

Supplementary Materials: “SATPose: Improving Monocular 3D Pose Estimation with Spatial-aware Ground Tactility”

Anonymous Authors

1 SOURCE CODE

We will release the source code and pretrained models of the pressure image reconstruction network and the multimodal 3D human pose estimation network upon acceptance.

2 DATA COLLECTION DETAILS

2.1 Action Explanations

Table 1 provides a detailed description of the 16 human actions and their distribution within the dataset. Throughout the experiments, we strategically arranged the sequence of high-intensity and low-intensity actions to interleave, thus mitigating excessive fatigue from continuous exercises.

Table 1: Explanation of the 16 human actions in the PVM dataset, along with their respective number of frames in the training and testing sets.

Actions	Explanation	Duration	Train	Test
Deep Squat	Raise both hands overhead and squat down until thighs are parallel to the ground.	90s	28041	7038
Standing on Toes	Lift the heels to balance on the balls of the feet.	90s	29319	7216
Hurdle Step	Lift one leg over an obstacle while maintaining balance and stability.	L/R leg each 6 times in L/R dir.	33843	7580
Inline Lunge	Extend one leg forward and kneel down with the other leg, keep both legs in a line.	L/R leg each 6 times in L/R dir.	34351	7836
Marching in Place	Lift the knees alternately while staying in the same spot.	90s	21469	5524
Arm-waving	Rhythmically swing the arms back and forth accompanied by slight knee bending.	90s	31206	7125
Direction	Hand movements accompanied by tilting the body towards the direction of the hand.	90s	40437	11072
Waist Turning	Twist the waist in a circular motion.	90s	29909	7585
Twisting	Twist the body back and forth from left to right while keeping both feet in place.	90s	30335	7682
Bending	Bend the body alternately toward the left and right toes.	90s	30380	7882
Sitting	Sit on the pressure mat in any comfortable position.	90s	29405	7577
Active Straight-Leg Raise	Lie on the back and alternately lift the legs straight up.	45s each in L/R dir.	31511	7798
Rotary Stability	Lift and extend the arms and legs in opposite directions in a kneeling prone position.	L/R leg each 6 times in L/R dir.	28939	6489
Push-ups	Push-ups / kneeling push-ups.	45s each in L/R dir.	27763	6790
Posing	Casually strike poses as if being photographed.	90s	28640	6883
Waiting	Stand straight or shift weight towards one leg.	90s	30188	7710
Total			485736	119787

2.2 Reflective Markers on the Mocap Suit

As shown in Fig.1, we utilized the 53-point human body marking method to label various joints of the human body, ultimately recording the 3D position information of 18 body joints.



Figure 1: An illustration of the 53-point human body marking method and the 18-joint human skeleton.

2.3 2D Skeleton Construction

The open-source human pose estimation library, OpenPose [1], is utilized to extract 2D skeletons from monocular images. Specifically, we use the OpenPose model with 25 skeleton keypoints, as it exhibits greater overlap with the ground truth 3D skeleton. For the extracted 25 skeleton keypoints, we:

- Remove 6 keypoints corresponding to eyes, ears, and heels;
- Remove 4 keypoints corresponding to big toes and little toes;
- Add 2 keypoints for the toes calculated from the midpoints of the left and right big toes and little toes;
- Add a new spine keypoint calculated as the midpoint between the neck and hip keypoints.

Finally, we obtain 2D skeleton with 18 keypoints corresponding one-to-one with the 3D skeleton keypoints.

3 COMPARISON WITH SOTAS ON PROTOCOL #1

Another variant of MPJPE (mean per joint position error), *i.e.*, Protocol #1, is a commonly used metric in previous works to measure the accuracy of 3D human pose estimation. It calculates the average error of each joint of the human body without considering global displacement. Specifically, it computes the Euclidean distance between the predicted joints and ground truth joints in the local coordinate system after aligning the root joint of the predicted skeleton with the root joint of the ground truth skeleton.

Table 2: Quantitative comparison results with state-of-the-art methods on protocol #0. Bold and underline indicate the best and second-best values.

Protocol #1	D.S.	S.T.	H.S.	I.L.	M.P.	A.W.	Direction	W.T.	Twisting	Bending	Sitting	A.S.L.R.	R.S.	P.U.	Posing	Waiting	Avg.
Shan et al. [5] MM2021	49.7	39.3	47.7	49.9	44.3	53.3	47.9	54.6	47.1	55.0	42.4	46.8	68.9	57.3	52.8	45	50.1
Li et al. [2] ToMM2022	44.2	35.6	51.2	51.1	44.7	49.0	<u>42.6</u>	43.3	48.4	44.8	43.7	46.3	68.1	56.6	51.0	42.3	47.7
Shan et al. [4] ECCV2022	43.7	36.0	47.0	<u>49.6</u>	<u>42.2</u>	<u>48.0</u>	<u>42.7</u>	<u>41.7</u>	44.3	47.9	44.6	62.8	69.4	62.3	<u>49.0</u>	42.0	48.3
Li et al. [3] CVPR2022	41.8	37.1	48.8	49.0	48.1	48.7	40.6	44.2	49.2	57.4	46.6	48.2	74.2	54.4	48.6	39.5	48.5
Zhang et al. [7] CVPR2022	46.4	35.6	54.0	59.7	46.0	56.5	45.2	57.6	52.9	52.6	46.7	55.4	77.1	56.3	52.7	42.5	52.3
Zhao et al. [8] CVPR2023	50.6	37.0	48.3	51.2	47.0	55.0	47.2	43.4	47.5	57.9	44.1	<u>46.6</u>	69.0	58.4	53.3	43.1	50.0
Yu et al. [6] ICCV2023	40.8	37.1	49.2	52.4	42.7	45.6	48.8	47.4	46.3	52.3	49.4	52.0	80.8	59.8	53.9	42.2	50.0
Ours (pred. pressure)	42.8	<u>33.9</u>	48.3	53	43.5	51.9	44.0	42.8	43.7	<u>45.3</u>	42.0	46.9	66.8	55.6	49.4	<u>37.8</u>	<u>46.7</u>
Ours	<u>41.4</u>	33.8	<u>47.4</u>	51.5	41.7	51.3	42.8	40.3	<u>43.9</u>	47.4	<u>42.1</u>	47.9	<u>66.9</u>	<u>54.8</u>	49.7	37.6	46.3

In comparison to protocol #0, protocol #1 eliminates the global displacement bias, focusing solely on the error in the pose itself. This results in a reduced error gap between different methods, yet our method still outperforms others. The position of pressure traces provides significant gains in depth information, enabling substantial improvements over monocular methods, especially in challenging actions like sitting and rotary stability amidst occlusions. The pressure distribution within these traces offers finer guidance on centroid shifts, aiding in achieving high-precision estimation of actions such as standing on toes, waist turning, twisting, and waiting. Additionally, the disparity between results from protocol #0 and protocol #1 reflects the exceptional spatial perception capabilities of pressure images, offering robust guidance for global 3D pose estimation.

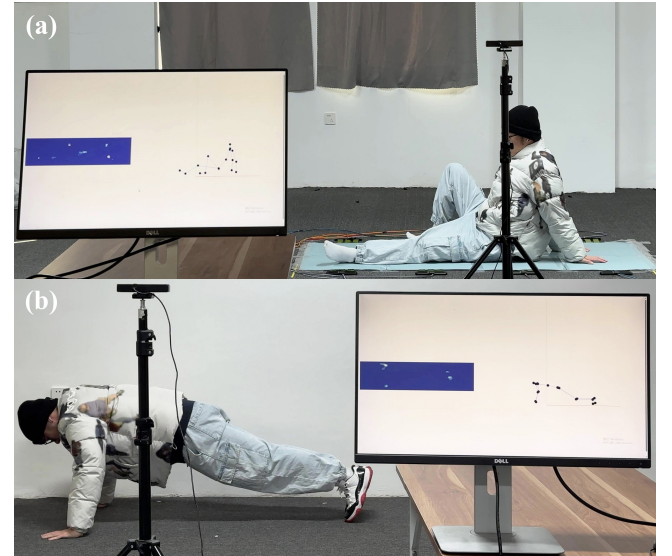
4 LIVE DEMO

To illustrate the viability of our multimodal 3D human pose estimation approach, we implemented a real-time demonstration. The workflow encompasses i) the real-time acquisition of 2D pose and pressure image sequence, ii) the input of these data into the 3D human pose estimation network to derive the output 3D pose, and iii) the transmission of the predicted 3D pose to the web interface for real-time visualization.

Two distinct implementations are employed for acquiring pressure images: one entails the real-time collection of sensor data from a pressure mat, and the other reconstructs pressure images in real-time from the 2D poses. As shown in Fig.2, the former achieves relatively precise pose estimation but necessitates a more intricate hardware setup (comprising a monocular camera and a pressure mat); the latter compromises a slight degree of accuracy to accommodate a lighter hardware system (solely requiring a monocular camera). These alternative implementations cater to diverse scenarios and applications, ensuring the adaptability of our methodology. Additionally, as real-time prediction lacks access to future data, and considering that the input to the pose estimation network during training comprises information from future frames, we address this limitation by duplicating future frames based on the content available in the current frame.

The real-time system estimates poses at a frame of approximately 10Hz. The time consumption for 2D pose extraction, pressure image reconstruction, and 3D pose prediction is reported to be approximately 0.098 seconds, 0.002 seconds, and 0.017 seconds, respectively. The most notable delay stems from the 2D pose extraction with

OpenPose, and future endeavors may explore swifter alternatives. Overall, our approach demonstrates reliable pose prediction results while ensuring a smooth user experience in system usage.

**Figure 2: Real-time demonstration with monocular images and real (a) / reconstructed (b) pressure images as input.**

5 VIDEO

We provide a video containing the action demonstrations from the dataset, the dynamic visualizations of the pressure image reconstruction and 3D pose estimation results, as well as the live demo of the multimodal pose tracking system.

REFERENCES

- [1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.
- [2] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. 2022. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia* 25 (2022), 1282–1293.
- [3] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. 2022. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13147–13156.
- [4] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. 2022. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *European Conference on Computer Vision*. Springer, 461–478.

- [5] Wenkang Shan, Haopeng Lu, Shanshe Wang, Xinfeng Zhang, and Wen Gao. 2021. Improving robustness and accuracy via relative information encoding in 3d human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3446–3454.
- [6] Bruce XB Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. 2023. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8818–8829.
- [7] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. 2022. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13232–13242.
- [8] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. 2023. PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8877–8886.

233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348