PoseSonic: 3D Upper Body Pose Estimation Through Egocentric Acoustic Sensing on Smartglasses

SAIF MAHMUD, Cornell University, USA
KE LI, Cornell University, USA
GUILIN HU, Cornell University, USA
HAO CHEN, Cornell University, USA
RICHARD JIN, Cornell University, USA
RUIDONG ZHANG, Cornell University, USA
FRANÇOIS GUIMBRETIÈRE, Cornell University, USA
CHENG ZHANG, Cornell University, USA

In this paper, we introduce PoseSonic, an intelligent acoustic sensing solution for smartglasses that estimates upper body poses. Our system only requires two pairs of microphones and speakers on the hinges of the eyeglasses to emit FMCW-encoded inaudible acoustic signals and receive reflected signals for body pose estimation. Using a customized deep learning model, PoseSonic estimates the 3D positions of 9 body joints including the shoulders, elbows, wrists, hips, and nose. We adopt a cross-modal supervision strategy to train our model using synchronized RGB video frames as ground truth. We conducted in-lab and semi-in-the-wild user studies with 22 participants to evaluate PoseSonic, and our user-independent model achieved a mean per joint position error of 6.17 cm in the lab setting and 14.12 cm in semi-in-the-wild setting when predicting the 9 body joint positions in 3D. Our further studies show that the performance was not significantly impacted by different surroundings or when the devices were remounted or by real-world environmental noise. Finally, we discuss the opportunities, challenges, and limitations of deploying PoseSonic in real-world applications.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Human pose estimation, Acoustic sensing, Smart/AR glasses, Deep learning, Cross-modal supervision

ACM Reference Format:
Saif Mahmud, Ke Li, Guilin Hu, Hao Chen, Richard Jin, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2023. PoseSonic: 3D Upper Body Pose Estimation Through Egocentric Acoustic Sensing on Smartglasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 111 (September 2023), 28 pages. https://doi.org/10.1145/3610895

1 INTRODUCTION

Smartglasses have become an important form of personal wearable computing, but they face limitations when it comes to sensing the user's body postures. Due to restrictions in weight, size, and battery consumption of

Authors' addresses: Saif Mahmud, Cornell University, Ithaca, NY, USA, sm2446@cornell.edu; Ke Li, Cornell University, Ithaca, NY, USA, kl975@cornell.edu; Guilin Hu, Cornell University, Ithaca, NY, USA, gh386@cornell.edu; Hao Chen, Cornell University, Ithaca, NY, USA, hc732@cornell.edu; Richard Jin, Cornell University, Ithaca, NY, USA, rj284@cornell.edu; Ruidong Zhang, Cornell University, Ithaca, NY, USA, rz379@cornell.edu; François Guimbretière, Cornell University, Ithaca, NY, USA, francois@cs.cornell.edu; Cheng Zhang, Cornell University, Ithaca, NY, USA, chengzhang@cornell.edu.

2474-9567/2023/9-ART111 $15.00
https://doi.org/10.1145/3610895

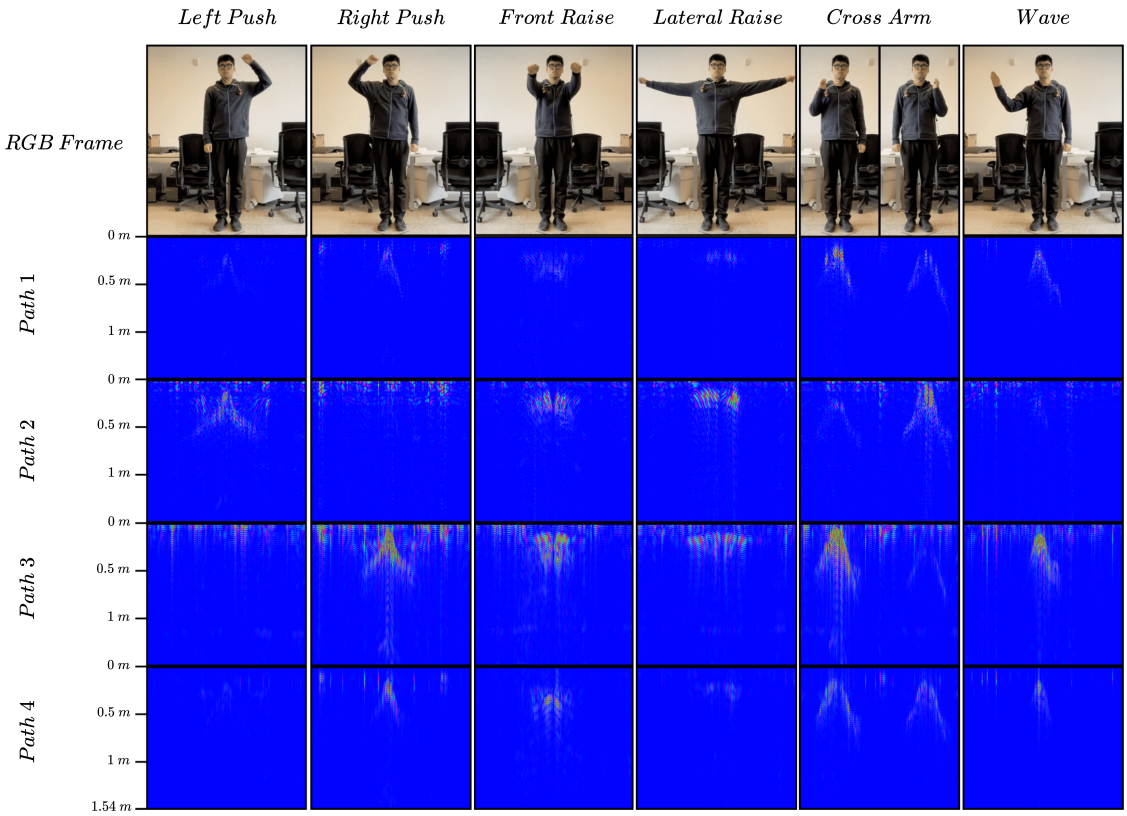


Fig. 1. Overview of the sensing technique adopted in PoseSonic: The first row denotes the RGB video frames of upper body movements. The bottom four rows indicate the differential echo profile computed for estimating body joint coordinates. Since we use pair of speakers and mics on both sides of the eyeglass, the acoustic signal can travel to any mic in four different paths. The paths are defined as follows: **Path-1:** Left Speaker to Right Mic (LR), **Path-2:** Left Speaker to Left Mic(LL), **Path-3:** Right Speaker to Right Mic (RR), **Path-4:** Right Speaker to Left Mic (RL)

the hardware components inside the glasses, most current smartglasses are only able to capture the motion of the head using the inertial measurement units (IMUs), touch input using a touchpad, or speech interactions using a microphone. However, none of these sensors are capable of capturing body poses critical for a range of applications, such as activity recognition [23], health monitoring [36], and virtual reality [4]. Upper body pose, for instance, is essential for virtual and augmented reality interactions and is involved in almost all types of physical activities. Tracking upper body poses allow computers to better understand and interpret a user's behavior and intention, such as detecting when the user is eating.

Tracking upper body pose has been proved to be an extremely challenging task for smartglasses, as they are not directly attached to any part of the limbs. One of the most popular approaches [39] to track body pose is using egocentric cameras to capture body images from the head, an approach that has been widely adopted on head-mounted devices such as virtual reality headsets. However, unlike VR headsets which have a large battery and form factor, the small form factor and battery size of smartglasses make it extremely challenging to add

an egocentric camera. Additionally, even if it is possible to add cameras to smartglasses, keeping the camera running would quickly exhaust the battery and CPU, and the camera may also capture images of the user and the surrounding environment, potentially raising privacy concerns. As a result, we have not seen any commodity smart glasses, other than VR headsets, incorporating camera-based solutions for upper body pose tracking. There is a clear need for a new lightweight sensing solution for smart glasses to track fine-grained upper-body poses.

In this paper, we present an intelligent upper body pose estimation solution on smartglasses that utilizes a low-power, minimally-obtrusive, and privacy-sensitive acoustic sensing technique. Our method allows the smartglasses to track upper body poses including the 3D positions of upper limbs without the need of collecting any training data from the new user. Two pairs of MEMS microphones and speakers need to be attached to the two hinges of an off-the-shelf glass frame. The speakers emit inaudible, Frequency Modulated Continuous Wave (FMCW)-encoded acoustic signals that are reflected by different body parts, such as the arms, hands, torso, and shoulders. The two microphones on the glass frame capture these reflected signals, which are then analyzed to extract the reflection information in the form of echo profiles (detailed discussion in Sec. 4.3), as illustrated in Figure 1. To learn the complex mappings between the captured acoustic signals and upper body poses, we develop a customized convolutional neural network. The network takes time series of both original and differential echo profiles extracted from acoustic signals as input and outputs the 3D coordinates of upper body parts, including nose, shoulders, wrists, elbows, and hips.

To evaluate the performance of the PoseSonic, we conducted an in-lab user study with 12 participants and a semi-in-the-wild study with 10 participants, where each participant performed a variety of pre-defined upper body poses wearing the commodity glass integrated with our solution. We conducted the studies in 5 rooms with different furniture settings and two places outside the lab (sidewalk and cafe). Each participant remounted the glasses multiple times during the experiment. The results show that, without the need of collecting any training data from a new user, PoseSonic can estimate the upper body poses including 9 (nine) 3D body joint positions for them with an average localization error of 6.169 cm in the lab setting and 14.119 cm in semi-in-the-wild setting in a leave-one-participant-out experiment. The performance can be further improved to 5.633 cm in the lab setting if we fine-tune the model with the participant's own training data. On the other hand, the performance in a naturalistic setting can be improved to 12.004 cm with an addition of 30 minutes of training data not necessarily from the same user. Furthermore, by using the data from different rooms as testing data, we show that there is no significant difference in performance when the rooms and furniture are different. Moreover, the PoseSonic system demonstrates robustness to noise in different real-world settings.

To the best of our knowledge, PoseSonic is the first acoustic-sensing technology on smartglasses that tracks upper body poses. Different from many data-driven sensing approaches which require training data from each user to fine-tune the model, our system works well even without the need for collecting training data from a new user, as shown by the promising performance in a user-independent evaluation. Additionally, the microphones and speakers are lightweight and have a low-power signature. Therefore, we believe that PoseSonic fills the gap of the smartglasses-based body pose tracking by showing the feasibility of a novel acoustic-based body-pose sensing solution that has the potential to be deployed on commodity smartglasses in the future.

Our contributions are summarized as follows:

- We presented the first acoustic sensing system on smartglasses that can estimate upper body poses.
- We designed and implemented a customized deep learning framework with cross-modal supervision to learn 9 key points on the upper body without any manual labeling.
- We conducted two user studies with a total of 22 participants in lab setting and the study demonstrates the promising performance of PoseSonic across different participants and environments without the need of collecting training data from the new participant nor the environment.

- We conducted a semi-in-the-wild study with 10 participants and the study demonstrates the robustness of PoseSonic system to real-world noise and random user motion.
- We discussed the opportunities, challenges, and limitations of PoseSonic within the context of deploying it in real-world applications.

2 RELATED WORKS

Research in human pose estimation can be broadly classified into two categories: non-wearable and wearable-based. This section provides an overview of the existing work in both categories and examines their relevance to our approach.

2.1 Non-wearable Based Pose Estimation

2.1.1 Computer Vision-based. The existing body of research on pose estimation relies mostly on computer vision-based approaches which require the installation of RGB or specialized cameras. Researchers apply deep learning techniques [6, 9, 21, 32, 43, 50, 52, 56] to estimate 2D human pose based on labeled pose databases. Meanwhile, the construction of 3D human pose from a single image [15, 17] or video [33, 53] also got attention from the computer vision community. In addition, the potential of depth or infrared sensors in estimating human pose has been explored in some recent research [27, 47, 51]. Microsoft Kinect [1] and LeapMotion [12, 42] utilizes depth or infrared sensors alongside RGB camera for estimating human pose and these systems are already in commercial use. Moreover, the use of built-in cameras and sensors of the smartphone [3, 5] shows promising results in approximating human body movements. Although vision-based systems offer fine-grained information about the human body, occlusion is a fundamental challenge to these systems. As a result, there is usually a degradation of performance in the scenario of non-line-of-sight or poor lighting conditions. Moreover, there is increasing concern about the privacy of the users as well as bystanders of the vision-based systems. For instance, the leakage of recorded videos of users and their environment is one of the major concerns.

2.1.2 Other Types of Non-wearable Sensors. In addition to computer vision-based systems, other types of sensors can contribute to body pose estimation. For example, RF-Pose [54] and RF-Pose3D [55] present approaches to estimate human poses using radio frequency signals. These systems accurately estimate human poses despite occlusions between the body and the installed radio. Furthermore, some other existing pose estimation systems [16, 34, 35] leverage the WiFi signal to track motions. In addition to that, some recent research [2, 18] has utilized off-the-shelf mmWave radar to estimate the body pose of multiple users simultaneously. These systems also utilize the reflection of different signals from the body to estimate users' body poses. However, most of the systems require the instrumentation of the environment through the installation of single or multiple sensors at a particular place. It is a major bottleneck to the movability of the subject and deployment in the wild.

2.2 Wearable-based Pose Estimation

In order to solve the problems of occlusions and movement of users of the aforementioned non-wearable systems, researchers have put much effort into developing wearable systems to track human poses. Wearable-based systems overcome the constraints of movability by widely adopting Inertial Measurement Units (IMUs) in pose estimation wearable systems [8, 14, 24, 26, 41]. However, IMU-based sensing techniques do not capture fine-grained data on the movement of different body parts [40]. On the other hand, some recent works exhibit the feasibility of wrist-worn cameras [11, 22] to track human body pose. Other works place regular or fish-eye cameras on hats [49], VR headsets [39] or chest mount [13, 31] to capture body pose. Furthermore, acoustic sensing has been deployed to capture the distance between a pair of wearables to infer human body pose [19]. For most of the wearable systems above, the user is required to wear multiple sensors at different body locations to obtain reasonable

performance, which is not practical in real-life scenarios. Apart from that, the size and power consumption of these systems are major setbacks in real-world deployment for continuous estimation.

Although smart glass has the potential to become a widely adopted wearable device in the future, the feasibility of building sensing systems on smart glasses for body pose estimation has not been fully explored. Additionally, RGB cameras on smart glass create privacy leakage issues and thus might not be acceptable as a sensing modality for pose estimation in many daily living and social interaction scenarios. Therefore, we propose PoseSonic which is an acoustic sensing system for smart glass to estimate the upper limb human body pose.

3 THEORY OF OPERATION

PoseSonic aims to estimate upper body posture through the use of acoustic sensors on eyeglasses, with a focus on privacy preservation, less intrusive design, and low power consumption. This idea of using the reflection patterns of acoustic signals to track human activities has been explored before for topics such as facial expression tracking [20, 46] and gesture tracking [30, 45]. However, there has not been any prior work utilizing acoustic sensing to track human poses. Similar to facial expressions and finger gestures, the upper body posture is composed of various movements involving the shoulders, arms and hips. Each component has a unique position and moves at different directions and speeds when people perform different gestures. As a result, the acoustic signal emitted from the speaker is reflected and diffracted differently at each component before being received by the microphones. This generates unique patterns in the reflected signal which will help PoseSonic estimate the upper body posture using a customized deep learning network.

To validate the feasibility of this approach, we conducted an experiment where three researchers performed various upper-body movements while wearing glasses with speakers and microphones on both hinges. By generating echo profiles through a signal processing pipeline (detailed in later sections), we have identified clear patterns in the reflected acoustic signal, as shown in Fig. 1, that demonstrated the feasibility of our approach. With the aforementioned objectives in mind, we conducted extensive research and experimentation on various acoustic sensing configurations and learning algorithms. After a thorough evaluation, we selected a system that best met our goals and documented the process and results in the following sections.

4 DESIGN AND IMPLEMENTATION OF SENSING SYSTEM

4.1 Frequency Modulated Continuous Wave (FMCW)

The naïve approach to using audio signals to measure distance is to transmit a sharp pulse and receive the signal reflected from objects. By measuring the time delay τ between the transmitted and received signals, the distance can be calculated with $d = v\tau/2$, where v is the speed of sound and d is the distance between the audio source and object. The result is divided by 2 since d is the round-trip distance that the signal travels between the source and the object. Although this approach is straightforward, the sharp pulse transmission requires a large bandwidth and thus consumes high power [44].

To tackle these challenges, we decided to use the Frequency Modulated Continuous Wave (FMCW) technique. FMCW is a type of signal that indirectly measures the distance between the signal source and objects by using the difference in frequency between the transmitted and received signal. An example of the FMCW signal is provided in Fig 2, where τ represents the time delay between the transmitted and received signals. The frequency of the signal increases linearly over time during each sweep period. The frequency of the acoustic FMCW signals can be represented as a linear function of time. At the time t, the frequency of FMCW is $f(t) = f_c + \frac{B}{T} \cdot t$, where f_c is the carrier frequency, B is the bandwidth and T is the sweep period. The resolution δR is the minimum distance that the FMCW signal can appreciate and is measured as $\delta R = \frac{v}{2f_s}$ where v is the velocity of sound and f_s is the sampling rate. The maximum theoretical range of the signal is computed as $R_{max} = \delta R \cdot N = \frac{v}{2f_s} \cdot N$, where N is the number of samples in one sweep period.

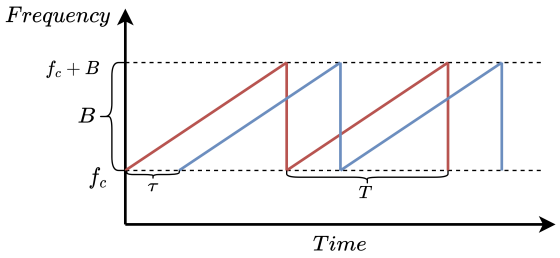


Fig. 2. Transmitted (red) and received (blue) signal of FMCW for two sweep periods

4.2 Acoustic FMCW Signal Configuration

The PoseSonic system is equipped with speakers emitting FMCW acoustic signals. The signal configurations of our system are determined as follows. Firstly, sounds around 17 KHz and above are inaudible to most adults [7]. Secondly, the microphones and speakers in the commodity wearables (e.g., smartglasses, earables) can be operated up to the Nyquist frequency of 22.05 KHz with a sampling rate of around 44.1 KHz. Thus, our choice of frequency aims to find a solution that is potentially applicable to commodity microphones and speakers without losing performance. Therefore, we chose the carrier frequency f_c as 18 KHz and 21.5 KHz respectively for the left and right channels for the two primary reasons above. The bandwidth B is set to be 3 KHz.

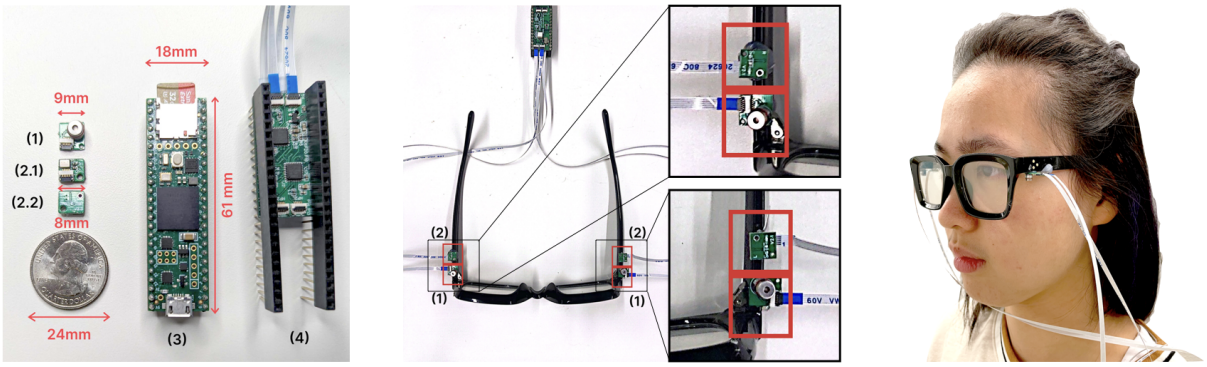
As a result, the FMCW chirp sweeps within the frequency range of 18 - 21 KHz for the left channel and 21.5 - 24.5 KHz for the right channel so that the two channels do not interfere with each other. The sampling rate f_s of our FMCW signal is 50 KHz. The sweep period of the signal is set to be $N = 900$ samples or 18 millisecond ($\frac{900}{50KHz}$). The speed of sound v in dry air at $20°C$ ($68°F$) is 343 m/s. Given the above parameter values, the resolution of our system is $\frac{343}{2 \cdot 50000} = 0.343$ cm, which is high enough to capture subtle movements of the body. Our sensing mechanism also covers the whole body since the maximum range of the signal is $\frac{343}{2 \cdot 50000} \cdot 900 = 3.087$ m or around 10 feet, which is higher than the average human height.

4.3 Echo Profile Calculation

Our system applies cross-correlation to measure the time delay between the transmitted and reflected signals. We used the cross-correlation function defined in [44]:

$$R(n) = \begin{cases} \frac{1}{N-n} \sum_{m=0}^{N-n-1} v_{tx}(m) \cdot v_{rx}(m+n), & \text{if } n \geq 0 \\ \frac{1}{N-|n|} \sum_{m=0}^{N-|n|-1} v_{rx}(m) \cdot v_{tx}(m+n), & \text{if } n < 0 \end{cases} \quad (1)$$

where N represents the number of samples in each sweep period, $n = -N+1, -N+2, ..., N-1$ and $R(n)$ is the similarity between transmitted signal v_{tx} and received signal v_{rx} shifted away by n samples. As defined in EarIO [20], the cross-correlation has the same period as the transmitted signal, and the output of the cross-correlation operation in that period can be named as an *echo frame*. Through transmitting FMCW acoustic signal repeatedly, we obtain an array of consecutive echo frames and this array is defined as *echo profile*. If we consider an echo frame as a column vector, the vertical axis of the echo profile represents the distance between the object and the sensor, and the horizontal axis is the temporal axis. Each bright spot on the echo profile means a strong reflection signal is received at the corresponding distance and time. Since the period of the signal is 900 samples and our device contains two pairs of sensors mounted on the left and right hinges of the glass frame respectively, the echo profile is a 3600-dimensional vector with four 900-dimensional channels stacked together (4 possible speaker-microphone links between 2 speakers and 2 microphones as depicted in Fig. 1).



(a) Sensors and Microcontroller (b) Form Factor Prototype (c) User Wearing the Prototype

Fig. 3. Hardware and form factor of PoseSonic: In Subfig. **(a)**, **1:** Transmitter or speaker, **2.1 and 2.2:** Two sides of the receiver or microphone board, **3:** Microcontroller (Teensy 4.1) to interface speakers and mics, **4:** Customized adapter PCB. In Subfig. **(b)**, Speakers and mics connected to the microcontroller using FPC cables.

We further process the echo profile to acquire a *differential echo profile* by calculating the absolute difference between two consecutive echo frames. The main advantage of a differential echo profile is that it preserves the changes in the echoes across consecutive frames and removes unchanged signals reflected from static objects. As a result, it contains less noise from the surroundings and yields a higher signal-to-noise ratio (SNR). In addition, a differential echo profile is particularly helpful in mitigating issues of remounting. Although a change in the position of the device alters the absolute distance between the sensors and targets, such difference will not impact the differential echo profile which only considers the change across consecutive echo frames. Therefore, the same physical action or movement will offer a similar differential echo profile irrespective of the physical characteristics of the user.

4.4 Implementation of Sensing System

To implement the FMCW-based sensing system we designed above, two speakers (OWR-05049T-38D[1]) and two microphones (ICS-43434[2]) are adopted in the system. A micro-controller Teensy 4.1[3] is used to manage the signal emission and reception. It has an SD card interface on board so the received data can be saved into an SD card. To better control the audio signal, we customize a Printed Circuit Board (PCB) with two SGTL5000 chips[4] that is compatible with the micro-controller. All the hardware components communicate with each other through the Inter-IC Sound (I2S) interface. The microcontroller, the speaker, the microphone, and the customized PCB are shown in Fig. 3(a).

Considering our goal is to track the posture of users' upper body, we chose a pair of glasses as our form factor and deployed our sensing system on it. As displayed in Fig. 3(b), we place one speaker and one microphone on each hinge of the glasses, pointing downwards. The system includes two pairs of speakers and microphones symmetrically so that each one contains information mainly from one side of the body. Besides, the sensors are pointing downwards so that the signals are directly emitted towards the upper body and the microphones have a better view of receiving signal reflections. The speakers and microphones are connected to the microcontroller

[1] https://www.digikey.com/en/products/detail/ole-wolff-electronics-inc/OWR-05049T-38D/13683703
[2] https://invensense.tdk.com/products/ics-43434/
[3] https://www.pjrc.com/store/teensy41.html
[4] https://www.mouser.com/new/nxp-semiconductors/nxp-sgtl5000-low-power-stereo-codec/

using Flexible Printed Circuit (FPC) cables. We believe this prototype is minimally-obtrusive and does not have a significant impact on users' daily activities. Besides, this form factor guarantees the stability of the system and is close enough to the users' upper body for tracking its posture.

5 DEEP LEARNING FRAMEWORK

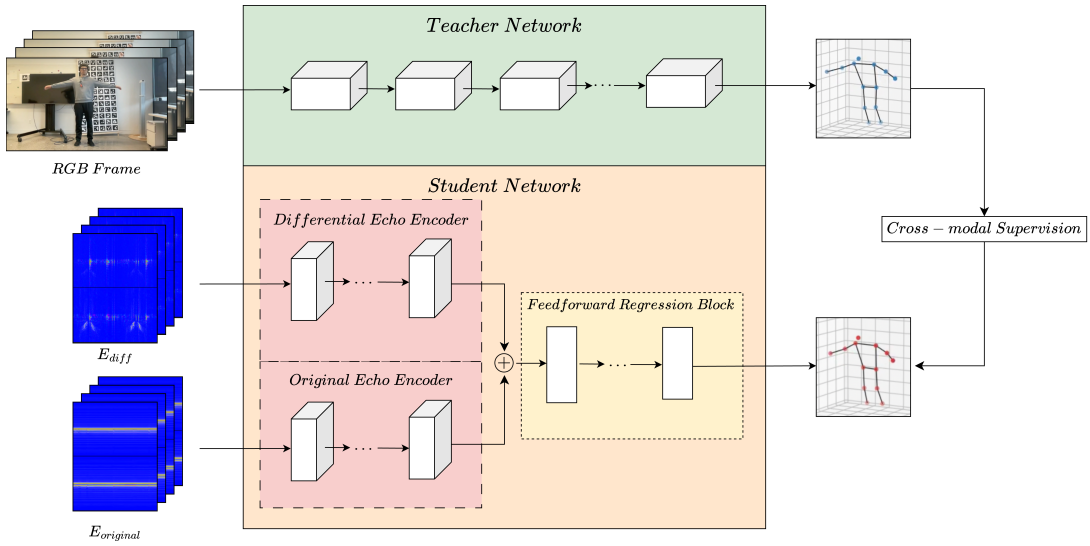


Fig. 4. Deep learning architecture of PoseSonic. **Teacher Network:** Pre-trained BlazePose GHUM 3D model [48] consisting of MobileNet-V2 [37] convolutional architecture, **Student Network:** Model with two branches for original and differential echo profiles respectively where each branch is a ResNet18 [10] encoder. The input to the feedforward regression network is the elementwise summation of representation vectors retrieved from two branches.

The architecture of the deep learning framework for estimating upper body pose is illustrated in Fig. 4. The design of this framework follows a teacher-student training strategy. The pipeline demonstrated in the upper part of Fig. 4 is the teacher network which provides supervision to the student network illustrated in the bottom part.

5.1 Cross-Modal Supervision

Since it is not possible to manually label the acoustic signal with corresponding human upper limb joint coordinates, the student network designed for acoustic data utilizes cross-modal supervisory signal obtained from the teacher network to learn this correspondence. The teacher network takes RGB video frames synchronized with the acoustic data as input. In this work, we utilize a pre-trained neural network [48] as the teacher network and it predicts nine 3D upper body joint coordinates from RGB video frames. On the other hand, the student network takes echo profiles as input and predicts upper limb joint coordinates. The training of the student network is guided by the prediction retrieved from synchronized video frames by the teacher network.

Let's consider a pair of synchronized RGB video frame and echo profile frames as (I, E) where I denotes an RGB frame and E denotes a pair of original echo profile $E_{original}$ and differential echo profile E_{diff}. The pre-trained teacher network $T(\cdot)$ predicts joint coordinates $T(I)$ from the RGB video frame I. The student network $S(\cdot)$ takes $E = (E_{original}, E_{diff})$ as input and predicts upper limb joints $S(E)$. The training objective of the student network is to minimize the difference between predictions $T(I)$ and $S(E)$ and it can be formulated as:

$$\min_{S} \sum_{\forall (I,E)} L(T(I), S(E)) \tag{2}$$

Here, $L(\cdot)$ denotes the loss function of the deep learning model and we defined it as the Mean Squared Error (MSE) between the joint coordinate prediction from the teacher network and the student network:

$$L(T, S) = \frac{1}{N} \sum_{i=1}^{N} (T_i - S_i)^2 \tag{3}$$

where T_i and S_i are the prediction of i-th batch from the teacher and student network respectively, and N is the total number of batches.

5.2 Network Architecture

The original and differential echo profiles (details in 4.3) represent movements of different body parts. Taking into consideration that the system is located on eyeglasses and we intend to predict upper body pose with it, we crop 450 pixels (which represent the range $\sim$ 1.5m) of each channel in the echo profiles. Then, we segment the echo profiles into overlapping sliding windows. As we have four channels in the echo profiles, the shape of each input original and differential echo profile will be $(4 \times \text{sliding window length} \times 450)$. The student network for estimating upper body pose takes these sliding windows of the original echo profile and differential echo profile as input.

We design the network consisting of two branches to encode original and differential echo profiles respectively. In this regard, we adopt ResNet18 [10] convolutional neural network as the encoder backbone in both branches followed by an average pooling layer on the spatial axis. Each encoder learns a 256 dimensional representation of the echo profiles. We compute the element-wise summation of these two representation vectors and it serves as the input to the feedforward regression layer. The regression layer outputs the prediction of the coordinates of nine 3D upper body joints as a vector with dimension $9 \times 3 = 27$.

5.3 Training and Implementation

The size of the input sliding window is tuned as a hyperparameter and set to 2 seconds (112 samples of echo frame) for this system. Therefore, the shape of the input echo profile will be $(4 \times 112 \times 450)$. In order to tune this hyperparameter, we create a holdout validation set using 20% of the training data. Then, we iterate over the range of sliding window sizes starting from 0.10 second to 7.50 second with a hop size of 0.10 second. For each sliding window size, we train the model with training data excluding the validation set for 30 epochs and evaluate the performance on the holdout validation set. We then chose the value of sliding window size that yielded the best performance as a tuned hyperparameter. After preprocessing the input data with the tuned value of the sliding window (2 seconds with 50% overlap), we have approximately $60,000$ data samples for each participant in the user study. We use batch normalization after each layer in the encoder followed by the sigmoid activation function. The dropout probability for the feedforward regression layer is set to 0.2. We train the model for 30 epochs with a batch size of 48. We use Adam as an optimizer and deploy a learning rate scheduler with an initial value of 10^{-5} and decay factor $\gamma = 0.1$ in each 10 epoch. The neural network architecture is implemented in PyTorch and PyTorch Lightning frameworks and trained on GeForce RTX 2080 Ti GPUs.

5.4 Evaluation Metric

We adopt Mean Per Joint Position Error (MPJPE) as the evaluation metric of the system. MPJPE is a widely adopted metric to evaluate 3D pose estimation performance [25] and has been reported in recent wearable [22] and non-wearable [35] based pose estimation techniques. Mean Per Joint Position Error (MPJPE) is defined as the

average of Euclidean distances between the predicted joint coordinates and ground truth joint coordinates. We compute MPJPE in the unit of centimeters and is defined as below:

$$Mean\ Per\ Joint\ Position\ Error\ (MPJPE) = \frac{1}{M}\sum_{j=1}^{j=M}\sqrt{\sum_{i=1}^{i=D}(q_i - p_i)^2} \tag{4}$$

where M is the total number of body joints to be tracked by the system ($M = 9$) and D is the dimension of joint coordinates which is $D = 3$ in our case.

MPJPE is computed after we align the estimated joint coordinates with ground truth coordinates. The ground truth acquisition method BlazePose GHUM 3D [48] in the teacher network provides normalized coordinates of upper body joints with the origin at the center of gravity of the human body which is the pelvis joint. We align the estimated upper body coordinates with the origin of ground truth coordinates using the Procrustes method [38]. Here, the Procrustes method [38] is an approach to finding a single orthogonal linear transformation such that the sum-of-square distances between the distribution point sets are minimized. The numerical value of the Mean Per Joint Position Error can be described with a geometric intuition. If we consider a sphere with a radius equal to localization error centering at the ground truth point of a particular joint, the prediction is expected to be any 3D point within the volume of this sphere of uncertainty.

6 IN-LAB USER STUDY

In this section, we provide the detail of the user study we conducted in the lab to evaluate the performance of PoseSonic. The goal of this study is to benchmark the performance of PoseSonic with respect to the variability across users and environments. In this regard, we design a set of action units to cover the range of upper-body movements and evaluate the estimated pose of PoseSonic in tracking those unit movements. We evaluate PoseSonic in two stages. In the first stage, we evaluate the performance of the system where the user is performing the action units while being static or standing in the same place. In the second stage, the users are in motion (i.e. walking) while moving their upper limbs according to the same set of action units.

6.1 Apparatus

In order to test the performance of our system on tracking upper body poses, we conducted a user study with the prototype introduced in Subsec. 4.4. According to Subsec. 5.1, the teacher network needs RGB video frames as input to train the model. Hence, we used laptops equipped with Intel RealSense D435 RGBD camera to record a video of participants when they performed different body gestures. In the meantime, we played instruction videos on the laptop, showing the body gestures to perform to the participants. The laptop was connected to a 24-inch LED monitor via cables to make it easier for participants to see the instruction video. In addition, the users were provided with Apple AirPods Pro such that they can listen to audio instructions while in motion or not facing the monitor. The study setup is demonstrated in Fig. 6 and Fig. 7.

6.2 Design Space of Action Units

The objective of the study is to validate the performance of our system in tracking upper-body poses. As a result, we designed 6 upper body action units involving movements of different parts of the upper body for participants to perform. Fig. 5 displays how these body gestures are performed. The body gestures according to this design are left push, right push, front raise, lateral raise, cross arm, and wave. Our design space was primarily influenced by the spatial location where upper limb movements take place. In order to evaluate a full range of upper body movements, we separate the body movements into three categories: asymmetric, symmetric, and complex. Asymmetric movements refer to activities that only involve one arm moving at a time. Symmetric

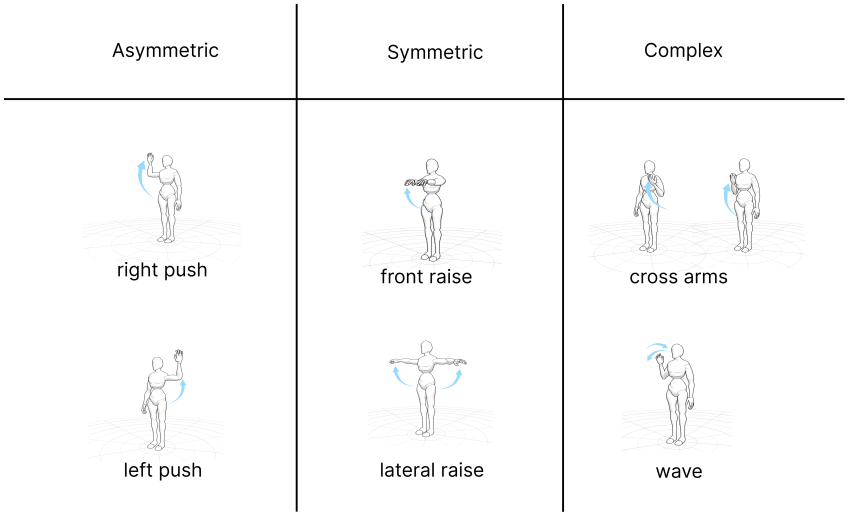
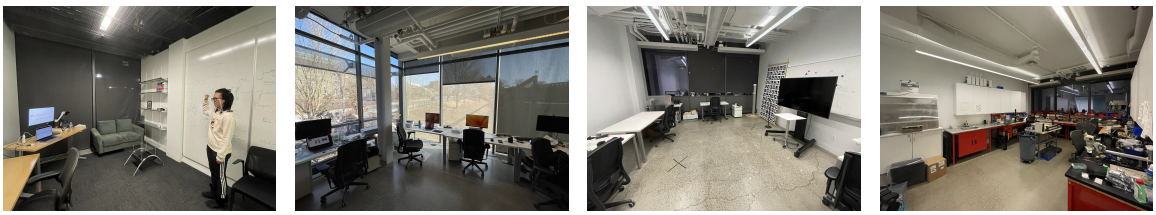


Fig. 5. Design space of upper body action units

movements refer to activities where both arms move in a similar pattern at the same time. We designed the action units in this way because our sensing system was placed on both hinges of smart glasses. Furthermore, we designed a third category - complex - which refers to activities involving both arms moving at the same time, but in different patterns. We designed the action units in this way because these complex movements are more similar to daily activities, such as waving one hand. Moreover, we can validate the performance of the sensing system under different movement synchronization conditions.

6.3 Study Design



(a) Main Study Room (b) Room A (c) Room B (d) Room C

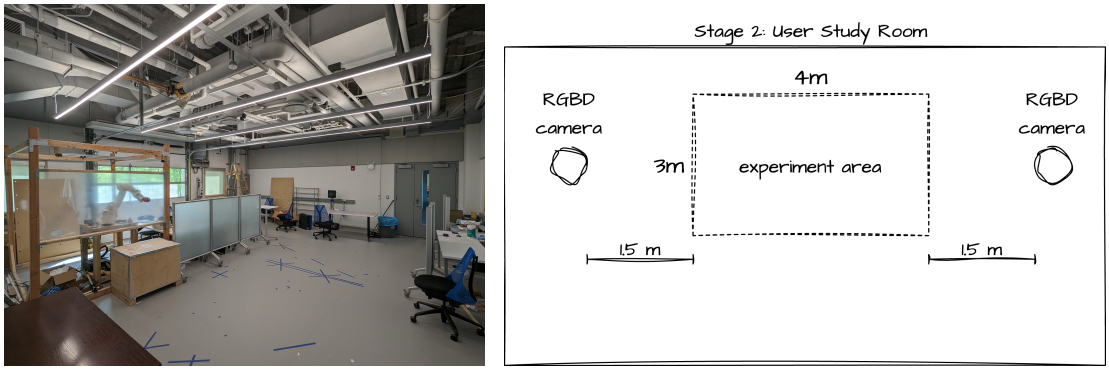
Fig. 6. User study setup of the first stage in four different rooms

6.3.1 First Stage: This stage of the study was conducted in four rooms in the same building on campus, as shown in Fig. 6. These rooms are usually used as offices, labs, and experiment rooms. When the study started, we introduced the study goal and procedures to the participants and obtained their consent to participate in the study. Then the participants were instructed to stand far enough from the laptop to make sure the camera

could capture the whole body of the participants. Afterward, we asked the participants to wear our prototype and perform physical movements defined by the action units (described in Subsec. 6.2) following the instruction video, as a practice session. After the participants were familiar with the gestures to perform and the remounting process, the formal study began.

In each session of the formal study process, the instruction video would show 5 repetitions of each of the 6 action units. The gestures were displayed in a random order and each gesture was shown for 6 seconds so that the participants had enough time to finish the gesture. Therefore, each session of our study lasted for 3 minutes (6s × 6 gestures × 5 repetitions). The laptop also played audio together with the instruction video in case the participants were not able to see the video clearly. After each session, we asked the participants to remount the device and take a break if needed.

A total of 13 sessions described above were collected in the main study room (depicted in Fig. 6(a)) for our study. This part was the major part of our study. Among these 13 sessions, 12 sessions were used as the training and calibration datasets of the model while the remaining one session was used for testing. Then we asked the participant to go to two randomly selected rooms from the remaining three (Room A, B, and C) to collect one testing session from each room. With this step, we would like to validate that the performance of our system is invariant to the change of location. In short, we collected 15 sessions of data in total (equivalent to 45 minutes), which contains 12 training sessions and 3 testing sessions. Our experiment design ensures that for each participant, we have one testing session for each of the 3 different rooms.



(a) Study Room (b) Schematic Diagram of Study Setup

Fig. 7. User study setup in the second stage

6.3.2 Second Stage: In the second stage of the study, we evaluated the performance of PoseSonic under the condition of random user motion. We followed a similar initial procedure of briefing participants as in the first stage. The study was conducted in a very large room as depicted in Fig. 7(a). We defined a 4 meter × 3 meter experiment area where the participants were allowed to randomly walk while performing the upper limb actions (as described in Subsec. 6.2). We instrumented the experiment area with two ground truth acquisition devices placed at 180° position as depicted in Fig. 7(b). Since the participants were walking during data collection, we provided the participants with both audio and visual instructions, which they listened to by wearing a pair of Apple AirPods Pro.

Each session of the formal study was of the same duration (3 minutes) and structure as in the first stage. After each session, we asked the participants to remount the PoseSonic device. We collected data from a total of 10 sessions where eight were used for training and two for testing.

6.4 Participants

The study was approved by the Institutional Review Board for Human Participant Research (IRB) of our institution. We successfully recruited 12 participants (4 females and 8 males) with an average age of (21.833 ± 2.368) years, ranging from 18 to 26 years old for the first stage of evaluation. In the second stage, we recruited a new group of 10 participants (6 females and 4 males) with an average age of (22.9 ± 3.247) years, ranging from 19 to 29 years

The overall duration of our study in any stage for each participant was no longer than 75 minutes. After the study, we collected demographic information from the participants.

7 PERFORMANCE EVALUATION OF IN-LAB USER STUDY

In this section, we outline the experiments carried out to assess PoseSonic's performance. First, we show the performance of the leave-one-participant-out cross-validation. Then, we present the results of the fine-tuned personalized model. Lastly, we detail the performance of PoseSonic under different environmental conditions by evaluating the performance of the test set collected in different rooms.

7.1 Evaluation of Leave-one-participant-out Experiment

We want to evaluate the generalizability of PoseSonic across different users and thus perform leave-one-participant-out cross-validation. As described in Sec. 6, we had 12 participants in the user study and recorded 15 sessions of data for each. In the leave-one-participant-out cross-validation experiment, we first hold out all 15 sessions of data for one participant and consider it as the test set. The remaining data from the other 11 participants is used for training the deep learning model. We then repeat the aforementioned process on each participant, generating 12 models in total. Please note that these are user-independent models which do not require training data from the user themselves. We report the mean per joint position errors for each participant of this cross-validation experiment in Table 1. In addition, we summarize the evaluation of the leave-one-participant-out experiment in Fig. 8. The blue dot in Fig. 8 represents the mean localization error in centimeters across all participants for the particular body joint denoted in the x-axis. The red line denotes the standard deviation of localization error for that body joint.

We observe that the localization error for both the right and left hip is very low compared to other body joints in Table 1. It can be attributed to the fact that the movement of hips compared to other upper body joints in our dataset is low. Therefore, it induces lower prediction error in the evaluation. If we exclude the mean localization errors of the hips, then the mean localization error across all participants is $7.751(\pm 4.754)$ cm in the leave-one-participant-out experiment.

7.2 Performance of Fine-tuned Personalized Model

The aforementioned leave-one-participant-out experiment validates the performance of PoseSonic with respect to user variability. Therefore, we can obtain the performance noted in Table 1 for the particular user without any requirement for training data. However, the performance for a particular user can be improved if the user-independent model described in Subsec. 7.1 is fine-tuned with data from that user. As mentioned in Sec. 6, we collect 15 sessions of data from each participant and each session lasts for 3 minutes. We hold out the last three sessions for each participant in the user study and include it in the test set. Hence, the test dataset of this experiment is one-fifth or 20% of the total data collected in the user study. The rest of the data is used for fine-tuning the user-independent pose estimation model (12 sessions per participant). The performance of

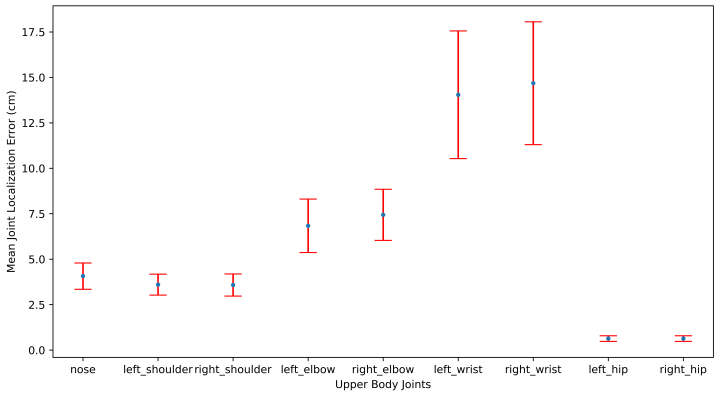


Fig. 8. Evaluation of the user-independent model for different upper body joints in leave-one-participant-out experiment

Table 1. Mean Per Joint Position Error (cm) for Each Participant in Leave-one-participant-out Experiment

Participant	nose	left shoulder	right shoulder	left elbow	right elbow	left wrist	right wrist	left hip	right hip	mean
P01	4.5811	3.8147	3.6610	7.0934	7.3905	14.2766	15.7059	0.5682	0.5645	6.4062
P02	3.9563	3.8035	3.5025	7.2667	7.6168	15.6616	16.1285	0.5831	0.5819	6.5668
P03	3.5808	3.2015	3.2244	5.8077	6.7344	11.7020	13.7535	0.5260	0.5234	5.4504
P04	4.0480	3.5361	3.1733	7.7231	8.6947	16.2302	18.9958	0.5703	0.5700	7.0602
P05	3.6278	2.7903	3.4093	5.4233	5.8295	10.4278	10.2144	0.4924	0.4919	4.7452
P06	3.6494	3.1759	3.0159	6.2114	7.3659	13.5301	15.4433	0.4557	0.4560	5.9226
P07	4.7267	3.9848	3.7620	8.1299	8.9405	17.1594	18.7400	0.5501	0.5488	7.3936
P08	6.2705	5.7956	6.2993	9.3760	9.7107	20.7617	16.8044	0.7048	0.6983	8.4913
P09	2.9244	2.8824	2.8157	6.5383	7.3927	12.9663	13.5415	0.5826	0.5802	5.5805
P10	4.7363	4.1569	4.1032	5.7571	6.7917	11.6215	12.6871	0.6795	0.6893	5.6914
P11	3.5270	3.3695	3.1099	6.7003	6.1610	12.7604	12.0721	1.0674	1.0599	5.5364
P12	3.2035	2.7143	2.8687	6.0212	6.6620	11.4247	12.1055	0.8287	0.8290	5.1842
Avg.	**4.0693**	**3.6021**	**3.5788**	**6.8374**	**7.4409**	**14.0435**	**14.6827**	**0.6341**	**0.6328**	**6.1691**
Std.	**0.9026**	**0.8376**	**0.9373**	**1.1498**	**1.1590**	**2.9641**	**2.7445**	**0.1697**	**0.1686**	**1.0639**

fine-tuned pose estimation models is summarized in Fig. 9. This figure represents the computed value of the evaluation metric Mean Per Joint Position Error (details in Subsec. 5.4) for the nine joints that PoseSonic is tracking. We find that the mean localization error across all participants for all nine body joints is 5.6336(±0.7831) cm. Moreover, we observe that the mean error and standard deviation are smaller for shoulders, elbows, and hips than the wrists. This phenomenon can be explained based on the sensing technique we are utilizing in PoseSonic. Since the input echo profiles fundamentally represent the distance of different moving parts of the body and wrists are the most dominant joint in most of the action units in our user study, it induces higher variability across sessions and participants. This lead to higher localization errors for the wrists.

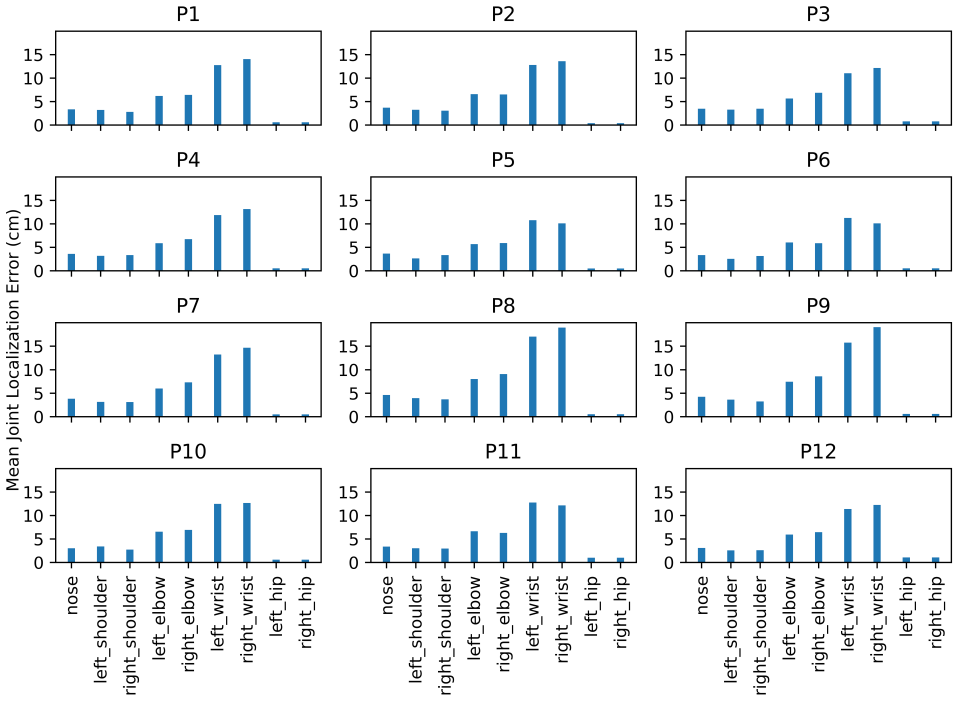


Fig. 9. Evaluation of the fine-tuned personalized model for different users across different upper body joints

7.3 Evaluation across Different Indoor Setting

We conducted a user study of 12 participants (details in Sec. 6) and collected data in four different rooms. Room-1 is a carpeted office space that is used as the main study room (image provided in Fig. 6(a)). Among the total of 15 sessions, we collected 13 sessions of data in that room. The other rooms are labs or office spaces without carpets and experiment rooms with electronics and prototyping instruments. If we label the three rooms except for the main study room with A, B, and C (as depicted in Fig. 6), then we can have $\binom{3}{2}$ choice of 2 room combinations which are the pairs (A, B), (B, C) and (C, A). We pick a random pair for any particular participant uniformly and independently and label them as Room-2 and Room-3. Afterward, we collect one session of data in each room. We construct a test set for each participant with one session of data collected in Room-1, Room-2, and Room-3 respectively, and train the model with the rest. We present the mean localization error in three different rooms across all body joints for each participant in Fig. 10. We conducted a one-way repeated measures ANOVA to compare the result in three different indoor settings. We compute that the f-ratio value is 0.271 (where degrees of freedom between groups $df_{between}$ and within groups df_{within} are 2 and 9 respectively) and the p-value is 0.765. Therefore, we do not find a statistically significant difference between tracking performance across rooms ($p = 0.765 > 0.05$). Hence, it could conceivably be hypothesized that different indoor settings such as the size of the rooms, furniture arrangement, floor surface, etc. do not have a significant impact on the performance of PoseSonic.

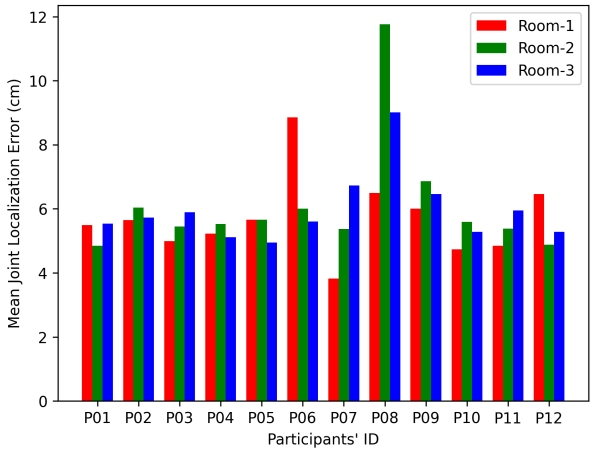


Fig. 10. Evaluation of PoseSonic across different rooms

Table 2. Mean Per Joint Position Error (cm) for Each Participant in Random User Motion Experiment

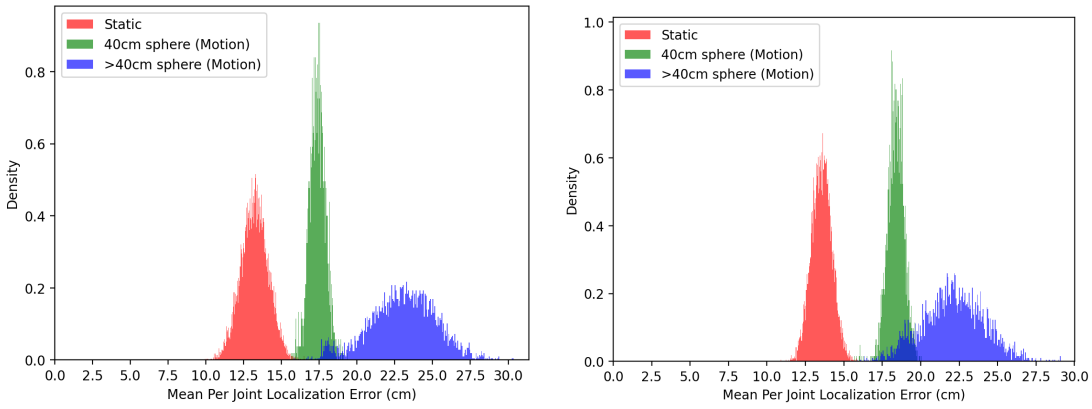
Participant	nose	left shoulder	right shoulder	left elbow	right elbow	left wrist	right wrist	left hip	right hip	mean
P01	5.513	16.674	17.240	15.169	14.344	21.852	22.313	9.103	9.109	14.591
P02	5.501	15.854	18.332	14.876	13.109	21.874	17.205	9.259	9.267	13.920
P03	4.521	16.477	17.418	14.748	14.033	20.156	21.381	9.144	9.145	14.114
P04	4.617	16.865	17.076	17.410	17.648	20.829	21.004	9.557	9.547	14.950
P05	2.281	16.968	16.398	16.139	16.716	19.506	18.214	9.415	9.427	13.896
P06	4.597	14.951	17.548	14.161	15.633	19.467	19.819	8.979	8.990	13.794
P07	3.678	15.666	16.721	17.209	16.667	19.638	19.969	9.145	9.141	14.204
P08	2.305	16.271	17.477	15.586	16.207	20.408	18.130	9.470	9.468	13.925
P09	4.607	16.971	17.602	15.643	15.206	18.971	21.848	9.726	9.732	14.478
P10	4.853	16.701	17.352	15.091	14.114	23.297	21.159	9.575	9.577	14.635
Avg.	**4.247**	**16.340**	**17.316**	**15.603**	**15.368**	**20.600**	**20.104**	**9.337**	**9.340**	**14.251**
Std.	**1.153**	**0.663**	**0.524**	**1.050**	**1.453**	**1.367**	**1.746**	**0.246**	**0.244**	**0.391**

7.4 Evaluation under the Condition of Random User Motion

We intend to evaluate the robustness of PoseSonic while the user is in motion. Here, the motion can be referred to as walking in a random pattern and with random head movements to observe the surrounding. As PoseSonic relies on the information encoded in the reflection of the ultrasonic acoustic signal, it captures reflection from the environment while the user is in motion. To evaluate this scenario, we designed the second stage of the in-lab user study as described in Subsec. 6.3.2. In this stage of the study, we collected 30 minutes of data in 10 sessions. We used data from 8 sessions (24 minutes) as training data and 2 sessions (6 minutes) of data for testing the performance. Here, we initialized the pose estimation model with the weights learned from training on the data of all participants in the first stage of the study. It should be noted that the training data in the first stage did not include any random user motion. We trained the model in a leave-one-participant-out fashion with the data collected in the second stage of the study. This user-independent model was evaluated with the testing data

collected in the second stage and Mean Per Joint Localization Error for each participant is reported in Table 2. The average MPJPE across all participants was 14.251 cm with a standard deviation of 0.391 cm. We observe that the wrists have the highest localization error of 20.6 cm and 20.104 cm for the left and right one respectively.

We observed in Table 2 that there was an increase in the error of tracking different body joints compared to the tracking in the first stage of the study. It can be attributed to the fact that when the user moves their head or walks, the original and differential echo profile captures reflection patterns from the objects in the environment if it is within the sensing range. However, these reflection patterns are mostly not similar to the ones created by moving upper limbs. Therefore, the joint coordinate estimation pipeline of PoseSonic shows robustness to environmental reflections and captures upper body poses under the scenario of random user motion.

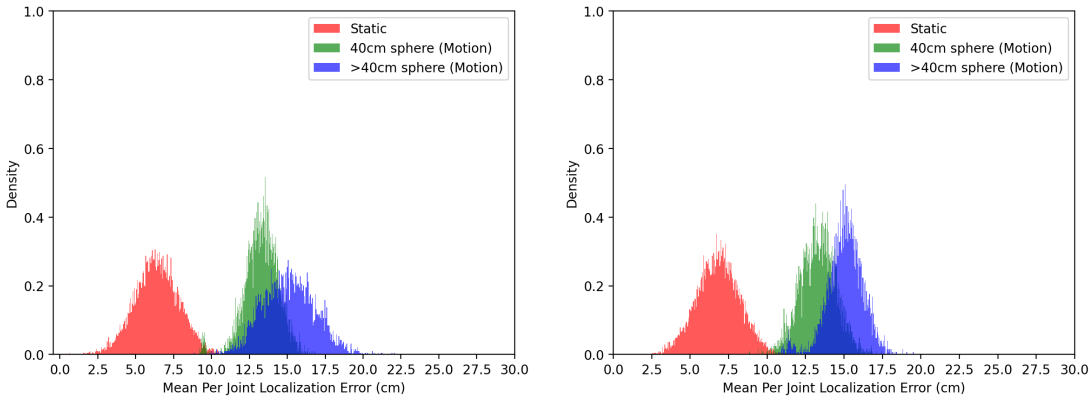


(a) Left Wrist

(b) Right Wrist

Fig. 11. Distribution of Mean Per Joint Position of Error (cm) of **wrist joints** across three different scenarios: **(1)** the participant is static while performing upper body movements, **(2)** the participant is in random motion and the wrist is being moved within the volume of a sphere with a radius of 40 cm and center at the bridge of PoseSonic, and **(3)** user in motion and wrist is being moved outside the sphere defined in (2)

When the participant is in random motion, the reflection from environmental objects is captured in echo profiles. Generally, these objects are not in very close vicinity (25 – 30 cm) of the sensing system instrumented at the hinges of PoseSonic. Therefore, if the wrists or elbows (which are the two body joints with the highest localization error) are moving closer to the sensing system, we observe less degradation in performance than wrist or elbow movement in further locations (greater than approximately 40 cm from PoseSonic). This phenomenon is illustrated in Fig. 11 and Fig. 12. For both wrists and elbows, if it is moving near the sensing system, it yields a lower localization error of the joints. In this regard, we define an imaginary sphere centering at the bridge of the PoseSonic frame and having a radius of 40 cm. Wrist and elbow movement within this sphere shows lower error in joint coordinate prediction (Mean Per Joint Position Error of approximately 17.5 cm for wrists and approximately 13 cm for the elbows) and it is similar to the error under the situation where the user is not in random motion. However, movements outside the aforementioned sphere yield higher localization errors as depicted in Fig. 11 and Fig. 12. Given that we do not observe a difference in behavior based on this boundary in the static setting, we believe that this could be attributed to the fact that the PoseSonic sensing system encounters different objects in the surrounding, and reflection from these objects create adversarial data samples that impact the performance of the deep learning model to capture body joint movements. Nevertheless, the performance in

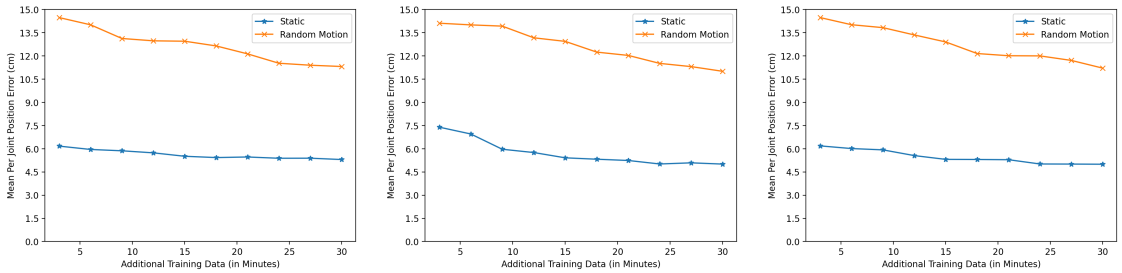


(a) Left Elbow (b) Right Elbow

Fig. 12. Distribution of Mean Per Joint Position Error (cm) of **elbows** across three different scenarios: **(1)** the participant is static while performing upper body movements, **(2)** the participant is in random motion and the wrist is being moved within the volume of a sphere with a radius of 40 cm and center at the bridge of PoseSonic, and **(3)** user in motion and wrist is being moved outside the sphere defined in (2)

this scenario can be improved with more data from users in different random motion situations and preprocessing pipeline for environmental noise elimination.

7.5 Impact of Additional Training Data on Estimation Performance



(a) P02 (b) P05 (c) P08

Fig. 13. Impact of additional training data on the performance of PoseSonic. The x-axis indicates the amount of additional training data in minutes and the y-axis indicates the Mean Per Joint Position Error in centimeters on the test set of that particular participant. The orange and blue lines demonstrate the evaluation metric in random user motion and static scenarios respectively.

In order to evaluate the impact of additional training data on the performance of PoseSonic, we conducted a follow-up experiment on three randomly selected participants. These participants were selected from the second stage of the in-lab user study and they are P02, P05, and P08. For each participant, we incrementally add training data and evaluate the performance on the test set of the participant in consideration. Here, the additional training

data did not come from the particular participant. There are two sources of this additional data: other participants' data in the formal user study and researchers' data in the pilot study which was conducted prior to the formal user study. The impact of adding more training data is depicted in Fig. 13. The inference evaluation plot for all three participants in Fig. 13 demonstrates that more training data lead to better inference performance on both static and random motion scenarios. However, the improvement of performance in the scenario of random user motion is more conspicuous. This experiment indicates that although PoseSonic shows promising performance without the need to collect any training data from a user, its performance can further be improved if it is provided with more training data. The purpose of this paper is to demonstrate the proof-of-concept of this novel sensing system on glasses. We will leave the study of how large-scale data sets improve the performance to future studies.

8 SEMI-IN-THE-WILD USER STUDY

In this section, we provide a detailed description of the semi-in-the-wild study that we conducted outside the lab setting where the users are exposed to environments with real-world noise. This study was intended to evaluate PoseSonic in an unconstrained naturalistic environment. In this regard, we designed three real-world scenarios where the participants perform upper body poses while walking:

- *Playing music in loudspeaker*: we asked the participants to select any music they want to listen to and play it in full volume on their phone (80dB to 90dB). While playing the music, they walked randomly in the same experiment area designed for the second stage of the in-lab user study (as described in Subsec. 6.3.2) and performed upper limb motions of action units.
- *Sidewalk of a road*: We selected a particular sidewalk in front of a building on the campus. There was generally light traffic on the road and moderate congestion of people on the sidewalk. Note that there was a construction site on the other side of the road and there was some noise from heavy machinery occasionally.
- *Cafe or restaurant*: We selected an indoor cafe in a building on campus. The cafe has two counters and a large seating area in a hallway. We conducted the study in the hallway of the cafe. We conducted the study in the cafe on both weekdays and weekends. Generally, the weekdays' meal time is the busiest hour and weekends are less crowded.

Since we used RGBD cameras as the ground truth acquisition method, we restricted the movement of the users to a certain area in all three scenarios. However, the participants walked randomly in the designated area and performed upper limb poses. In each real-world scenario, we collected data for two sessions where each session has the same structure and duration (3 minutes) as the in-lab user study (details in Sec. 6). There was a remounting of the device between each session of data recording. All the six sessions recorded in this study are used for testing the inference of PoseSonic under real-world noise scenarios in a naturalistic setting.

The same set of ten participants took part in this semi-in-the-wild user study after finishing the second stage of the in-lab user study. The users received audio instruction through a pair of Apple AirPods Pro and performed upper body actions accordingly. We did not impose any constraints on external factors such as congestion in the surrounding environment, users' clothing, study time selection, walking speed, etc. in this semi-in-the-wild study.

9 EVALUATION OF PERFORMANCE IN SEMI-IN-THE-WILD STUDY

9.1 Impact of Real-world Noise in Unconstrained Environment

In order to evaluate the robustness of PoseSonic under the condition of exposure to real-world noise, we designed a semi-in-the-wild study as described in Sec. 8. We collected data in three different scenarios of real-world noise. To train the pose estimation model, we undertook a data augmentation strategy. We recorded noises defined in the particular scenarios of real-world experiments and overlayed that information on the training data of the

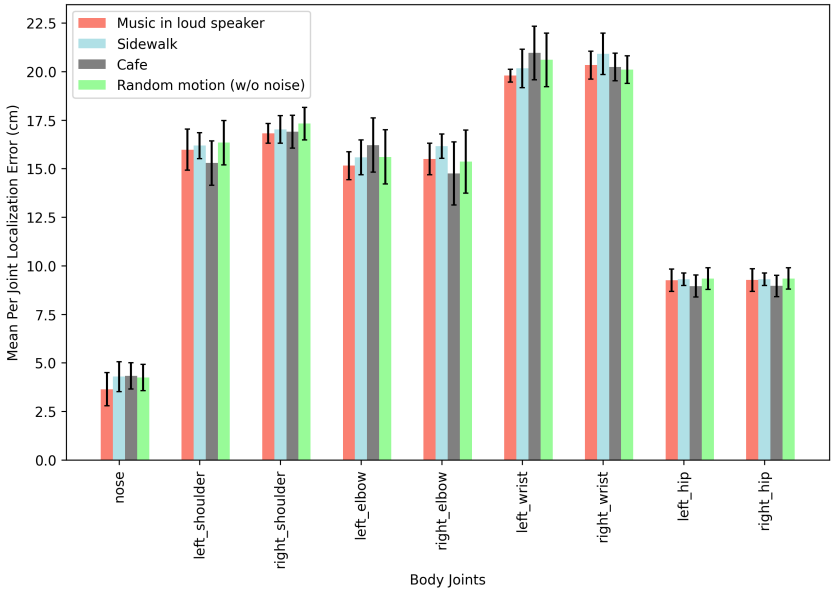


Fig. 14. Evaluation of PoseSonic under real-world noise in semi-in-the-wild study

second stage of the in-lab user study where the participants performed upper limb movement while in random motion. We then tested the model on the six sessions of data recorded as part of the semi-in-the-wild user study. Here, these data represent three real-world noise scenarios, which are music played on a loudspeaker, sidewalk, and cafe or restaurant.

We compare the performance of PoseSonic in predicting upper body joints in these scenarios. The comparison is depicted in Fig. 14. The Mean Per Joint Position Error across all users in the scenarios of music in a loudspeaker, sidewalk, and cafe are 13.967, 14.324, and 14.065 cm respectively. Please note that the same 10 participants participated in both the second stage of the in-lab study and this semi-in-the-wild study. The difference between the second stage of in-lab performance with random motion noise and evaluation in all three scenarios at the semi-in-the-wild study was less than 1 cm. It is evident in Fig. 14 that the pose estimation performance of PoseSonic is not significantly impacted by the real-world noises presented in these locations and settings. This highly promising result indicates the potential robustness of PoseSonic towards various real-world noises if deployed at scale in real-world settings.

9.2 Noise Injection Experiment

The PoseSonic system uses acoustic sensing to track human poses, so one might question whether the noises in the environment will have an impact on the performance of the system. In order to validate the research question, we recorded three different kinds of noises in different environments using our system and injected them into the testing sessions of the first stage of the user study data. We believe this is a good way to approximate the study in real noisy environments because the noise should be additive to the data in a linear system like ours.

The three kinds of noises we recorded are (1) Cafe noise (recorded at a cafe with people talking around, noise level: 63.8 dB(A)), (2) Curbside noise (recorded at the curbside with vehicles and people passing by, noise level: 69.0 dB(A)), and (3) Music noise (recorded when a song was played, noise level: 71.5 dB(A)). The evaluation

Table 3. Evaluation of Mean Per Joint Position Error (cm) of PoseSonic in Noise Injection Experiment

Noise Injection Scenario	Mean Per Joint Position Error (cm)
Without Noise	6.0096
Cafe Noise	6.7989
Curbside Noise	6.3086
Music Noise	5.6724

results on the testing sessions overlaid with the three noises are displayed in Tab. 3. As shown in the table, the performance of our system does not change a lot in different noisy environments, proving that our system is resilient to noises in everyday surroundings.

10 DISCUSSION

10.1 Validation of Ground Truth Acquisition Method

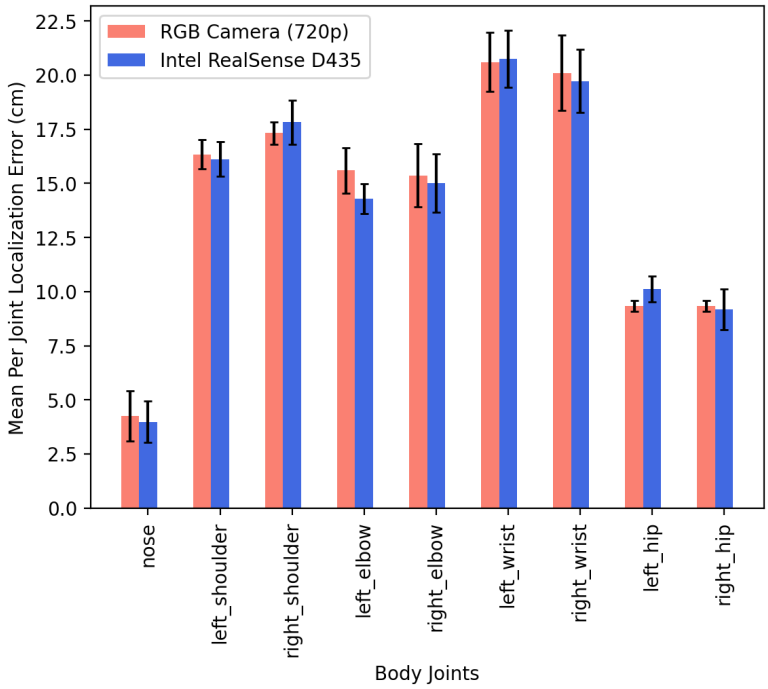


Fig. 15. Comparison of PoseSonic performance under the condition of two different ground truth acquisition modalities (720p RGB Camera and Intel RealSense D435 RGBD Camera)

In order to validate the ground truth retrieved by using the teacher network of PoseSonic in the cross-modal supervision setting (details in Sec. 5), we included Intel RealSense D435 as a ground truth acquisition modality in the second stage of in-lab and semi-in-the-wild user study. Using the RGBD sensing in the RealSense cameras, we computed normalized joint coordinates of the upper body. We then utilized these retrieved coordinates as ground

truth to train the deep learning model to estimate upper body poses from the acoustic signal. The comparison between the performance of PoseSonic using 720p RGB camera and Intel RealSense D435 RGBD camera as ground truth acquisition method is illustrated in Fig. 15. It is evident in the performance comparison that the pose estimation errors for both ground truth modalities are very similar. Furthermore, we validated the ground truth pose estimation of the GHUM [48] model (which is used as the teacher network in deep learning architecture) by using the Intel RealSense RGBD camera. In this regard, we estimated upper body joint coordinates utilizing the GHUM [48] model taking 720p RGB frames as input and skeleton pose estimation pipeline of the Intel RealSense camera. The margin of difference between the estimation from the two modalities in all testing scenarios is less than 1cm. However, one point worth noting is that the RGBD camera does not perform reliably under sunlight outdoors. Therefore, we relied only on the RGB frames to extract ground truth upper body joint coordinates in the sidewalk testing scenario of the semi-in-the-wild user study. Nevertheless, the margin of difference in different testing scenarios validates that the teacher network in cross-modal supervision setting (described in Sec. 5) provided reliable ground truth to the PoseSonic system to estimate upper body pose.

10.2 Comparison with Other Pose Estimation Methods

Table 4. Comparison of performance and sensing technique with other pose estimation methods

Method	**Sensing Technique**	**Mean Per Joint Position Error**	**Model Type**	**Notes on usage**
Pose-on-the-Go [3]	Depth camera and IMU on a smartphone	Shoulder: 10 cm Elbow: 12 cm Wrist: 27 cm	User independent	Includes poses while being static and in random motion
BodyTrak [22]	RGB Camera on a Wristband	Shoulder: 1.12 cm Elbow: 9.44 cm Wrist: 14.9 cm	User-dependent	All the motion in the study are static
GoPose [35]	Multiple WiFi antenna	Shoulder: ~5 cm Elbow: ~6 cm Wrist: ~8 cm	User-dependent	Pose estimation system installed in a room
PoseSonic	Acoustic Sensing	Shoulder: ~3 cm Elbow: ~6 cm Wrist: ~14 cm	User-independent	User can freely move wearing the system installed on eyeglasses

The goal of designing PoseSonic is to estimate the upper body human pose with a low power and privacy-preserving sensing technique on a minimally obtrusive form factor. To the best of our knowledge, PoseSonic is the first acoustic sensing system on the smartglasses form factor to estimate upper body pose. To help readers to better situate the performance of PoseSonic in comparison to prior work, we presented some of the most relevant prior work which also estimated body postures. As discussed in Sec. 2, there are numerous methods proposed to estimate human pose. Most of these methods rely on computer vision-based techniques. We note the performance of a few state-of-the-art methods in tracking upper limb movements in Table 4. We also note the usage and actions included in their evaluation. As mentioned in Table 4, Wifi-based pose estimation such as [35] requires extensive training data from each user and does not work when the user is not in an indoor setting. The most recent wearable-based pose estimation techniques are Pose-on-the-Go [3] and BodyTrak [22]. BodyTrak uses an RGB camera on the wrist and only evaluated the performance in a controlled lab study which was similar to our stage 1 user study in the lab. Our performance of 6.17 cm is significantly better than BodyTrak. Furthermore, BodyTrak requires training data from each user while PoseSonic work without any training data from users. The most recent work [3] also reported the performance in a user-independent evaluation under different scenarios. Our

performance was slightly better than theirs. Furthermore, [3] requires several devices including smartphones and other wearables to be worn at the same time. Our PoseSonic only needs glasses embedded with two pairs of cheap and low-power microphones. Therefore, all the factors considered, PoseSonic does advance the wearable-based pose tracking with strong performance, low energy consumption, and minimally-obtrusive form factor settings.

10.3 Power Consumption

We measured the power consumption of our system with a current ranger[5]. The system was powered with a DC voltage of 3.3 V and the current was measured to be 175.1 mA. Thus the power consumption of our system is 577.8 mW. This was measured with two speakers and two microphones operating, and with the data being written to the SD card on the microcontroller. This guarantees our system is low-power. With a battery of AR glasses like Google Glass (800 mAh), our system can run for around 4.5 hours if it is used alone.

10.4 Privacy Concern

The sensing system we developed in PoseSonic utilizes sound waves in the ultrasonic range (18KHz to 24.5KHz) to capture upper-body human movements. We apply a bandpass filter on the audio recorded by the receiver or microphone to segment signals within the frequency range of our interest. Therefore, PoseSonic does not require any sound to be recorded that includes our conversation or environmental sound in daily life to estimate upper body pose. This can mitigate the risk of user privacy leakage. However, the microphone may record audio that is not utilized by PoseSonic. This privacy threat can be addressed by processing the acoustic signals on edge devices locally. Although we did not investigate the deployment of PoseSonic on edge devices in this work, the required input data shape to the deep learning framework shows that the memory footprint of sensor data captured by PoseSonic is low, and thus the possibility of edge device deployment for more robust privacy protection can be explored in future studies.

10.5 Health Implication

PoseSonic constantly emits ultrasonic soundwaves to track human body movements. This might be viewed as a health concern. According to National Institute for Occupational Safety and Health (NIOSH) guidelines [29], it is possible for an individual to be subjected to a constant 85 decibels (dB) for a duration of 8 hours in the workplace before exceeding the maximum allowable daily limit. NIOSH recommendation is applicable for noises below 16kHz, which we did not use in PoseSonic. Moreover, we employed a CDC[6]-provided smartphone application to gauge the noise level of PoseSonic. Activating the speaker and placing the smartphone just beside it result in the app displaying a recording of around 68 decibels (dB). In addition, the frequency range of PoseSonic is not audible to most adults. However, it might still be audible to some animals and kids with a higher range of audibility. Moyano et al. [28] investigated the impact of ultrasound on human health under different conditions. The ultrasound waves can range from KHz to MHz range. There have been some impact on human body muscles such as reduction in muscle tone, release of muscle fluid, and increased membrane permeability of cells have been observed through the absorption of ultrasonic acoustic wave operating in MHz range. However, the literature does not present any concrete evidence of impact on human health with the ultrasound waves just above the human audibility range ($18 - 24$ KHz) where PoseSonic is operating. Nevertheless, the usage of ultrasound acoustic wave in commodity devices such as smartglasses require more attention and needs to be thoroughly evaluated for possible health implications in future works.

[5] https://lowpowerlab.com/guide/currentranger/
[6] https://blogs.cdc.gov/niosh-science-blog/2014/04/09/sound-apps/

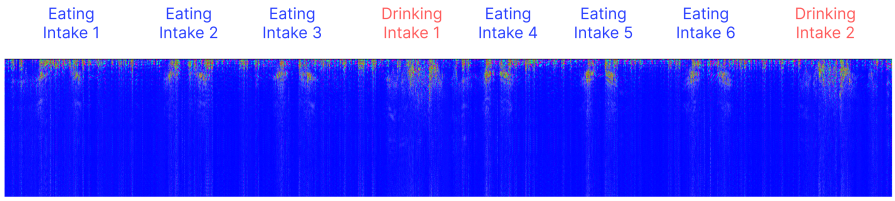


Fig. 16. Differential echo profile for eating and drinking

10.6 Potential Applications

With the promising performance of our system shown in Sec. 7, we discuss several potential applications of our system below.

10.6.1 Eating and Drinking Detection. Since our system is able to track upper body poses accurately, it is very suitable for tracking users' eating and drinking behavior because these activities usually involve the movements of hands and arms. If the users wear our device all day long, it can analyze their daily eating and drinking habits for a long period of time. Moreover, our system has the potential to track more fine-grained information, such as the type of food and drinks that the users are taking. With this information obtained, our system can provide constructive suggestions for users on their eating and drinking habits. For example, it can remind users if they are not drinking enough water every day or if they should have healthier food. In order to verify the feasibility of our system for tracking users' eating and drinking behavior, one researcher performed some eating and drinking activities when wearing our device. The recorded differential echo profile is demonstrated in Fig. 16. As we can see in the figure, there are clear patterns when the researcher is eating and drinking.

10.6.2 Disease Monitoring. Many neurodegenerative diseases, e.g. Parkinson's Disease, incur irregular body movements. The measurement of this kind of irregular movement is usually done at a clinical institution every several months. However, the patients may also need to monitor the progression of their diseases on a daily basis. In this case, it becomes inconvenient for patients to frequently go to the clinical institution. However, with our system, patients can easily record and analyze their body movements on a daily basis. This information is very helpful and valuable for doctors to monitor the progression of their diseases. We leave this for future exploration.

10.7 Model Compression and Latency

We designed a customized deep learning model for estimating pose from the acoustic signal (details in Sec. 5). The model takes both original and differential echo profiles as input to infer upper body pose. The model utilizes 32-bit floating point weight tensors to learn the parameters. In order to deploy the model locally, we reduce the memory footprint of the deep learning model through floating point quantization. In this regard, we adopt a post-training quantization strategy on saved weights of the model and reduce the weight tensors to 8-bit integer precision. This approach reduces the ResNet18 backbone deep learning model size from 46.76 Megabytes (1MB = 10^6 Bytes) to 10.28 MB. For each sliding window that represents 2 seconds of acoustic signal data, the inference time on arm64 CPU with the quantized model is 209.53 milliseconds. Using this quantized model, we can estimate the upper body pose from streaming acoustic data with a latency of 41 milliseconds. This compression approach can be viewed as a proof of concept for local deployment and the inference time serves as a performance upper bound. The deployment is the model in AI-accelerated Micro Controller Units (MCUs) can be further explored in future works.

10.8 Potential Challenges with the Sensing Technique

The sensing system in PoseSonic relies on transmitting FMCW acoustic signal and deciphering the information encoded in the reflection. The challenge of using ultrasonic FMCW signal is that it requires a frequency range (details in Sec. 4) to operate, unlike other signal transmission techniques such as chirps or GSM. The commodity microphones generally have a sampling rate of 50 KHz and thus a Nyquist frequency of 25 KHz. Therefore, the sensing range of PoseSonic should be restricted under 25 KHz to work reliably without any artifact in the received signal. This may create an issue if we extend PoseSonic to a sophisticated version with more transmitters with different frequency ranges. On the other hand, the sweep period of FMCW reflects the sensing range of PoseSonic and it is calibrated for upper body pose estimation. Unless there is a drastic change in the physical structure of the users, the calibration of the FMCW signal will not require any change.

11 LIMITATIONS AND FUTURE WORK

Although PoseSonic offers fine-grained pose estimation with low power, less obtrusive, and privacy-preserving sensing technique, it has some limitations like other wearable devices. In this section, we discuss the limitations and evaluate possible workaround from our perspective.

11.1 Head Motion

PoseSonic is a smartglasses form factor with an acoustic sensing system on both hinges. The sensing technique relies on sound waves reflected on different body parts. In this regard, the reflection varies if the user is having substantial head movement. Since the deep learning pipeline is trained mostly with action units containing static or slight head movement, PoseSonic does not perform well in the scenario of larger head movement. This limitation can be addressed to some extent with the data augmentation technique. We need to include head motions to the dataset and adopt a transfer learning approach to learn those patterns in the reflection patterns obtained through PoseSonic.

11.2 Collection of Ground Truth in the Wild

PoseSonic adopt cross-modal supervision as the learning technique for estimating upper-body human pose. Therefore, it requires a supervisory signal from the synchronized video to train the end-to-end pose estimation model. However, it is difficult to acquire video ground truth in the wild deployment scenario. Hence, there is a bottleneck in improving PoseSonic with in-the-wild video data. This limitation can be addressed with two approaches. First, we can utilize chest-mounted cameras such as GoPro to record the synchronized video. Nevertheless, this approach cannot capture all the upper body joints with this sort of on-body camera installation. And, secondly, we can utilize high-precision Inertial Measurement Units (IMUs) as the source of cross-modal supervision. Although IMUs might not be able to provide information as fine-grained as video frames, it may help in addressing real-life deployment performance.

11.3 Evaluation on Different Styles of Glasses Frames

During the evaluation of our system, we deployed our system on a glasses frame shown in Fig. 3(b). However, we did not evaluate our system on different styles of glass frames. Because we place sensors on the hinges of the glasses, we believe the sensor position is usually similar for most glasses. There might be some special kinds of glasses with very different frame styles, which will cause the sensor position to shift from that on the current glass frame we are using. However, we believe the performance of our system will not be severely impacted as long as the sensors are not blocked and have a good angle of view to emit and receive signals. The performance of our system on different frame styles is worth exploring in future work.

12 CONCLUSION

In this paper, we present PoseSonic which is a low-power and minimally-obtrusive acoustic sensing system on smart glasses. PoseSonic utilizes the sound wave reflection coming from different body parts to infer the upper limb joint coordinates of the human body. For that, we transmit high-frequency acoustic signals and compute the cross-correlation between the received signal and the transmitted signal. This enables PoseSonic to capture the position and movement of upper body joints. Extensive experiments with 22 participants in the lab and naturalistic setting demonstrate that PoseSonic is able to infer upper body posture across different persons and environments. We also present the robustness and generalizability of PoseSonic in the scenario of noise. PoseSonic has the potential to be included as a pose estimation pipeline in future smart glass interfaces. We discuss the possibility and limitations of those aspects here.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2239569. We thank Tianhong Catherine Yu for the valuable feedback on the manuscript. We also appreciate all the participants who took part in the user study.

REFERENCES

[1] 2021. Kinect2 for Windows. https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows

[2] Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. 2021. Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 292, 10 pages. https://doi.org/10.1145/3411764.3445138

[3] Karan Ahuja, Sven Mayer, Mayank Goel, and Chris Harrison. 2021. Pose-on-the-Go: Approximating User Pose with Smartphone Sensor Fusion and Inverse Kinematics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 9, 12 pages. https://doi.org/10.1145/3411764.3445582

[4] Karan Ahuja, Vivian Shen, Cathy Mengying Fang, Nathan Riopelle, Andy Kong, and Chris Harrison. 2022. ControllerPose: Inside-Out Body Capture with VR Controller Cameras. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 108, 13 pages. https://doi.org/10.1145/3491102.3502105

[5] Teo Babic, Florian Perteneder, Harald Reiterer, and Michael Haller. 2020. Simo: Interactions with Distant Displays by Smartphones with Simultaneous Face and World Tracking. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3334480.3382962

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2016. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CoRR* abs/1611.08050 (2016). arXiv:1611.08050 http://arxiv.org/abs/1611.08050

[7] Purves D, Augustine GJ, Fitzpatrick D, et al., and editors. 2001. The Audible Spectrum. In *Neuroscience. 2nd edition.*

[8] Mahmoud El-Gohary and James McNames. 2012. Shoulder and Elbow Joint Angle Tracking With Inertial Sensors. *IEEE Transactions on Biomedical Engineering* 59, 9 (2012), 2635–2641. https://doi.org/10.1109/TBME.2012.2208750

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR* abs/1703.06870 (2017). arXiv:1703.06870 http://arxiv.org/abs/1703.06870

[10] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 770–778.

[11] Ryosuke Hori, Ryo Hachiuma, Hideo Saito, Mariko Isogawa, and Dan Mikami. 2021. Silhouette-Based Synthetic Data Generation For 3D Human Pose Estimation With A Single Wrist-Mounted 360° Camera. In *2021 IEEE International Conference on Image Processing (ICIP)*. 1304–1308. https://doi.org/10.1109/ICIP42928.2021.9506043

[12] Dun-Yu Hsiao, Min Sun, Christy Ballweber, Seth Cooper, and Zoran Popović. 2016. Proactive Sensing for Improving Hand Pose Estimation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2348–2352. https://doi.org/10.1145/2858036.2858587

[13] Dong-Hyun Hwang, Kohei Aso, and Hideki Koike. 2019. MonoEye: Monocular Fisheye Camera-based 3D Human Pose Estimation. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 988–989. https://doi.org/10.1109/VR.2019.8798267

[14] Stephen S. Intille, Ling Bao, Emmanuel Munguia Tapia, and John Rondoni. 2004. Acquiring in Situ Training Data for Context-Aware Ubiquitous Computing Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria)

(CHI '04). Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/985692.985693

[15] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. 2020. Coherent Reconstruction of Multiple Humans from a Single Image. In *CVPR*.

[16] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D Human Pose Construction Using Wifi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (London, United Kingdom) *(MobiCom '20)*. Association for Computing Machinery, New York, NY, USA, Article 23, 14 pages. https://doi.org/10.1145/3372224.3380900

[17] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2017. End-to-end Recovery of Human Shape and Pose. *CoRR* abs/1712.06584 (2017). arXiv:1712.06584 http://arxiv.org/abs/1712.06584

[18] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. M3Track: <U>Mm</U>Wave-Based <U>M</U>Ulti-User 3D Posture Tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services* (Portland, Oregon) *(MobiSys '22)*. Association for Computing Machinery, New York, NY, USA, 491–503. https://doi.org/10.1145/3498361.3538926

[19] Jinjiang Lai and Chengwen Luo. 2021. AcousticPose: Acoustic-Based Human Pose Estimation. In *Wireless Sensor Networks*, Li Cui and Xiaolan Xie (Eds.). Springer Singapore, Singapore, 57–69.

[20] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-Power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. 6, 2, Article 62 (jul 2022), 24 pages. https://doi.org/10.1145/3534621

[21] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. 2021. MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation. https://doi.org/10.48550/ARXIV.2111.12707

[22] Hyunchul Lim, Yaxuan Li, Matthew Dressa, Fang Hu, Jae Hoon Kim, Ruidong Zhang, and Cheng Zhang. 2022. BodyTrak: Inferring Full-Body Poses from Body Silhouettes Using a Miniature Camera on a Wristband. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 154 (sep 2022), 21 pages. https://doi.org/10.1145/3552312

[23] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 143–152.

[24] Saif Mahmud, M. T. H. Tonmoy, Kishor Kumar Bhaumik, A. M. Rahman, M. A. Amin, M. Shoyaib, Muhammad Asif Hossain Khan, and A. Ali. 2020. Human Activity Recognition from Wearable Sensor Data Using Self-Attention. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain*.

[25] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. arXiv:1611.09813 [cs.CV]

[26] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation Using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 529, 12 pages. https://doi.org/10.1145/3544548.3581392

[27] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. 2018. V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] David Baeza Moyano, Daniel Arranz Paraiso, and Roberto Alonso González-Lezcano. 2022. Possible Effects on Health of Ultrasound Exposure, Risk Factors in the Work Environment and Occupational Safety Review. In *Healthcare*, Vol. 10. MDPI, 423.

[29] William Murphy and John Franks. 2002. NIOSH Criteria for a Recommended Standard: Occupational Noise Exposure, Revised Criteria 1998. *Journal of The Acoustical Society of America - J ACOUST SOC AMER* 111 (05 2002), 2397–2397. https://doi.org/10.1121/1.4778162

[30] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1515–1525.

[31] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. 2020. You2Me: Inferring Body Pose in Egocentric Video via First and Second Person Interactions. *CVPR* (2020).

[32] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin P. Murphy. 2017. Towards Accurate Multi-person Pose Estimation in the Wild. *CoRR* abs/1701.01779 (2017). arXiv:1701.01779 http://arxiv.org/abs/1701.01779

[33] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[34] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2022. Winect: 3D Human Pose Tracking for Free-Form Activity Using Commodity WiFi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 176 (dec 2022), 29 pages. https://doi.org/10.1145/3494973

[35] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2022. GoPose: 3D Human Pose Estimation Using WiFi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 69 (jul 2022), 25 pages. https://doi.org/10.1145/3534605

[36] J. Roggendorf, S. Chen, S. Baudrexel, S. van de Loo, C. Seifried, and R. Hilker. 2012. Arm swing asymmetry in Parkinson's disease measured with ultrasound based motion analysis during treadmill gait. *Gait Posture* 35, 1 (2012), 116–120. https://doi.org/10.1016/j.gaitpost.2011.08.020

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2019. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381 [cs.CV]
[38] Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31, 1 (1966), 1–10.
[39] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. 2019. xR-EgoPose: Egocentric 3D Human Pose from an HMD Camera. In *Proceedings of the IEEE International Conference on Computer Vision*. 7728–7738.
[40] Catherine Tong, Shyam A. Tailor, and Nicholas D. Lane. 2020. Are Accelerometers for Activity Recognition a Dead-End?. In *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications* (Austin, TX, USA) *(HotMobile '20)*. Association for Computing Machinery, New York, NY, USA, 39–44. https://doi.org/10.1145/3376897.3377867
[41] M. Tanjid Hasan Tonmoy, Saif Mahmud, A. K. M. Mahbubur Rahman, M. Ashraful Amin, and Amin Ahsan Ali. 2021. Hierarchical Self Attention Based Autoencoder for Open-Set Human Activity Recognition. In *Advances in Knowledge Discovery and Data Mining*. Springer International Publishing, Cham, 351–363.
[42] ultraleap. 2021. leap motion controller. https://www.ultraleap.com/product/leap-motion-controller/
[43] Bastian Wandt, James J. Little, and Helge Rhodin. 2021. ElePose: Unsupervised 3D Human Pose Estimation by Predicting Camera Elevation and Learning Normalizing Flows on 2D Poses. https://doi.org/10.48550/ARXIV.2112.07088
[44] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. 1, 4, Article 170 (jan 2018), 20 pages. https://doi.org/10.1145/3161188
[45] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*. 82–94.
[46] Wentao Xie, Qian Zhang, and Jin Zhang. 2021. Acoustic-Based Upper Facial Action Recognition for Smart Eyewear. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, Vol. 5. 1–28.
[47] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Zhou Tianyi, and Junsong Yuan. 2019. A2J: Anchor-to-Joint Regression Network for 3D Articulated Pose Estimation from a Single Depth Image. In *Proceedings of the IEEE Conference on International Conference on Computer Vision (ICCV)*.
[48] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6184–6193.
[49] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, P. Fua, Hans-Peter Seidel, and Christian Theobalt. 2018. Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera. *IEEE Transactions on Visualization and Computer Graphics* 25 (2018), 2093–2101.
[50] Jackie (Junrui) Yang, Tuochao Chen, Fang Qin, Monica S. Lam, and James A. Landay. 2022. HybridTrak: Adding Full-Body Tracking to VR Using an Off-the-Shelf Webcam. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 348, 13 pages. https://doi.org/10.1145/3491102.3502045
[51] Fukun Yin and Shizhe Zhou. 2020. Accurate Estimation of Body Height From a Single Depth Image via a Four-Stage Developing Network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8264–8273. https://doi.org/10.1109/CVPR42600.2020.00829
[52] Yu Zhan, Fenghai Li, Renliang Weng, and Wongun Choi. 2022. Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization. https://doi.org/10.48550/ARXIV.2203.11471
[53] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. 2019. Semantic Graph Convolutional Networks for 3D Human Pose Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3425–3435.
[54] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-Wall Human Pose Estimation Using Radio Signals. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7356–7365. https://doi.org/10.1109/CVPR.2018.00768
[55] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-Based 3D Skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication* (Budapest, Hungary) *(SIGCOMM '18)*. Association for Computing Machinery, New York, NY, USA, 267–281. https://doi.org/10.1145/3230543.3230579
[56] Weixi Zhao, Yunjie Tian, Qixiang Ye, Jianbin Jiao, and Weiqiang Wang. 2021. GraFormer: Graph Convolution Transformer for 3D Pose Estimation. https://doi.org/10.48550/ARXIV.2109.08364