Intelligent Carpet: Inferring 3D Human Pose from Tactile Signals

Yiyue Luo, Yunzhu Li, Michael Foshey, Wan Shou, Pratyusha Sharma,
Tomás Palacios, Antonio Torralba, Wojciech Matusik
Massachusetts Institute of Technology
{yiyueluo, liyunzhu, mfoshey, wanshou, pratyuss, tpalacios, torralba, wojciech}@mit.edu

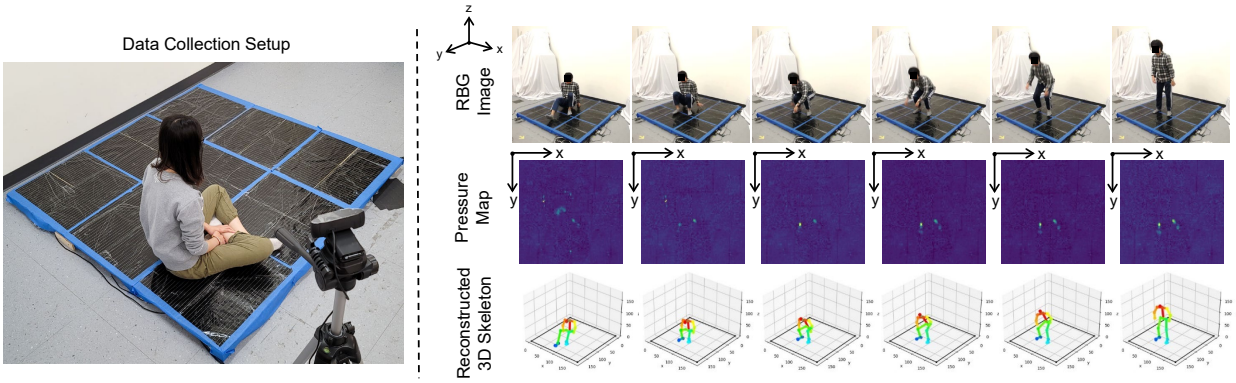


Figure 1. **Left:** A low-cost, high-density, large-scale intelligent carpet system to capture the real-time human-floor tactile interactions. **Right:** The inferred 3D human pose from the captured tactile interactions of a person standing up from a sitting position.

Abstract

Daily human activities, e.g., locomotion, exercises, and resting, are heavily guided by the tactile interactions between the human and the ground. In this work, leveraging such tactile interactions, we propose a 3D human pose estimation approach using the pressure maps recorded by a tactile carpet as input. We build a low-cost, high-density, large-scale intelligent carpet, which enables the real-time recordings of human-floor tactile interactions in a seamless manner. We collect a synchronized tactile and visual dataset on various human activities. Employing a state-of-the-art camera-based pose estimation model as supervision, we design and implement a deep neural network model to infer 3D human poses using only the tactile information. Our pipeline can be further scaled up to multi-person pose estimation. We evaluate our system and demonstrate its potential applications in diverse fields.

1. Introduction

Human pose estimation is critical in action recognition [30, 52], gaming [26], healthcare [64, 36, 22], and robotics [34]. Significant advances have been made to estimate human pose by extracting skeletal kinematics from images and videos. However, camera-based pose estimation remains challenging when occlusion happens, which is inevitable during daily activities. Further, the rising demand for privacy also promotes development in non-vision-based human pose estimation systems [63, 62]. Since most human activities are dependent on the contact between the human and the environment, we present a pose estimation approach using tactile interactions between humans and the ground. Recently, various smart floor or carpet systems have been proposed for human movement detection [11, 2, 48, 7, 3, 60, 40, 16, 1], and posture recognition [25, 50]. Previous work has also demonstrated the feasibility of using pressure images for pose estimation [6, 9, 8]. However, these studies mainly focus on the estimation of poses where a large portion of the body is in direct contact with the sensing surface, e.g., resting postures. A more challenging task is to infer 3D human pose from the limited pressure imprints involved in complicated daily activities, e.g., using feet pressure distribution to reconstruct the pose of the head and limbs. So far, complex 3D human pose estimation and modeling using tactile information, spanning a diverse set of human activities including locomotion, resting, and daily exercises, have not been available.

In this study, we first develop an intelligent carpet – a large integrated tactile sensing array consisting of over 9,000 pressure sensors, covering over 36 square feet, which can be seamlessly embedded on the floor. Coupled with readout circuits, our system enables real-time recordings

of high-resolution human-ground tactile interactions. With this hardware, we collect over 1,800,000 synchronized tactile and visual frames for 10 different individuals performing a diverse set of daily activities, e.g., lying, walking, and exercising. Employing the visual information as supervision, we design and implement a deep neural network to infer the corresponding 3D human pose using only the tactile information. Our network predicts the 3D human pose with the average localization error of less than 10 cm compared with the ground truth pose obtained from the visual information. The learned representations from the pose estimation model, when combined with a simple linear classifier, allow us to perform action classification with an accuracy of 98.7%. We also include ablation studies and evaluate how well our model generalizes to unseen individuals and unseen actions. Moreover, our approach can be scaled up for multi-person 3D pose estimation. Leveraging the tactile sensing modality, we believe our work opens up opportunities for human pose estimation that is unaffected by visual obstructions in a seamless and confidential manner.

2. Related Work

2.1. Human Pose Estimation

Thanks to the availability of large-scale datasets of annotated 2D human poses and the introduction of deep neural network models, human pose estimation from 2D images or videos has witnessed major advances in recent years [57, 56, 45, 38, 39, 59, 46, 5, 17, 12, 20, 54]. Recovering 3D information from 2D inputs is intrinsically ambiguous. Some recent attempts to recover 3D human pose from 2D images either require explicit 3D supervision [33] or rely on a discriminator and adversarial training to learn a valid pose distribution [23, 24] or performing semi-supervised learning by leveraging the temporal information [44]. Still, 3D pose estimation remains a challenging problem due to the underlying ambiguity. Many methods do not perform well in the presence of a severe occlusion. Another alternative is to use multi-camera motion capture systems (e.g., VICON) or multiple cameras to obtain a more reliable pose estimation [51, 63, 21].

Our work builds upon the past advances in computer vision by using OpenPose [4] to extract the 2D keypoints from multiple cameras and triangulate them to generate the ground truth 3D pose. Our system predicts 3D pose from only the tactile signals, which does not require any visual data and is fundamentally different from all past work in computer vision. The introduced tactile carpet has a lower spatial resolution than typical cameras. However, it functions as a camera viewing humans from the bottom up. This type of data stream does not suffer from occlusion problems that are typical for camera systems. Furthermore, it provides additional information, such as whether humans are in contact with the ground and the pressure they exert.

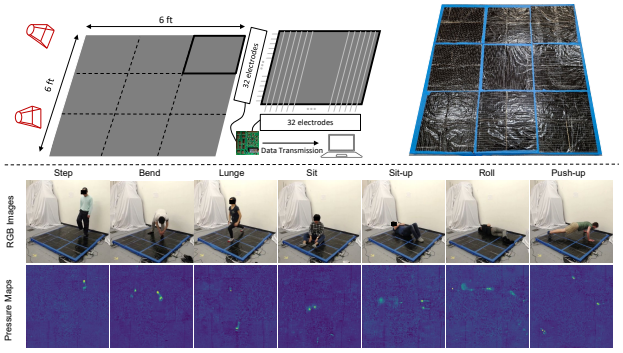


Figure 2. **Tactile data acquisition hardware. Top:** Our recording set-up includes a tactile sensing carpet spanning 36 ft^2 with 9,216 sensors (upper right), the corresponding readout circuits, and 2 cameras. **Bottom:** Typical pressure maps captured by the carpet from diverse human poses and activities.

2.2. Tactile Sensing

Recent advances in tactile sensing have benefited the recording, monitoring, and modeling of human-environment interactions in a variety of contexts. Sundaram et al. [55] have investigated the signatures of human grasp through the tactile interaction between human hands and different objects, while other researchers have proposed to connect vision and touch using cross-domain modeling [31, 61, 28]. Davoodnia et al. [10] transform annotated in-bed human pressure map [41, 15] to images containing shapes and structures of body parts for in-bed pose estimation. Furthermore, extensive works on biomechanics, human kinematics, and dynamic motions [13, 29, 35, 32] have explored the use of the foot pressure maps induced by daily human movement.Previous studies have demonstrated human localization and tracking by embedding individual pressure sensing units in the smart floor systems [53, 50]. Furthermore, using large-scale pressure sensing matrices, researchers have been able to capture foot pressure patterns when humans are standing and walking and develop models that provide gait analysis and human identification [43, 58, 37]. Based on the fact that human maintains balance through redirecting the center of mass by exerting force on the floor [19], Scott et al. [49] have predicted foot pressure heatmap from 2D human kinematics.

Different from previous works, which include only limited actions due to the limited size and resolution of the tactile sensing platform, we record and leverage high-resolution tactile data from diverse human daily activities, e.g., exercises, to estimate 3D human skeleton.

3. Dataset

In this section, we describe details of our hardware setup for tactile data acquisition, pipeline for ground truth 3D keypoint confidence map generation, as well as data augmentation and synthesis for multi-person pose estimation.

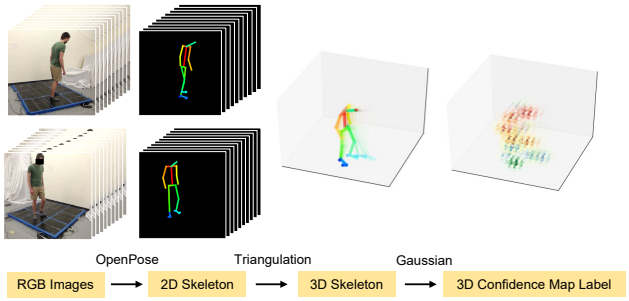


Figure 3. **3D keypoint confidence map generation.** The ground truth voxelized 3D keypoint confidence maps are annotated by first extracting 2D skeleton keypoints from RGB images using OpenPose [4], then generating 3D keypoints through triangulation and optimization, and finally applying a 3D Gaussian filter.

3.1. Tactile Data Acquisition

Our tactile data acquisition is based on a custom, high-density, large-scale piezoresistive pressure sensing carpet, which spans over 36 ft^2 and contains 9,216 sensors with a spacing of 0.375 inches. The carpet is composed of a piezoresistive pressure sensing matrix fabricated by aligning a network of orthogonal conductive threads as electrodes on each side of the commercial piezoresistive films. Each sensor locates at the overlap of orthogonal electrodes and is able to measure pressure up to 14 kPa with the highest sensitivity of 0.3 kPa. Our tactile sensing carpet is low-cost (∼$100), easy to fabricate, and robust for large-scale data collection. Using the coupled readout circuit, we collect the tactile frames with 9,216 individual sensing readouts at a rate of 14 Hz.

With such a large-scale high-resolution tactile sensing platform, we can not only capture people's foot pressure maps, which most of the previous work focused on, but also capture the full tactile interactions between the human and the floor when people are performing complex activities. As shown in Figure 2, our carpet captures the feet pressure maps when people perform activities in upright positions, as well as the physical contacts between the human body (e.g., hands, limbs) and the floor when people perform exercises and complex actions (e.g., push-ups, sit-ups, and rolling).

We have collected over 1,800,000 synchronized tactile and visual frames for 10 volunteers performing 15 actions. More details are included in supplementary materials. Our tactile acquisition set-up and the captured dataset are open-sourced to facilitate future research in the field.

3.2. 3D Pose Label Generation

We design and implement a pipeline to capture and generate the training pairs, i.e., synchronized tactile frames and 3D keypoint confidence maps. We capture visual data with 2 cameras that are synchronized and calibrated with respect to the global coordinate of the tactile sensing carpet using standard stereo camera calibration techniques. In order to annotate the ground truth human pose in a scalable manner, we leverage a state-of-the-art vision-based system, OpenPose [5], to generate 2D skeletons from the images captured by the cameras.

Once we have obtained the 2D skeletons generated from the calibrated camera system, we triangulate the keypoints to generate the corresponding 3D skeletons. The triangulation results may not be perfect in some frames due to perception noise or misdetection. To resolve this issue, we add a post-optimization stage to constrain the length of each link. More specifically, we first calculate the length of the links in the skeleton using the median value across the naively triangulated result for each person. For this specific person, we denote the length of the i^{th} link as K_i. We then use $\boldsymbol{q}^{\text{A}}$ and $\boldsymbol{q}^{\text{B}}$ to represent the detected N keypoints at a specific time step from the two cameras, which lie in a 2D space, where $\boldsymbol{q}^{\text{A}} = \{\boldsymbol{q}_1^{\text{A}}, \dots, \boldsymbol{q}_N^{\text{A}}\}$ and $\boldsymbol{q}_k^{\text{A}} = (u_k^{\text{A}}, v_k^{\text{A}})$. We then calculate the length of each link $\hat{K}_i$ from the naive triangulation result and optimize the 3D location of the keypoints $\boldsymbol{p} = \{\boldsymbol{p}_1, \dots, \boldsymbol{p}_N\}$ by minimizing the following loss function using stochastic gradient descent:

$$\begin{aligned}\mathcal{L}^{\text{skeleton}} = &\sum_{k=1}^{N} \|P^{\text{A}}\boldsymbol{p}_k - \boldsymbol{q}_k^{\text{A}}\| + \sum_{k=1}^{N} \|P^{\text{B}}\boldsymbol{p}_k - \boldsymbol{q}_k^{\text{B}}\| \\ &+ \sum_{i=1}^{N-1} \|\hat{K}_i - K_i\|\end{aligned} \tag{1}$$

where there are N keypoints and $N-1$ links, $\boldsymbol{p} = \{\boldsymbol{p}_1, \dots, \boldsymbol{p}_N\}$ lie in 3D space spanned by the world coordinate, $\boldsymbol{p}_k = (x_k, y_k, z_k)$. P^{A} and P^{B} are the camera matrices that project the 3D keypoints onto the 2D image frame. In our experiments, we use $N = 21$. The accuracy of the 3D pose label and the effectiveness of our optimization pipeline are further analyzed in supplementary materials.

Given the optimized 3D positions of the 21 keypoints on the human skeleton, we further generate the 3D keypoint confidence maps by applying a 3D Gaussian filter over the keypoint locations on a voxelized 3D space (Figure 3).

3.3. Data Augmentation and Multi-person Dataset Synthesis

When projecting the human skeletons to the x-y plane (Figure 1), we find a spatial correspondence between the projection and the tactile signals, which allows us to augment our dataset by rotating and shifting the tactile frames and the corresponding human skeletons.

Due to the restriction of social distancing and the size of the sensing carpet, we conduct the data collection with only one person at a time. The multi-person dataset, however, is synthesized by combining multiple single-person clips. In other words, we add up the synchronized tactile frames and the generated 3D keypoint confidence maps from different

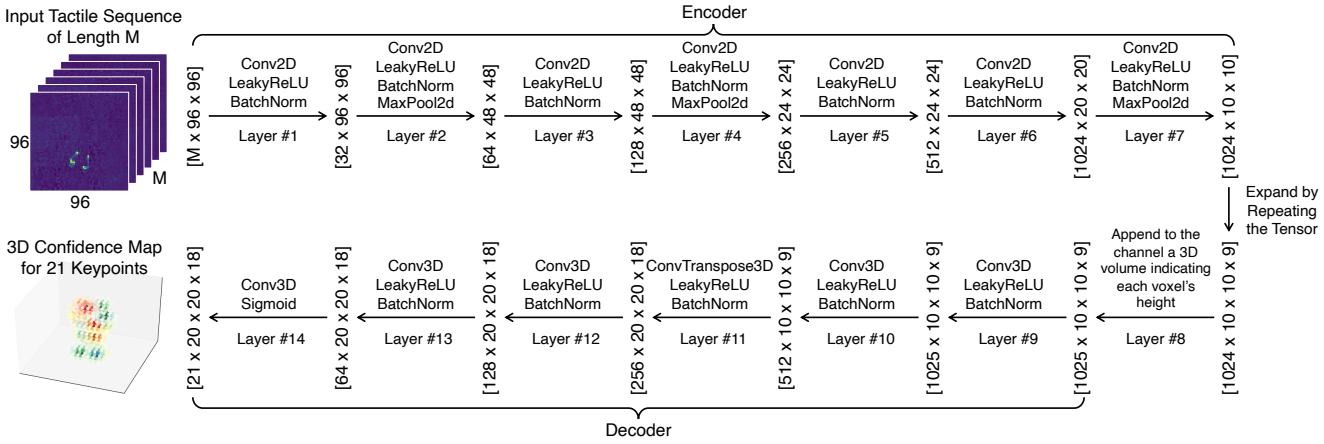


Figure 4. **Overview of the model for 3D human pose estimation.** Our model consists of an encoder and a decoder. The encoder maps the input tactile sequence into a 10×10 feature through 7 blocks of Conv2D-LeakyReLU-BatchNorm. We then expand the feature and repeat along the last dimension to transform the 2D feature map into a 3D feature volume. After appending an indexing volume indicating the height of each voxel, the feature goes through a set of decoding layers to generate the predicted confidence map for each keypoint.

recording clips. More specifically, since people rarely perform actions with one on top of the other, we assume that the pressure maps induced by the actions of different people will not overlap at the given time. We specify the location of each person by creating anchor boxes of the human skeleton projected onto the floor plane. We remove frames with the Intersection over Union (IoU) larger than 0.1 to ensure that the skeletons and tactile signals from different people do not overlap with each other. Note that the training of our models in the experiments is entirely based on the single-person dataset and the synthetic multi-person variants. We also record synchronized visual and tactile data for multiple people but only for evaluation purposes.

4. Method

In this section, we present the details of our pose estimation model. We first discuss how we transform the tactile frames into 3D volumes indicating the confidence map of the keypoints. We then describe how we extend it to multi-person scenarios and present the implementation details.

4.1. Keypoint Detection using Tactile Signals

The goal of our model is to take the tactile frames as input and predict the corresponding 3D human pose. We take the ground truth human pose estimated from our multi-camera setup as the supervision and train our model to predict the 3D confidence map of each of the 21 keypoints, including head, neck, shoulders, elbows, waists, hips, knees, ankles, heels, small and big toes.

To include more contextual information and reduce the effects caused by the sensing noise, instead of taking a single tactile frame as input, our model takes a sequence of tactile frames spanning a temporary window of length M as input (Figure 4). For each input segment, the model processes the spatio-temporal tactile information and outputs the keypoint confidence maps in 3D that correspond to the middle frame.

As shown in Figure 1, the input tactile frames lie in 2D space, which has a nice spatial correspondence with the human skeleton over the x-y plane (the floor plane). Our model builds on top of a fully convolutional neural network to exploit such spatial equivariance. The encoder of our model uses 2D convolution to process the tactile frames. Then, to regress the keypoints in 3D, we expand the feature map by repeating it along a new dimension in the middle of our network (Figure 4), which essentially transforms the 2D feature map into a 3D feature volume. However, naive 2D to 3D expanding via repetition will introduce ambiguities as subsequent convolutional layers use shared kernels to process the feature - it is impossible to tell the height of a specific voxel, making it hard to regress the keypoint location along the z-axis. To resolve this issue, we add a new channel to the 3D feature map with a 3-dimensional indexing volume, indicating the height of each voxel. We then use 3D convolution to process the feature and predict the 3D keypoint confidence map for each of the 21 keypoints. The detailed architecture and the size of the feature maps along the forwarding pass are shown in Figure 4.

We optimize the model by minimizing the Mean Squared Error (MSE) between the predicted keypoint heatmap and the ground truth using Adam optimizer [27]. We also use spatial softmax to transform the heatmap into the keypoint location and include an additional loss term $\mathcal{L}^{\text{link}}$ to constrain the length of each link in the skeleton to lie in the range of normal human limb length. For each data point, the loss function is defined as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\|H_i - \hat{H}_i\| + \frac{1}{N-1}\sum_{i=1}^{N-1}\mathcal{L}_i^{\text{link}}, \qquad (2)$$

where N denotes the number of keypoints, $N-1$ is the number of links in the skeleton, H_i and $\hat{H}_i$ represent the ground truth and the predicted 3D keypoint confidence

maps. The link loss is defined as the following:

$$\mathcal{L}_i^{\text{link}} = \begin{cases} K_i^{\min} - \hat{K}_i, & \text{if } \hat{K}_i < K_i^{\min}. \\ \hat{K}_i - K_i^{\max}, & \text{if } \hat{K}_i > K_i^{\max}. \\ 0, & \text{otherwise}, \end{cases} \quad (3)$$

where $\hat{K}_i$ is the link length calculated from our prediction, $K_i^{\min}$ and $K_i^{\max}$ represent the 3$^{\text{th}}$ and 97$^{\text{th}}$ percentile of each of the body limb length in the training dataset.

4.2. Keypoint Detection for Multiple People

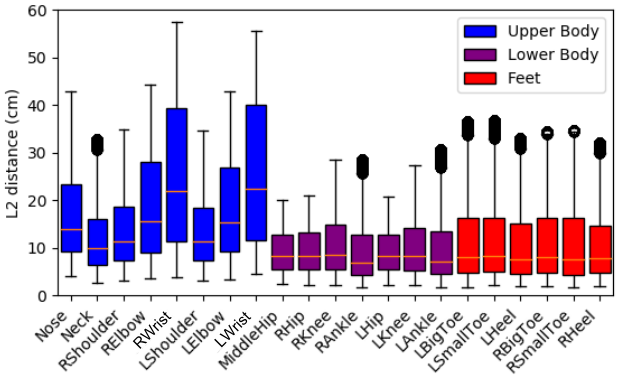
When moving into multi-person scenarios, each keypoint confidence map can contain multiple regions with high confidence that belong to different people. Therefore, we threshold the keypoint confidence map to segment out each of these high confidence regions, and then calculate the centroid of each region to transform it into the 3D keypoint location. To associate the keypoints that belong to the same person, we start from the keypoint of the head and traverse through the person's skeleton (represented as a tree) to include the remaining keypoints. Every time we want to add a new keypoint to the person, e.g., the neck, we select the one among multiple extracted keypoint candidates with the closest L2 distance to its parent, e.g., head, which has already been added to the person on the skeleton tree. This method is simple but works well when people are kept at a certain distance from each other. More complicated and effective techniques could be used to handle cases where people are very close to each other [5]. Since it is not the main focus of this paper, we plan to explore this direction in the future.

4.3. Implementation Details

Our network is implemented using PyTorch [42]. We train the model by minimizing Eq. 2 using a learning rate of $1e^{-4}$ and a batch size of 32. As shown in Figure 4, the encoding part of our network consists of 7 groups of layers. The Conv2D in the first 5 and the 7$^{\text{th}}$ layers use 3×3 kernels and 1×1 padding. The 6$^{\text{th}}$ uses 5×5 kernels and zero padding. A 2×2 MaxPool2D is also applied in the 2$^{\text{nd}}$, 4$^{\text{th}}$, and 7$^{\text{th}}$ layers to reduce the resolution of the feature maps.

We expand the tactile feature maps to 3D by repeating the tensor 9 times along the last dimension, and then append the channel with a 3D indexing volume indicating the height of each voxel. The decoding network takes in the resulting tensor and predicts the 3D confidence maps of the keypoints.

The decoder is composed of 5 layers of 3×3 ×3 3D convolution with a padding of 1×1 ×1. The 11$^{\text{th}}$ layer uses a kernel size of 2×2×2 with a stride of 2 to increase the resolution. We also apply batch normalization and Leaky ReLU after each layer except the last one, where we use the Sigmoid activation function to regress the confidence value.



Axis	Ave.	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Feet
X	6.8	6.4	6.3	8.9	10.9	4.6	5.8	5.6	6.4
Y	7.2	8.0	6.5	8.8	10.9	5.2	5.8	5.7	6.7
Z	6.8	9.6	7.0	8.9	14.4	4.0	4.0	3.1	3.5

Figure 5. **Results on single person pose estimation (unit: cm). Top:** The Euclidean distance between the predicted single-person 3D skeleton (21 keypoints) and the ground truth label. **Bottom:** Average keypoint localization error of body parts along the X, Y, and Z axis in the real-world coordinate. Since the changes in pressure maps are dominated by the movements of the lower body and the torso, their predictions are more accurate.

5. Experiments

5.1. Single Person Pose Estimation

Single-person pose estimation is trained with 135,000 pairs of tactile and visual frames and validated on 30,000 pairs of frames. The performance is tested on a held-out test set with 30,000 tactile frames. We use Euclidean distance (L2) as the evaluation metric to compare the predicted 3d human pose to the corresponding ground truth human pose retrieved from the visual data. The L2 distance of each keypoint and the localization error of each body part are listed in Figure 5. Generally, keypoints on the lower body (e.g., knee and ankle) and the torso (e.g., shoulder and hip) hold higher accuracy compared with the keypoints on the upper body (e.g., waist and head). The observation agrees with our intuition that changes in pressure maps are primarily determined by the positions of the lower body and the torso. We also note that the model obtains better predictions if the keypoints are closer to the torso on the skeleton tree - the prediction error increases as we move further away from the torso, e.g., shoulders to elbows, and then to the waist. Figure 6 shows some qualitative results on the 3D human pose predictions, along with the input tactile frames, ground truth pose extracted from the RGB image, over a continuous time sequence.

We perform ablation studies on the sensing resolution of the intelligent carpet. To ablate the tactile sensing resolution, we reassign the value in each 2 × 2 grid with the average of the four values, which reduces the effective resolution from 96 × 96 to 48 × 48. We then use the same train-

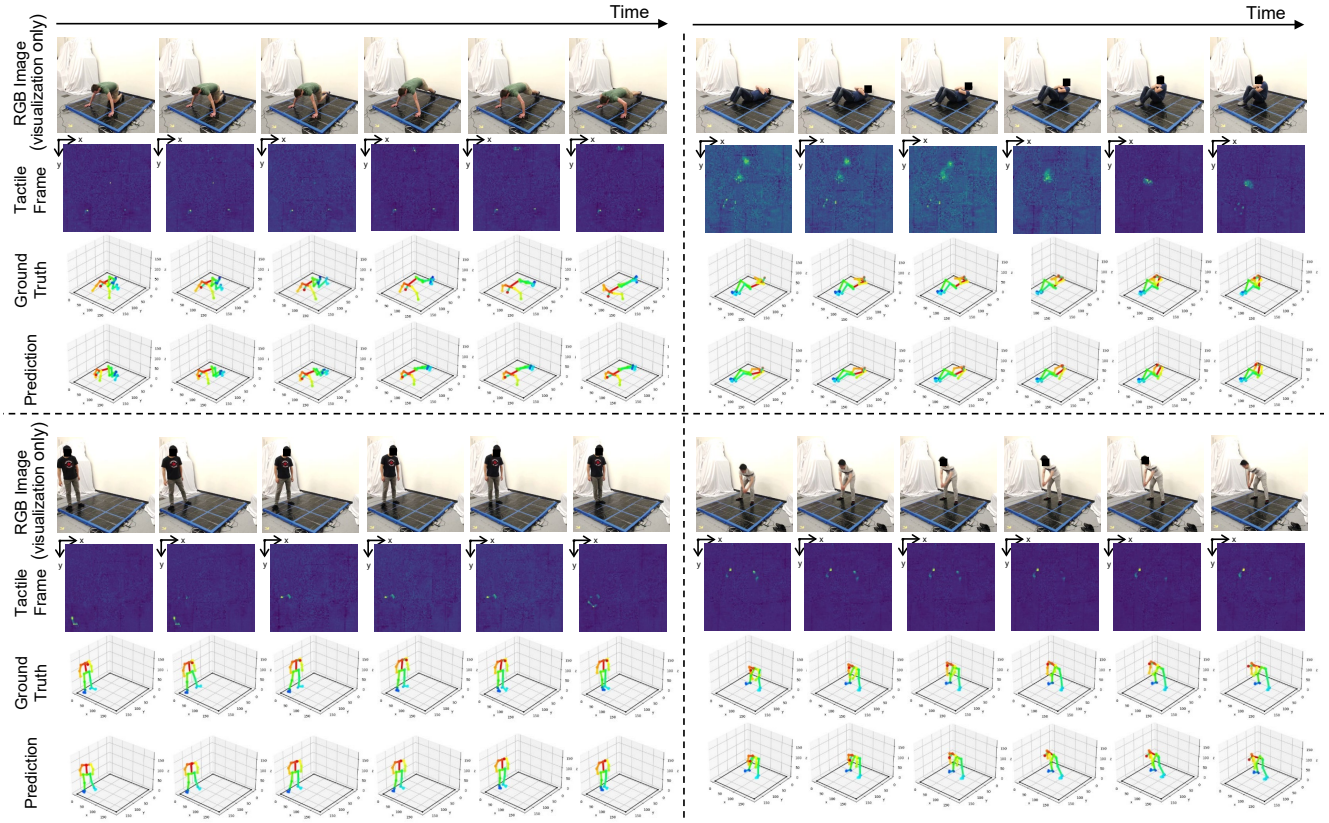


Figure 6. **The qualitative results of single-person 3D pose estimation across time steps.** For each sequence, from top to bottom, we show the RGB image as ground truth annotation (only used here for visualization purpose), the captured tactile frame, ground truth 3D skeleton, and predicted 3D skeleton from our model using only the tactile frames (unit: cm). The predicted poses are consistent over time with a smooth transition along the corresponding trajectories.

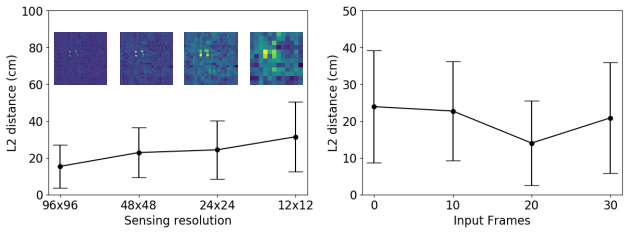


Figure 7. **Ablation studies.** Model performance with different sensing resolution (left) and the number of input frames (right).

ing pipeline to derive the predictions. A similar procedure is employed for evaluating the model performance with the effective sensing resolution of 24 $\times$ 24 and 12 $\times$ 12. As Figure 7 illustrates, the prediction accuracy decreases with the decrease of sensing resolution, which reiterates the importance of our high density, large-scale tactile sensing platform. We also perform an ablation study on the number of input frames, where the best performance was obtained with 20 input frames ($\sim$1.5 sec, Figure 7). We include additional ablation studies on our pose estimation model, i.e. the 3D indexing volume, repeating tensor, and link length loss, in supplementary materials.

We evaluate how well the model generalizes to unseen individuals and activities. As demonstrated in Figure 8, our model generalizes to unseen people with a negligible increase of the keypoint localization error. On the other hand, our model has a varying performance on different types of unseen tasks. The learned model easily generalizes to poses with the pressure maps similar to what the model is trained on but delivers poor performance with tactile imprints that the model has never encountered before. For example, our model generalizes to the lunging pose, where the pressure maps are mainly directed by the human's center of mass; however, it fails to predict the push-up pose, which induces pressure imprints that are vastly different from the training distribution. When deploying the system for practical use in real life, it is essential to perform a more systematic data collection procedure covering more typical human activities to achieve a more reliable pose estimation performance.

5.2. Action Classification

To obtain a deeper understanding of the learned features in the pose estimation network, we perform action classification by applying a linear classifier on the downsampled tactile feature maps. We use the dataset on one single person performing 10 different actions, where 80% is used for training, 10% for validation, and 10% for testing. As demonstrated in Figure 9, we obtain an accuracy of 97.8%,

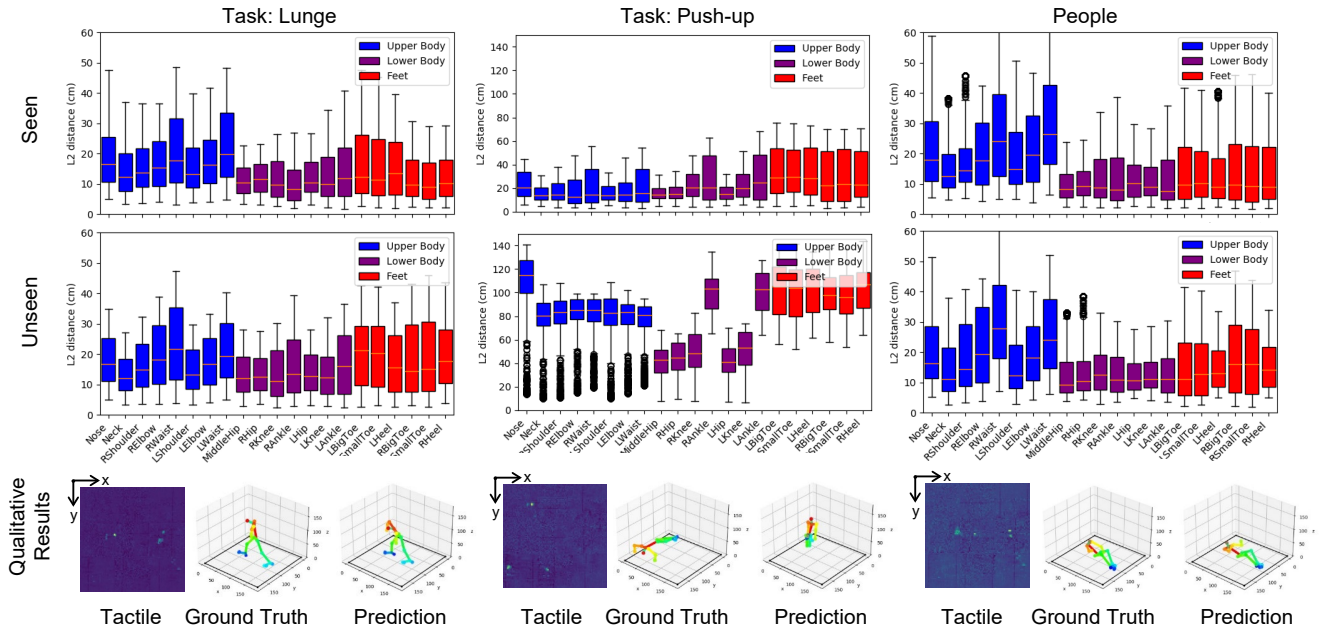


Figure 8. **Generalization results (unit: cm). Top:** Localization error of predictions on seen tasks and individuals, where the training was performed on the full dataset including all tasks and individuals. **Middle:** Localization error of predictions on unseen tasks and individuals, where the training was performed on a split dataset excluding specific actions and individuals. **Bottom:** Qualitative results on unseen tasks and individuals.

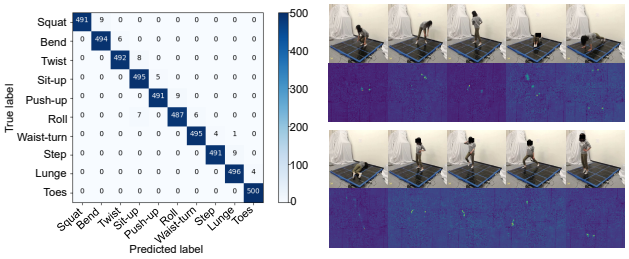
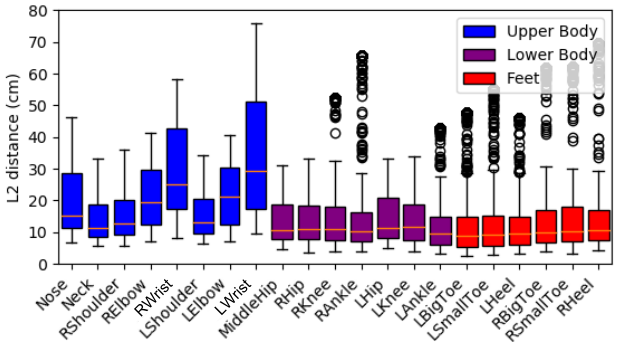


Figure 9. **Results on action classification. Left:** Confusion matrix of action classification using a linear classifier on the learned features from the pose estimation model. The linear model achieves good accuracy, suggesting that the learned features contain semantically meaningful information on the input tactile frames. **Right:** Representative tactile frames from different actions.

which demonstrates the capability of our model to facilitate downstream classification tasks.

5.3. Multi-person Pose Estimation

We further extend our model for multi-person pose estimation. As discussed in Section 4.2, the multi-person pose estimation model is trained and validated with 112,000 and 10,000 pairs of synthesized tactile frames and keypoint confidence maps. Performance is evaluated with 4,000 recorded tactile frames of two people performing stepping, sitting, lunging, twisting, bending, squatting, and standing on toes. The L2 distance of each keypoint and the localization error of each body part are listed in Figure 10. Examples of the multi-person pose prediction are shown in Figure 11. Purely from the tactile information, our network successfully localizes each individual and predicts his or her 3D pose with a localization error of less than 15 cm. The predictions do not rely on any visual information and, therefore, are unaffected by visual obstructions or a limited field of view, which are common challenges in vision-based human pose estimation.



Axis	Ave.	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Feet
X	14.5	14.1	10.1	15.3	24.7	10.2	12.6	14.1	14.9
Y	12.9	13.9	10.8	15.9	21.6	10.1	11.0	9.7	9.9
Z	12.7	16.6	13.2	17.3	23.9	10.0	8.0	6.5	6.4

Figure 10. **Results on multi-person scenarios (unit: cm). Top:** Euclidean distance between the predicted multi-person 3D skeleton and the ground truth. **Bottom:** Average keypoint localization error of body parts along the X, Y, and Z axis in the real-world coordinate.

5.4. Failure Cases

As demonstrated in Figure 12, the typical failure cases can be categorized into three main types. First, the model

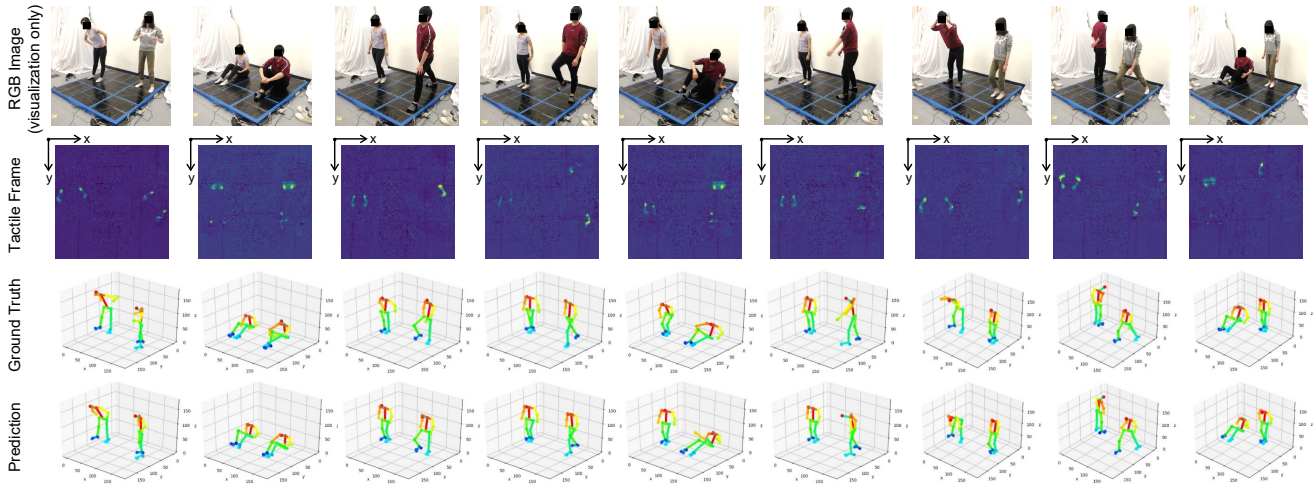


Figure 11. **Qualitative results of multi-person 3D human pose estimation.** From top to bottom, we show the RGB image for ground truth annotation, the captured tactile frame, ground truth 3D skeleton, and the predicted 3D skeleton from our model using only the tactile frames (unit: cm). Our network learns to localize each individual and predict the corresponding 3D pose.

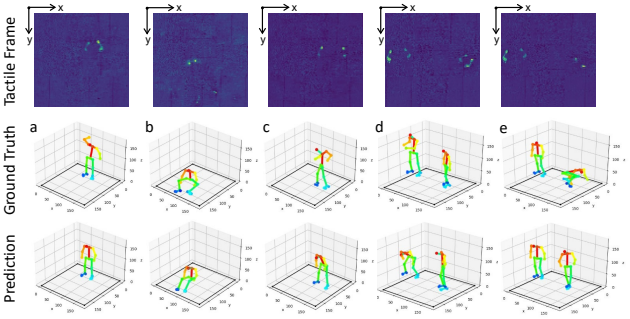


Figure 12. **Failure cases.** Our model fails due to the lack of discernible physical contact with the floor (a-b) or the ambiguity of the tactile signal (c-e).

fails to predict the position of the waist and the head (Figure 12 a). This is expected as we observe that the pressure distributions of the tactile maps are rarely or not affected by the movement of the head and wrist when a person is standing on feet. Also, the model fails to predict the poses where actions are performed without notable physical contact with the floor, e.g., free-floating legs during sit-ups and twisted torso during the standing-up process (Figure 12 b and e). Furthermore, different actions may induce very similar pressure imprints, e.g., bending and twisting, causing trouble for the model to distinguish the activities due to the intrinsic ambiguity of the tactile signal (Figure 12 c and d). As for the multi-person pose estimation, additional errors can happen because of the ambiguity underlying the tactile signals from different individuals, where the model fails when two people are too close to each other. This type of data is not included in our synthetic training dataset.

6. Limitations and Future Work

We observe that even with the constraint on the human body link lengths, some predicted human poses appear unrealistic in real life. Therefore, adversarial prior can be imposed to further constrain the predicted 3D human pose. Also, we currently use the same model for the single-person and multi-person pose estimation, which suffers from the ambiguity of the tactile signal induced by multiple people that are too close to each other. To obtain more accurate predictions on multi-person pose estimation, a region proposal network can be applied to localize the tactile information belonging to each of the individuals, which will then respectively pass through the pose estimation network to predict the pose of each person [14, 47, 18]. Estimation and modeling of multi-person interactions from tactile information would be another interesting future direction.

7. Conclusion

We built a low-cost, high-density, large-scale tactile sensing carpet and captured a large tactile and visual dataset on humans performing daily activities. Leveraging the perception results from a vision system as supervision, our model learns to infer single-person and multi-person 3D human skeletons with only the tactile readings of humans performing a diverse set of activities on the intelligent carpet. This work introduces a sensing modality that is different and complementary to the vision system, opening up new opportunities for human pose estimation unaffected by visual obstructions in a seamless and confidential manner, with potential applications in smart homes, healthcare, and gaming.

Acknowledgement: We are grateful to all volunteers for their contributions to our dataset and R. White for the administration of the project. We thank the anonymous reviewers for their insightful comments.

References

[1] Ibrahim Al-Naimi and Chi Biu Wong. Indoor human detection and tracking using advanced smart floor. In *2017 8th International Conference on Information and Communication Systems (ICICS)*, pages 34–39. IEEE, 2017.

[2] Myra A Aud, Carmen C Abbott, Harry W Tyrer, Rohan Vasantha Neelgund, Uday G Shriniwar, Ashrafuddin Mohammed, and Krishna Kishor Devarakonda. Smart carpet: Developing a sensor system to detect falls and summon assistance. *Journal of gerontological nursing*, 36(7):8–12, 2012.

[3] Leticia M Avellar, Arnaldo G Leal-Junior, Camilo AR Diaz, Carlos Marques, and Anselmo Frizera. Pof smart carpet: a multiplexed polymer optical fiber-embedded smart carpet for gait analysis. *Sensors*, 19(15):3356, 2019.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[6] Leslie Casas, Nassir Navab, and Stefanie Demirci. Patient 3d body pose estimation from pressure imaging. *International journal of computer assisted radiology and surgery*, 14(3):517–524, 2019.

[7] Kabalan Chaccour, Rony Darazi, Amir Hajjam el Hassans, and Emmanuel Andres. Smart carpet using differential piezoresistive pressure sensors for elderly fall detection. In *2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 225–229. IEEE, 2015.

[8] Henry M Clever, Zackory Erickson, Ariel Kapusta, Greg Turk, Karen Liu, and Charles 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.

[9] Henry M Clever, Ariel Kapusta, Daehyung Park, Zackory Erickson, Yash Chitalia, and Charles 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.

[10] Vandad Davoodnia, Saeed Ghorbani, and Ali Etemad. In-bed pressure-based pose estimation using image space representation learning. *arXiv preprint arXiv:1908.08919*, 2019.

[11] Scott Elrod and Eric Shrader. Smart floor tiles/carpet for tracking movement in retail, industrial and other environments, Mar. 29 2007. US Patent App. 11/236,681.

[12] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017.

[13] C Giacomozzi, V Macellari, A Leardini, and MG Benedetti. Integrated pressure-force-kinematics measuring system for the characterisation of plantar foot loading during locomotion. *Medical and Biological Engineering and Computing*, 38(2):156–163, 2000.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[15] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[16] Chuan He, Weijun Zhu, Baodong Chen, Liang Xu, Tao Jiang, Chang Bao Han, Guang Qin Gu, Dichen Li, and Zhong Lin Wang. Smart floor with integrated triboelectric nanogenerator as energy harvester and motion sensor. *ACS Applied Materials & Interfaces*, 9(31):26126–26133, 2017.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[19] At L Hof. The equations of motion for a standing human reveal three mechanisms for balance. *Journal of biomechanics*, 40(2):451–457, 2007.

[20] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[21] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7718–7727, 2019.

[22] Ahmad Jalal, Shaharyar Kamal, and Daijin 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.

[23] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[24] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

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

[26] Shian-Ru Ke, LiangJia Zhu, Jenq-Neng Hwang, Hung-I Pai, Kung-Ming Lan, and Chih-Pin 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.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] Jet-Tsyn Lee, Danushka Bollegala, and Shan Luo. "touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4276–4282. IEEE, 2019.

[29] Edward D Lemaire, Ajoy Biswas, and Jonathan Kofman. Plantar pressure parameters for dynamic gait stability analysis. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4465–4468. IEEE, 2006.

[30] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.

[31] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[32] Yiyue Luo, Yunzhu Li, Pratyusha Sharma, Wan Shou, Kui Wu, Michael Foshey, Beichen Li, Tomás Palacios, Antonio Torralba, and Wojciech. Matusik. Learning human–environment interactions using conformal tactile textiles. *Nature Electronics*, 4(3):193–201, 2021.

[33] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[34] Pauline Maurice, Adrien Malaisé, Clélie Amiot, Nicolas Paris, Guy-Junior Richard, Olivier Rochel, and Serena Ivaldi. Human movement and ergonomics: An industry-oriented dataset for collaborative robotics. *The International Journal of Robotics Research*, 38(14):1529–1537, 2019.

[35] Marnee J McKay, Jennifer N Baldwin, Paulo Ferreira, Milena Simic, Natalie Vanicek, Elizabeth Wojciechowski, Anita Mudge, Joshua Burns, 1000 Norms Project Consortium, et al. Spatiotemporal and plantar pressure patterns of 1000 healthy individuals aged 3–101 years. *Gait & posture*, 58:78–87, 2017.

[36] Keyu Meng, Shenlong Zhao, Yihao Zhou, Yufen Wu, Songlin Zhang, Qiang He, Xue Wang, Zhihao Zhou, Wenjing Fan, Xulong Tan, et al. A wireless textile-based sensor system for self-powered personalized health care. *Matter*, 2020.

[37] Lee Middleton, Alex A Buss, Alex Bazin, and Mark S Nixon. A floor sensor system for gait recognition. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, pages 171–176. IEEE, 2005.

[38] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in neural information processing systems*, pages 2277–2287, 2017.

[39] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[40] Robert J Orr and Gregory D Abowd. The smart floor: A mechanism for natural user identification and tracking. In *CHI'00 extended abstracts on Human factors in computing systems*, pages 275–276, 2000.

[41] Sarah Ostadabbas, Maziyar Baran Pouyan, Mehrdad Nourani, and Nasser Kehtarnavaz. In-bed posture classification and limb identification. In *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*, pages 133–136. IEEE, 2014.

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[43] Todd C Pataky, Tingting Mu, Kerstin Bosch, Dieter Rosenbaum, and John Y Goulermas. Gait recognition: highly unique dynamic plantar pressure patterns among 104 individuals. *Journal of The Royal Society Interface*, 9(69):790–800, 2012.

[44] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019.

[45] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015.

[46] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016.

[47] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[48] Domnic Savio and Thomas Ludwig. Smart carpet: A footstep tracking interface. In *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*, volume 2, pages 754–760. IEEE, 2007.

[49] Jesse Scott, Christopher Funk, Bharadwaj Ravichandran, John H Challis, Robert T Collins, and Yanxi Liu. From kinematics to dynamics: Estimating center of pressure and base of support from video frames of human motion. *arXiv preprint arXiv:2001.00657*, 2020.

[50] Qiongfeng Shi, Zixuan Zhang, Tianyiyi He, Zhongda Sun, Bingjie Wang, Yuqin Feng, Xuechuan Shan, Budiman Salam, and Chengkuo Lee. Deep learning enabled smart mats as a scalable floor monitoring system. *Nature Communications*, 11(1):1–11, 2020.

[51] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010.

[52] Amarjot Singh, Devendra Patil, and SN Omkar. Eye in the sky: Real-time drone surveillance system (dss) for violent individuals identification using scatternet hybrid deep learning network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1629–1637, 2018.

[53] Miguel Sousa, Axel Techmer, Axel Steinhage, Christl Lauterbach, and Paul Lukowicz. Human tracking and identification using a sensitive floor and wearable accelerometers. In *2013 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 166–171. IEEE, 2013.

[54] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018.

[55] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698–702, 2019.

[56] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014.

[57] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[58] Ruben Vera-Rodriguez, John SD Mason, Julian Fierrez, and Javier Ortega-Garcia. Comparative analysis and fusion of spatiotemporal information for footstep recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(4):823–834, 2012.

[59] Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3073–3082, 2016.

[60] Aifang Yu, Wei Wang, Zebin Li, Xia Liu, Yang Zhang, and Junyi Zhai. Large-scale smart carpet for self-powered fall detection. *Advanced Materials Technologies*, 5(2):1900978, 2020.

[61] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[62] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018.

[63] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. Rf-based 3d skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 267–281, 2018.

[64] Zhihao Zhou, Sean Padgett, Zhixiang Cai, Giorgio Conta, Yufen Wu, Qiang He, Songlin Zhang, Chenchen Sun, Jun Liu, Endong 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.