Smart Insole: Predicting 3D human pose from foot pressure

Isaac Han[1] **Seoyoung Lee**[1] **Sangyeon Park**[1] **Ecehan Akan**[1] **Yiyue Luo**[2]

Kyung-Joong Kim[1]

Abstract

Footwear is typically worn during a range of daily activities, offering a valuable opportunity for integrating technologies like human pose estimation using embedded sensors. This study introduces a novel method of 3D human pose estimation using foot pressure data captured by a low-cost, high-resolution smart insole equipped with over 600 pressure sensors per foot. In contrast to previous works that used carpet-type sensors, which are limited to functioning only within a localized environment, our wireless smart insole enables pose estimation regardless of the user's location. We collect synchronized tactile and visual data across various actions. Utilizing a camera-based pose estimation model for supervision, we design a deep neural network to predict 3D human poses using only foot pressure data. Furthermore, integrating a simple linear classifier with our model's learned representations achieves successful classification of various daily activities.

1 Introduction

Estimating human pose is crucial for a range of applications, including robotics [Maurice et al., 2019, McColl et al., 2011, Ji et al., 2019], healthcare [Obdržálek et al., 2012, Zhou et al., 2020, Meng et al., 2020, Jalal et al., 2014, Liu et al., 2022], and gaming [Ke et al., 2010]. Recent methodologies employing images and videos to predict human poses have seen substantial advancements. Nonetheless, these vision-based methods face issues like occlusion and privacy concerns. Tactile-based approaches offer a solution to these challenges by providing non-intrusive sensing without relying on visual data and continuous sensing, thereby ensuring privacy and not being affected by occlusion issues.

Previous works have considered carpet-type sensors to estimate 3D human poses [Luo et al., 2021, Liu et al., 2024, Tripathi et al., 2023, Clever et al., 2020, Kasman and Moshnyaga, 2017, Shi et al., 2020, Casas et al., 2019, Clever et al., 2018]; however, such sensors are constrained by spatial limitations as they require a large area for installation and movement range is determined by the sensor's size. In contrast, insole-type wearable sensors offer greater movement flexibility without size restrictions. Human pose estimation in this context holds significant potential for various applications, as shoes are commonly worn in daily activities. Although some studies have begun to explore lower-body pose prediction from tactile signals [Alemayoh et al., 2023, Tam et al., 2019], previous efforts using plantar pressure sensors were limited by low-resolution technology, impacting performance. A particularly challenging problem is predicting 3D human poses involved in complicated daily activity and exercises, based solely on foot pressure, which represents only a small part of the body.

To address this, we developed a smart insole—a tactile sensing array embedded within a pair of wireless insoles, featuring over 600 pressure sensors per foot, providing real-time, high-resolution

[1]Gwangju Institute of Science and Technology
[2]University of Washington

NeurIPS 2024 Workshop on Touch Processing: From Data to Knowledge.

recordings of foot pressure. Using this hardware, we collected over 105,000 synchronized tactile and visual frames from five individuals performing various activities, including walking, squatting, and lunging. Unlike previous works that have utilized carpet-type sensors which only work locally in the environment, our wireless smart insole enables pose estimation regardless of where the person locates. We designed a deep learning model that predicts full-body 3D human poses directly from raw tactile frames. Our model demonstrates precise predictions, achieving an average localization error of 7.43 cm compared to the ground truth pose derived from visual data. By integrating a simple linear classifier with our model's learned representations, we attain a 96% accuracy rate in classifying daily activities. Our results also indicate that our model successfully generalizes to unseen actions and individuals. We further analyze these findings through ablation studies.

2 Method

2.1 Data collection

Our smart insole system incorporates a wireless piezoresistive pressure sensor insole, featuring over 600 sensors per foot, designed for high-resolution tactile data acquisition. Each insole consists of a grid formed by orthogonally aligned copper threads as electrodes on both sides of piezoresistive films. The intersections of these electrodes are the sensors, allowing for precise pressure mapping. Costing around $50 per unit, these insoles are cost-effective and easy to manufacture, ideal for large-scale data collection. The tactile data is captured through a wireless readout circuit that records frames at a frequency of 14 Hz, with the batteries and boards attached to ankles using elastic bands.

To derive 3D human keypoints from visual frames, we employed XRmocap [XRMoCap, 2022], an open-source multi-view motion capture framework. Data was collected using six strategically positioned cameras to capture diverse viewing angles. We extracted 19 keypoints for each person, including head, neck, shoulders, elbows, wrists, hips, knees, ankles, heels, small toes, and big toes.

We normalized the x-y positions relative to the pelvis keypoint to establish a consistent reference point. To eliminate rotational redundancy, we normalized the rotation of keypoints by aligning the shoulder keypoints with the x-axis. We exclude z-axis from normalization to maintain height information, as foot pressure distribution is highly correlated to the absolute height of feet. The normalized position P_i' of each keypoint P_i can be calculated as follows:

$$P_i' = R \cdot (P_i - P_{\text{pelvis}})$$

Where R is the rotation matrix required to align the shoulders with the x-axis, and P_{pelvis} is the raw position of the pelvis keypoint.

Our high-resolution system accurately captures foot pressure, improving 3D human keypoints' prediction from tactile inputs alone. We have collected over 105,000 synchronized tactile and visual frames. This dataset was generated from recordings of five participants, each performing seven distinct actions: squatting, lunging, warm-up, walking, walking-in-place, backward walking, and side walking. We collected data from exercises like squatting, lunging, and walking, including variations. As these actions primarily involve lower-body movements, we incorporated warm-up exercises with video guidelines to ensure the inclusion of dynamic upper-body movements in our data.

2.2 Pose estimation from foot pressure

Our model leverages foot pressure data to predict 3D human poses, utilizing temporal information and minimizing noise effects. Specifically, the model inputs $M = 40$ sequences of tactile frames from sensors embedded in two distinct insoles, and learns to predict keypoints corresponding to the central frame of the input sequence. The model outputs a confidence heatmap for 19 keypoints. Final keypoint prediction is produced by applying softargmax operation on the generated heatmaps.

The architecture includes dual CNN encoders—one for each foot—which process respective pressure signals independently. The features extracted by each encoder are concatenated to form a unified feature vector. As in previous work [Luo et al., 2021], we add a 3D feature consisting of z-axis indices indicating the height of each voxel to incorporate height information. We then use 3D convolution to produce 3D heatmaps for 19 keypoints. The overview of the architecture is shown in Figure 1.

Model optimization involves two loss terms. Heatmap loss is calculated using the Mean Squared Error (MSE) between the predicted heatmaps and label heatmaps, constructed from label keypoints

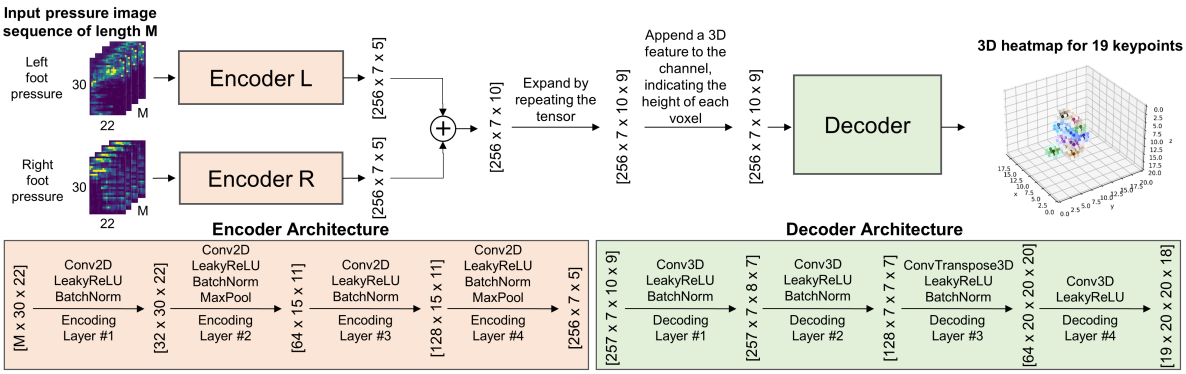


Figure 1: **CNN model for 3D human pose estimation.** Our model utilizes separate encoders for the tactile inputs from the left and right foot. After features extracted from these encoders are concatenated and expanded, a 3D feature that indicates the height of each voxel is then appended. This combined feature is fed into a decoder, which finally generates a 3D heatmap for 19 keypoints.

using a Gaussian distribution. Keypoint loss, similarly, is determined by the MSE between the predicted keypoints and the labeled keypoints. The final model loss is a weighted sum of these loss components. We employ the Adam optimizer [Kingma and Ba, 2014] to optimize model parameters.

3 Experiments

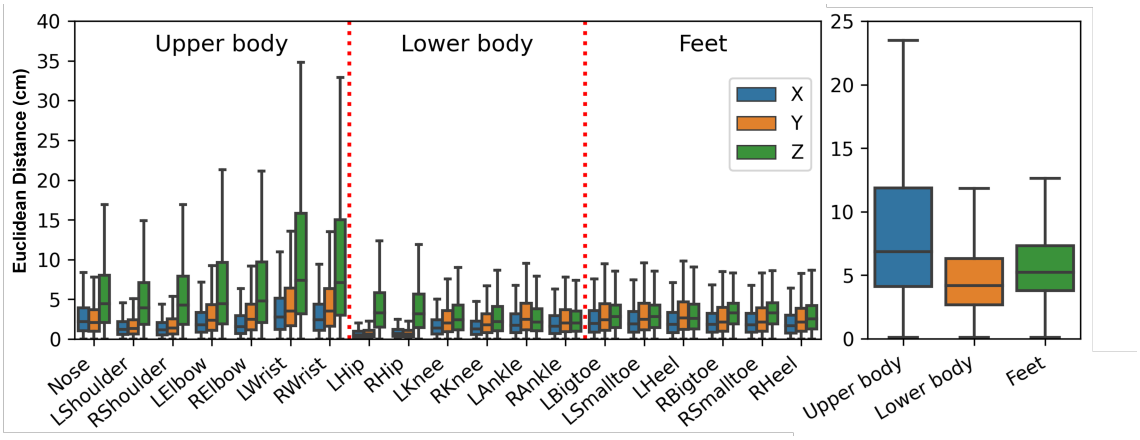


Figure 2: **Results on pose estimation (unit: cm).** Average Euclidean distance error by keypoints and axes (**left**) and the average value for each part (**right**). Lower body and feet have lower errors due to their stronger link to foot pressure.

For 3D pose estimation, we used 63,428 pairs of tactile and keypoint frames, validated on 21,156 pairs, and tested on 21,162 pairs. The evaluation metric is based on the Euclidean distance between the ground-truth 3D keypoints and the predicted 3D keypoints. Figure 2 illustrates the Euclidean distance error for each keypoint and axes of the body parts. The average Euclidean distance is 7.43cm, which indicates that the model predicts keypoints well only with foot pressure data. The result is consistent with our intuition that foot pressure is strongly influenced by lower body and foot motion. Since our model predicts human pose from foot tactile signals, the upper body shows a higher Euclidean distance error. Notably, the keypoint errors for the elbow and wrist, which are less related to foot pressure, are larger. We illustrate 5 consecutive images with keypoints and tactile insole visualization, true label of 3d keypoint and prediction of model over time in Figure 3.

To validate the importance of the two loss terms used in training the model and the significance of high-resolution tactile data, we conducted ablation studies. When both the heatmap loss and keypoint

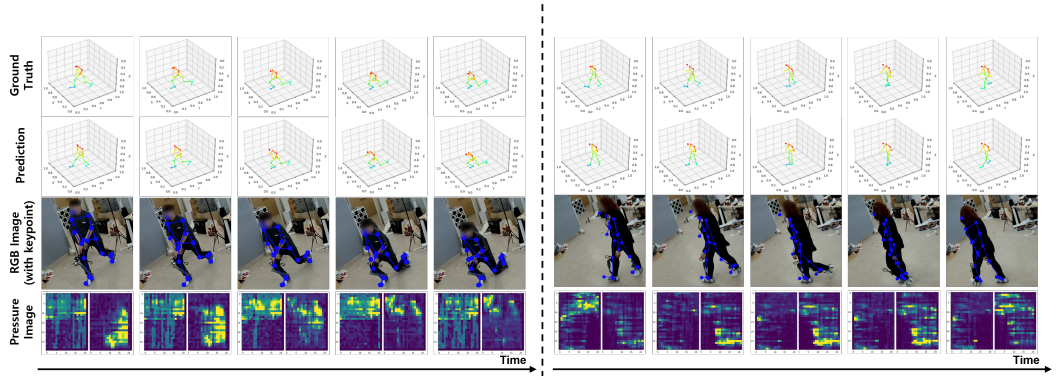


Figure 3: **Qualitative result of pose estimation.** Visualization sequence of the collected data and the predicted 3D keypoints for lunging (**left**) and walking (**right**). From top to bottom: Ground truth 3D keypoints, predicted 3D keypoints, RGB image used for visualization purposes only, and collected pressure data of the feet. The 3D pose is successfully predicted based on the tactile data used as input.

Generalization	Seen	Unseen
Squat	8.30	14.78
Warm-up	9.93	11.11
Walking	6.42	7.44
Walking-in-place	5.54	6.39
Side Walking	9.54	7.33
Backward walking	5.69	6.56
Lunge	6.71	14.37
Individual	8.00	9.43

Action classification	Accuracy (%)
Squat	98.68
Warm-up	99.30
Walking	88.43
Walking-in-place	99.84
Side Walking	96.83
Backward walking	97.19
Lunge	98.35
Avg	96.88

Figure 4: **Average Euclidean distance table of generalization task (unit: cm) (left)**. Average Euclidean distance for the generalization task across unseen actions and individuals. **Accuracy table of Action classification result (right)**. The action classification result of a linear classifier on the representation learned from the pose estimation model.

loss are used, the average Euclidean distance (cm) is 7.43. In contrast, using only the heatmap loss results in 8.18, and using only the keypoint loss results in 7.98. Additionally, as the tactile resolution was reduced by a quarter (from 30×22 to 15×11, 8×6 and 4×3) during training, the average Euclidean distance was 8.41, 17.50, and 25.98, respectively. It demonstrates that both loss terms and high-resolution data are essential for low test error.

We evaluated the generalization ability of the proposed model on unseen individuals and actions. As shown in the table in Figure 4, the model shows strong generalization performance, with an insignificant increase in Euclidean distance error when applied to data from unseen individuals. For unseen actions, the generalization performance varies by action. In the case of backward walking, the difference in average Euclidean distance between seen and unseen data is negligible, while the lunging data shows a larger difference. This is likely due to the lack of actions in the dataset that have pressure distributions similar to those of lunging. Walking and side walking are included in the data distribution, which allows the model to successfully generalize to backward walking. These findings emphasize the need for more diverse action data to improve the generalization performance, especially for further complex applications.

To analyze the representational capability of the trained model, we conducted an action classification experiment using the model's learned representations. We added a simple linear layer on the top of the model's encoders. The features from the two encoders are concatenated and flattened, finally used as an input for the linear classifier. We froze the encoders and trained the classifier using the same training and validation datasets used for human pose estimation. The classification accuracy for each action is shown in Figure 4. Our model achieved 96.88% accuracy on 21,162 pairs of test data, suggesting the model learned semantically meaningful representations of tactile frames.

4 Conclusion

We developed a high-resolution wireless tactile sensing system integrated into everyday footwear, enabling accurate 3D human pose estimation from foot tactile signals using a deep learning model. By incorporating a simple linear classifier, the system also classifies daily activities with high accuracy. This privacy-preserving approach offers a novel solution for pose estimation with potential applications in robotics, healthcare, and gaming.

5 Acknowledgement

This work was supported by the GIST-MIT Research Collaboration grant funded by the GIST in 2024. This research was supported by 'Project for Science and Technology Opens the Future of the Region' program through the INNOPOLIS FOUNDATION funded by Ministry of Science and ICT (Project Number: 2022-DD-UP0312)

6 Supplementary Material

6.1 Implementation details

We implemented our deep learning model using PyTorch [Paszke et al., 2019], a famous deep learning framework. Our model is trained with a learning rate of $1e^{-4}$ and a batch size of 32. As shown in Figure 1, our model consists of two encoders and a decoder with 4 layers. The encoder layers use a $(3, 3)$ kernel with padding size 1, and the 2nd and 4th layers use max pooling with size $(2, 2)$. The decoder consists of three conv3D layers and one transposed conv3D layer. Each decoder layer has a different kernel size to adjust the shape of the features to produce a $(20, 20, 18)$ heatmap from $(7, 10, 9)$ voxels. From the 1st to the 4th layer, they have kernel sizes of $(3, 5, 5)$, $(3, 4, 3)$, $(2, 2, 2)$, $(3, 3, 5)$, respectively. All encoder and decoder layers use batch normalization and leakyReLU activation. We clip the output heatmap values to a range between 0 and 1 to perform the softargmax operation.

6.2 Hardware details

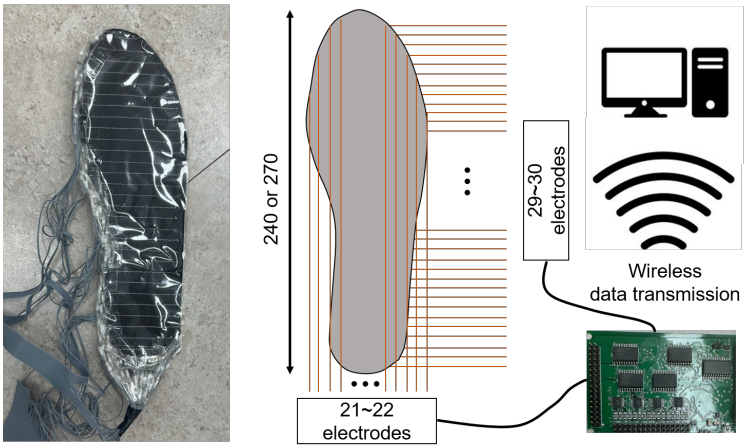


Figure 5: **Tactile sensing hardware.** Our insole-type tactile sensor consisting over 600 pressure sensors, with two insole sizes. Custom readout circuits send tactile data in a wireless manner.

The insole-type tactile sensor used for the data collection is shown in Figure 5. We prepared two insole sizes to accommodate most foot lengths: the smaller insole measures 240 mm with a 30×22 sensor grid (660 sensors), and the larger one is 270 mm with a 29×21 sensor grid (609 sensors). Both models are designed to maintain a consistent sensor density to ensure uniform data quality across different foot sizes. The tactile data from the insole with a 29×21 grid is padded to a 30×22 array to have the same input size for training.

6.3 Detailed action classification result

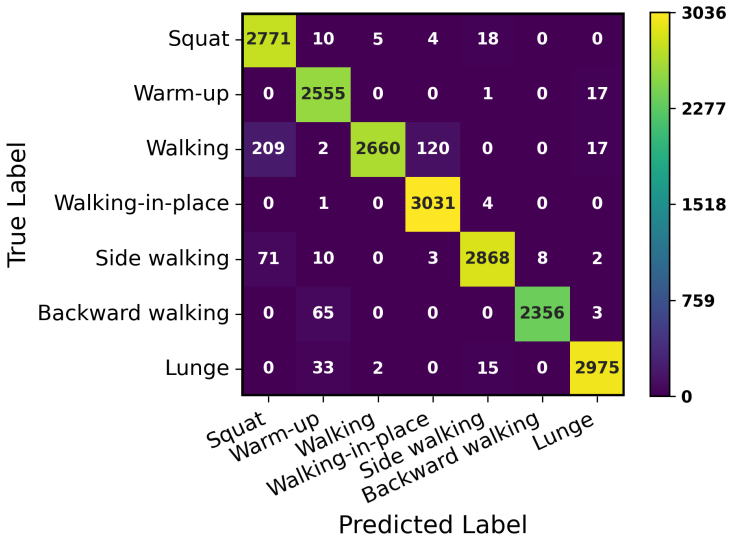


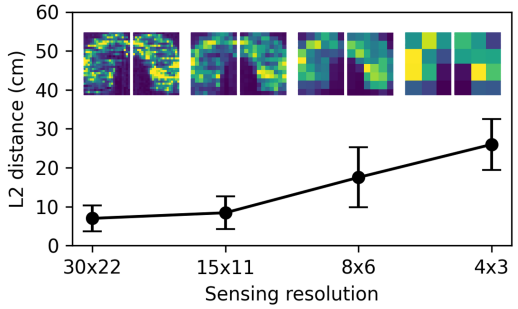
Figure 6: **Confusion matrix of action classification result**. Confusion matrix showing the results of action classification using the linear classifier on the learned representations from the pose estimation model.

The detailed result of evaluating the action classification model on the test data is shown in Figure 6. The model is implemented by downsampling the features extracted from the encoder through one max-pooling layers and feeding them into a linear classifier. The result suggests that our model successfully learned rich representations of tactile frames.

6.4 Ablation study

We conducted ablation studies to demonstrate the crucial roles of the high-resolution tactile data and loss terms that we use to train.

6.4.1 Estimation performance with different sensing resolution



Loss term	Avg L2 dist. (cm)
HM+KP	**7.43**
HM	8.18
KP	7.98

Figure 7: **Ablation study.** L2 distance, representing the difference in keypoints prediction (**left**), increases when input data resolution decreases. The x-axis represents the sensing resolution for a single foot. Total average of L2 distances with using different loss term (**right**). HM refers to the heatmap loss, while KP refers to the keypoint loss. The L2 distance error is minimized when both types of losses are used.

We reduced the resolution of the tactile data without modifying the existing model architecture by applying average pooling with a kernel size of 2 and a stride of 2 to the input data, followed by

interpolation. Depending on the degree of average pooling applied, we reduced the resolution from 30×22 to 15×11, 8×6, and 4×3, and trained the model with these resolutions. As shown in Figure 7, it can be observed that the Euclidean distance of the predicted keypoints increases as the resolution is reduced. This indicates that high-resolution data is a critical factor for the model performance.

6.4.2 Model performance with different loss function

As mentioned in Section 2.2, we utilized a weighted sum of the heatmap loss and keypoint loss. To demonstrate that both losses are critical components of the training process, we conducted an experiment that compares Euclidean distances of models by loss terms to train. HM and KP refer to the heatmap loss and the keypoint loss, respectively. The table in Figure 7 shows the average Euclidean distance in the test set and indicates that both losses play a crucial role in reducing the test error.

6.5 Generalization on unseen actions and individuals

Figure 8 shows the performance of the model on the two unseen actions (backward walking, lunging) and unseen individuals. In the case of seen actions and individuals, the models are trained on a dataset that contains all actions and individuals. In the case of unseen actions and individuals, the model was trained on a dataset where a specific task or individual is excluded. In the case of backward walking, since the training dataset contains movements with relatively similar pressure distributions, such as walking and side walking, the Euclidean distance shows an insignificant increase, demonstrating the strong generalization ability of the model. For lunging, however, the dataset lacks movements with similar pressure distributions, increasing Euclidean distance. This is particularly noticeable for the upper body keypoints, which are less correlated to the foot pressure data. Our model also successfully generalizes to unseen individuals.

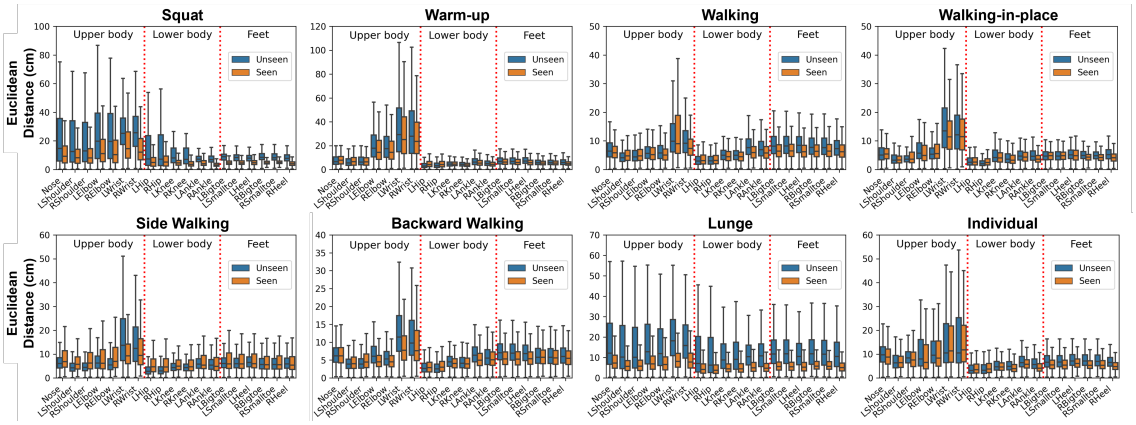


Figure 8: **Generalization results.** Results of evaluating a model trained with the full dataset on three generalization tasks (**top**). Results of evaluating a model trained with the dataset excluding the generalization tasks (**bottom**).

References

T. T. Alemayoh, J. H. Lee, and S. Okamoto. A neural network-based lower extremity joint angle estimation from insole data. In *2023 20th International Conference on Ubiquitous Robots (UR)*, pages 787–791. IEEE, 2023.

L. Casas, N. Navab, and S. Demirci. Patient 3d body pose estimation from pressure imaging. *International journal of computer assisted radiology and surgery*, 14:517–524, 2019.

H. M. Clever, A. Kapusta, D. Park, Z. Erickson, Y. Chitalia, and C. C. Kemp. 3d human pose estimation on a configurable bed from a pressure image. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 54–61. IEEE, 2018.

H. M. Clever, Z. Erickson, A. Kapusta, G. Turk, K. Liu, and C. C. Kemp. Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6215–6224, 2020.

A. Jalal, S. Kamal, and D. Kim. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. *Sensors*, 14(7):11735–11759, 2014.

Y. Ji, Y. Yang, F. Shen, H. T. Shen, and X. Li. A survey of human action analysis in hri applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):2114–2128, 2019.

K. Kasman and V. G. Moshnyaga. New technique for posture identification in smart prayer mat. *Electronics*, 6(3):61, 2017.

S.-R. Ke, L. Zhu, J.-N. Hwang, H.-I. Pai, K.-M. Lan, and C.-P. Liao. Real-time 3d human pose estimation from monocular view with applications to event detection and video gaming. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 489–496. IEEE, 2010.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv e-prints*, pages arXiv–1412, 2014.

C. Liu, Z. Dong, L. Huang, W. Yan, X. Wang, D. Fang, and X. Chen. Tagsleep3d: Rf-based 3d sleep posture skeleton recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–28, 2024.

S. Liu, X. Huang, N. Fu, C. Li, Z. Su, and S. Ostadabbas. Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1106–1118, 2022.

Y. Luo, Y. Li, M. Foshey, W. Shou, P. Sharma, T. Palacios, A. Torralba, and W. Matusik. Intelligent carpet: Inferring 3d human pose from tactile signals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11255–11265, 2021.

P. Maurice, A. Malaisé, C. Amiot, N. Paris, G.-J. Richard, O. Rochel, and S. Ivaldi. Human movement and ergonomics: An industry-oriented dataset for collaborative robotics. *The International Journal of Robotics Research*, 38(14):1529–1537, 2019.

D. McColl, Z. Zhang, and G. Nejat. Human body pose interpretation and classification for social human-robot interaction. *International Journal of Social Robotics*, 3:313–332, 2011.

K. Meng, S. Zhao, Y. Zhou, Y. Wu, S. Zhang, Q. He, X. Wang, Z. Zhou, W. Fan, X. Tan, et al. A wireless textile-based sensor system for self-powered personalized health care. *Matter*, 2(4): 896–907, 2020.

Š. Obdržálek, G. Kurillo, J. Han, T. Abresch, and R. Bajcsy. Real-time human pose detection and tracking for tele-rehabilitation in virtual reality. In *Medicine Meets Virtual Reality 19*, pages 320–324. IOS Press, 2012.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Q. Shi, Z. Zhang, T. He, Z. Sun, B. Wang, Y. Feng, X. Shan, B. Salam, and C. Lee. Deep learning enabled smart mats as a scalable floor monitoring system. *Nature communications*, 11(1):4609, 2020.

W.-k. Tam, A. Wang, B. Wang, and Z. Yang. Lower-body posture estimation with a wireless smart insole. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3348–3351. IEEE, 2019.

S. Tripathi, L. Müller, C.-H. P. Huang, O. Taheri, M. J. Black, and D. Tzionas. 3d human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4713–4725, 2023.

XRMoCap. Openxrlab multi-view motion capture toolbox and benchmark. https://github.com/openxrlab/xrmocap, 2022.

Z. Zhou, S. Padgett, Z. Cai, G. Conta, Y. Wu, Q. He, S. Zhang, C. Sun, J. Liu, E. Fan, et al. Single-layered ultra-soft washable smart textiles for all-around ballistocardiograph, respiration, and posture monitoring during sleep. *Biosensors and Bioelectronics*, 155:112064, 2020.