cv3	261.4	245.0	291.2	31.5	222.2	54.8
cv4	261.3	243.4	285.5	31.6	231.7	56.8
cv5	270.6	250.2	292.5	35.0	246.2	61.0
Cv std	5.9	4.7	5.5	3.2	11.6	2.5
Cv mean	261.1	244.0	286.4	34.6	229.4	56.8
Test	285.0	264.4	319.7	31.9	240.0	56.2
CanonPose [10] + 3D sv.						
cv1	163.6	155.7	160.2	33.7	173.5	49.1
cv2	158.5	153.0	161.0	25.8	157.4	48.0
cv3	157.9	150.0	161.5	21.8	156.5	47.3
cv4	157.3	154.1	155.5	22.7	162.1	49.0
cv5	175.1	168.9	169.3	25.4	187.9	55.5
Cv std	7.5	7.3	5.0	4.7	13.3	3.3
Cv mean	162.5	156.3	161.5	25.9	167.5	49.8
Test	163.6	155.9	161.3	22.2	171.4	47.4
Jointformer [5]						
cv1	121.5	114.8	100.9	34.3	158.6	55.9
cv2	112.5	104.9	93.4	25.2	147.5	56.6
cv3	110.5	101.2	94.3	20.6	141.9	56.2
cv4	123.5	115.7	104.5	21.1	158.7	58.2
cv5	129.4	116.0	107.9	22.5	171.6	61.1
Cv std	7.9	7.0	6.3	5.6	11.5	2.1
Cv mean	119.5	110.5	100.2	24.7	155.7	57.6
Test	109.2	103.0	82.4	19.8	155.9	53.5

Table 3. Results for I2D→3D task on each 5-fold and test sets. Results are shown for MPJPE metric. All results are pelvis aligned, except † and ‡ show nose and wrist aligned results for face and hands, respectively. Sv. is supervision. We observe that *CanonPose* fails to generalize to new subject in the test set and performs worse on the test set.

method	All	Body	Face		Hand	
				†		‡
SHN [7]+SimpleBaseline [6]						
cv1	191.0	177.9	159.8	41.4	248.7	66.1
cv2	159.4	151.0	135.8	30.4	202.3	62.6
cv3	170.8	169.9	157.0	25.8	193.9	64.7
cv4	204.8	202.3	192.9	27.7	225.5	68.0
cv5	204.8	192.7	173.8	30.2	261.5	71.8
Cv std	20.4	20.0	21.2	6.1	29.0	3.5
Cv mean	186.2	178.8	163.9	31.1	226.4	66.6
Test	182.5	189.6	138.7	32.5	249.4	64.3
CPN [1]+Jointformer[5]						
cv1	100.8	101.6	75.5	29.9	141.3	53.5
cv2	91.9	89.8	70.6	22.8	127.5	52.9
cv3	75.7	77.5	62.5	14.9	96.0	51.0
cv4	78.1	82.0	58.4	16.9	107.6	52.7
cv5	100.8	98.3	73.1	19.5	147.2	59.0
Cv std	12.1	10.3	7.3	5.9	21.8	3.0
Cv mean	89.5	89.8	68.0	20.8	123.9	53.8
Test	132.6	142.8	91.9	20.7	192.7	56.9
Resnet50 [3]						
cv1	123.8	117.7	97.6	34.9	169.6	58.3
cv2	111.8	107.1	88.3	25.3	152.4	57.2
cv3	102.5	103.8	81.9	20.0	135.0	57.8
cv4	113.5	114.5	89.9	21.2	151.3	58.5
cv5	122.8	119.5	91.7	23.1	175.1	62.6
Cv std	8.8	6.8	5.78	5.9	16.0	2.1
Cv mean	114.9	112.5	89.9	24.9	156.7	58.9
Test	166.7	151.6	123.6	26.3	244.9	63.1

Table 4. Results for RGB→3D task on each 5-fold and test sets. Results are shown for MPJPE metric. All results are pelvis aligned, except † and ‡ show nose and wrist aligned results for face and hands, respectively.

References

[1] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *CVPR*, 2017. 6

[2] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 2, 4

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2015. 6

[4] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 1, 2

[5] Sebastian Lutz, Richard Blythman, Koustav Ghosal, Matthew Moynihan, Ciaran Simms, and Aljosa Smolic. Jointformer: Single-frame lifting transformer with error prediction and refinement for 3d human pose estimation. *ArXiv*, 2022. 2, 5, 6

[6] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 2, 5, 6

[7] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *ECCV*, 2016. 6

[8] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed Osman, Dimitrios Tzionas, and Michael Black. Expressive body capture: 3D hands, face, and body from a single image. *CVPR*, 2019. 2, 4

[9] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. *ICCV*, 2021. 2, 4

[10] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *CVPR*, 2021. 5, 6