Ring-a-Pose: A Ring for Continuous Hand Pose Tracking

TIANHONG CATHERINE YU, Cornell University, USA
GUILIN HU, Cornell University, USA
RUIDONG ZHANG, Cornell University, USA
HYUNCHUL LIM, Cornell University, USA
SAIF MAHMUD, Cornell University, USA
CHI-JUNG LEE, Cornell University, USA
KE LI, Cornell University, USA
DEVANSH AGARWAL, Cornell University, USA
SHUYANG NIE, Cornell University, USA
JINSEOK OH, Cornell University, USA
FRANÇOIS GUIMBRETIÈRE, Cornell University, USA
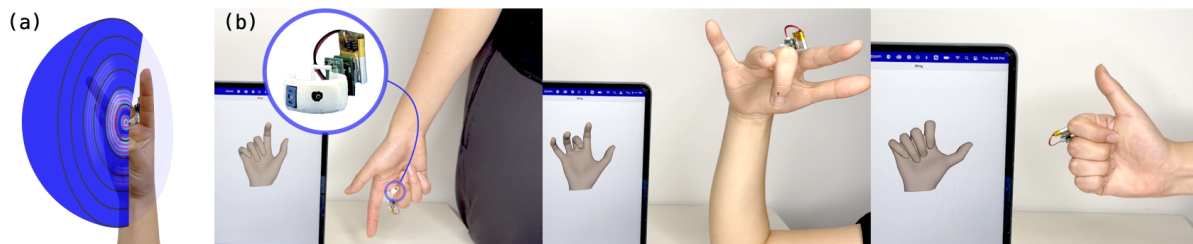CHENG ZHANG, Cornell University, USA

Fig. 1. Ring-a-Pose is an untethered ring that tracks hand poses continuously. With active acoustic sensing, we analyze the reflection strengths at different distances from the ring to reconstruct hand poses. (a) shows an example reflection profile. (b) shows examples of Ring-a-Pose's real-time inferences at different wrist and forearm orientations.

We present Ring-a-Pose, a single untethered ring that tracks continuous 3D hand poses. Located in the center of the hand, the ring emits an inaudible acoustic signal that each hand pose reflects differently. Ring-a-Pose imposes minimal obtrusions on the hand, unlike multi-ring or glove systems. It is not affected by the choice of clothing that may cover wrist-worn systems.

Authors' Contact Information: Tianhong Catherine Yu, ty274@cornell.edu, Cornell University, Ithaca, New York, USA; Guilin Hu, gh386@cornell.edu, Cornell University, Ithaca, New York, USA; Ruidong Zhang, rz379@cornell.edu, Cornell University, Ithaca, New York, USA; Hyunchul Lim, hl2365@cornell.edu, Cornell University, Ithaca, New York, USA; Saif Mahmud, sm2446@cornell.edu, Cornell University, Ithaca, New York, USA; Chi-Jung Lee, cl2358@cornell.edu, Cornell University, Ithaca, New York, USA; Ke Li, kl975@cornell.edu, Cornell University, Ithaca, New York, USA; Devansh Agarwal, da398@cornell.edu, Cornell University, Ithaca, New York, USA; Shuyang Nie, sn429@cornell.edu, Cornell University, Ithaca, New York, USA; Jinseok Oh, jo293@cornell.edu, Cornell University, Ithaca, New York, USA; François Guimbretière, fvg3@cornell.edu, Cornell University, Ithaca, New York, USA; Cheng Zhang, chengzhang@cornell.edu, Cornell University, Ithaca, New York, USA.

In a series of three user studies with a total of 36 participants, we evaluate Ring-a-Pose's performance on pose tracking and micro-finger gesture recognition. Without collecting any training data from a user, Ring-a-Pose tracks continuous hand poses with a joint error of 14.1mm. The joint error decreases to 10.3mm for fine-tuned user-dependent models. Ring-a-Pose recognizes 7-class micro-gestures with a 90.60% and 99.27% accuracy for user-independent and user-dependent models, respectively. Furthermore, the ring exhibits promising performance when worn on any finger. Ring-a-Pose enables the future of smart rings to track and recognize hand poses using relatively low-power acoustic sensing.

1 Introduction

The smart ring is a promising and rising wearable for everyday use [43]. However, due to the spatial constraint, the sensing capability of the ring has been limited: commodity rings integrate miniaturized and low-power sensors (*e.g.*, heart rate sensors, IMUs). Despite being worn on the finger, the existing smart ring, on its own, can not track hand poses continuously, which is crucial in understanding the user's actions and intentions (e.g., VR/AR interaction, human activity recognition). To address this problem, prior work has explored different sensing modalities to recognize hand/finger gestures on a ring [7, 18, 26, 42, 54, 62, 72–74], but multiple rings along with a wristband [79, 80] are required to track full hand poses, not desirable for everyday uses. Continuous hand pose tracking is particularly complex as hand joint movements have 21 degrees of freedom involving finger flexions/extensions and abductions/adductions [79]. The key challenge with a ring is to sense enough information on the pose and movement of each finger from a single ring-worn position, which is challenging to capture using traditional motion sensors like IMUs. Z-Pose explores bio-impedance measurements from a ring for hand pose tracking, but generalizing across sessions and users proves difficult due to inconsistent impedance measurements [63]. In this paper, we seek to answer the research question of **whether we can continuously track hand poses across sessions and users with just a single untethered ring**.

To this end, we present Ring-a-Pose (Fig. 1), a single untethered ring that tracks full hand poses continuously using active acoustic sensing. It only requires a microphone and a speaker embedded into the ring, which already (partially) exist in commodity smart rings (*e.g.*, [43]). Facing inside the palm, the speaker emits inaudible frequency-modulated continuous waves (FMCW) around the fingers, which are reflected by the palm and fingers with unique patterns and received by the co-located microphone. Using the correlation-based frequency modulated continuous wave method (C-FMCW) [66], we obtain the reflection strengths at different echo path lengths from the ring. In addition to capturing the movement of the instrumented finger like IMUs, our ring sensor also captures the movements of other fingers as their movements alter the reflection even when not directly instrumented. Ring-a-Pose is low-cost (~US$30) with a low profile. Furthermore, the battery consumption is relatively low power (148.0 mW) with the possibility of further optimization.

Active acoustic sensing on wearables exhibited exciting potential for capturing fine-grained body part movements from echos [29, 31, 32], but the ring form factor and worn location imposes additional sensing challenges: (a) *limited physical space*: to preserve the ring's low-profile nature, we embed only one pair of microphone and speaker into the ring (differing from multiple pairs of speakers and microphones from prior works), at the cost of reduced sensed information for a complex reconstruction task; and (b) *sensor movement&occlusion caused by finger movements*: the ring form factor has the unique benefit of being "inside" the hand that avoids the orientation disparity between the wrist and the sensor suffered by wrist-mounted devices [12, 28, 29], but the sensor moves along with the worn finger and can be occluded by the fingers. Embracing the challenges in pursuit of a compact yet powerful ring, we develop a lightweight deep-learning pipeline. The unique sensor placement, inside the

hand, allows the ring to (a) capture rich information about the entire hand with just one input channel; and (b) turn finger occlusions into repeatable information for the reconstruction task. We further discuss this in Sec. 3.5.

To thoroughly evaluate the performance of Ring-a-Pose, we conducted three studies. In the first study with 10 participants, we compared the performance when the ring was worn on different fingers. The results showed that Ring-a-Pose can track the hand poses well on any worn finger, which was underexplored in prior ring-based sensing systems [60]. In the second study with 12 participants, we showed that Ring-a-Pose achieves a mean per-joint position error (MPJPE) of 14.1 mm and 10.3 mm in user-independent and user-dependent evaluations, respectively. In the third user study with 10 participants, Ring-a-Pose was able to recognize 7 micro-finger gestures with accuracies of 90.60% and 99.27% in user-independent and user-dependent evaluation respectively.

Unlike most data-driven hand pose tracking systems that have significant performance degradation when the sensor is remounted, Ring-a-Pose is independent of wearing sessions. Additionally, Ring-a-Pose exhibits exciting performance even when used out-of-the-box, i.e. no calibration data from the new user, further highlighting the practicality of ring-based hand pose tracking systems. In summary, the main contributions of this paper are:

- To the best of our knowledge, Ring-a-Pose is the first single-ring system to demonstrate continuous hand pose tracking across sessions and participants: we applied the C-FMCW method to the ring to reconstruct hand poses based on the reflection strengths at different echo path lengths from the ring.
- We validated that the ring effectively tracks the hand pose when worn on any finger with finger-dependent models.
- We evaluated Ring-a-Pose's performance in continuous hand pose tracking and micro-finger gesture recognition with user studies.
- We further discuss the opportunities and challenges of integrating Ring-a-Pose to the future commodity ring platforms.

2 Related Work

Motivated by attractive applications of hand pose tracking in natural interaction, rehabilitation, immersive AR/VR experiences, sign language translation, etc., the computer vision community has explored various robust optical sensing approaches off-the-body: multi-camera systems [61], monocular RGB cameras [38, 75], and depth cameras [47, 58]. However, sensor instrumentation in the environment lacks portability and the occlusion issue limits the users' movements. Thus, there is a need for reliable wearable approaches that are portable and suffer less from occlusions from the environment. Here, we focus on reviewing wearable approaches that are pertinent to our contribution. In the rest of the section, we divide related works into three key areas: (1) hand pose sensing, the functionality of Ring-a-Pose, (2) acoustic sensing, the sensing technique of Ring-a-Pose, and (3) sensing rings, the form factor of Ring-a-Pose.

2.1 Hand Pose Sensing on Wearables

Mounting the sensor directly on the user allows mobile hand sensing. Cameras mounted on the heads are suitable for VR uses [1, 37], but are not practical for day-to-day uses. Wrist is a popular instrumentation site. Wrist-worn devices have low proximity to the hand and do not fully cover the hand like data gloves [10, 15] which are unrealistic to be worn at all times. Cameras mounted on the wrists [22, 68, 70] and other range-finding sensors [12, 24, 29, 54] can reconstruct hand poses or recognize hand gestures from limited viewpoints, but they require a clear line-of-sight from the wrist that cannot be covered by long-sleeve clothing. Other hand pose tracking sensing principles that do not suffer from sensor occlusion include impedance characteristics [28, 63] that do not generalize across users and electromyography (EMG) [34] with a bulky form factor. Our ring is not affected by clothing choices, generalizes well across users, and is low-profile.

Table 1. A High-Level Overview of Hand Sensing Rings. Ring-a-Pose is the first to track continuous hand poses with just a single ring. In the labels, "UI" stands for user-independence and "UD" stands for user-dependence.

System	Tracking Output	Sensor	Power Consumption	# of Components	Un-tethered	UI	UD
ElectroRing, 2021 [27]	Discrete Tap	Electric Field	220mW	1 Ring	✓	×	✓
ThumbTrak, 2018 [54]	Discrete Gesture	Proximity	120mW	1 Ring	×	×	✓
FingerPing, 2018 [74]	Discrete Gesture	Active Acoustic	-	1 Ring, Wristband	×	×	✓
Boldu et al., 2018 [4]	Discrete Gesture	Capacitive	475mW	1 Ring, Wristband	✓	✓	✓
CyclopsRing, 2015 [7]	Discrete Gesture	Camera	-	1 Ring	×	✓	✓
EFRing, 2022 [9]	Discrete Gesture 1D Continuous	Electric Field	-	1 Ring	×	✓	✓
Z-Ring, 2023 [62]	Discrete Gesture 2D Continuous	Bio-Impedance	2.4W	1 Ring, Wristband	×	✓	✓
Zhou et al., 2023 [80]	Continuous Pose	IMU and PPG	44mW/ring	2-5 Rings, Wristband	✓	✓	✓
ssLOTR, 2022 [79]	Continuous Pose	IMU	198mW	2-5 Rings, Wristband	✓	✓	✓
Ring-a-Pose	**Continuous Pose**	**Active Acoustic**	**148mW**	**1 Ring**	✓	✓	✓

Hand gesture recognition is an easier sensing task than hand pose tracking but is crucial to gestural controls. Prior systems of full-hand gesture recognition are implemented with impedance [78], wrist pressure [11], ultrasonic beamforming [24], capacitive sensing [57], etc. To enable more discrete and acceptable natural interactive contorl [21], micro-finger gesture recognition that requires fine-grained sensing is gaining traction [4, 16, 40, 52, 65, 73, 74]. In our work, we demonstrate that Ring-a-Pose effectively recognizes micro-finger gestures.

2.2 Acoustic Sensing on Wearables

On-body acoustic sensing has emerged as a reliable approach for tracking contexts and movements due to its robustness to noise factors like lighting conditions and electric fields. The applications span hand gesture sensing [19, 24, 29, 74], motion tracking [5, 36, 64], pose tracking[35], activity recognition [69], food intake recognition [3], facial expression tracking [31, 32], teeth and tongue gesture recognition[55, 77], gaze tracking [30], silent speech recognition [25, 53, 76, 77], etc.

The most recent work, EchoWrist [29] is the closest to our system. It used two pairs of microphones and speakers on a wristband to track hand poses and recognize hand-object interactions. For a similar sensing task, designing a sensing system for a ring exhibits much greater challenges than for a wristband as the ring has a smaller physical space. For example, a ring can not afford two pairs of microphones and speakers due to spatial constraints and energy consumption considerations. Although both systems employ acoustic sensing, EchoWrist and Ring-a-Pose adapt different sensing principles to infer hand poses: EchoWrist infers hand poses and hand-object interaction from the contour shape around the wrist captured by acoustic sensors, while Ring-a-Pose infers hand poses and recognizes hand gestures from the acoustic reflection from the fingers directly. The ring benefits from its unique worn location. Long-sleeved clothing easily covers the wristband but not the ring. The ring's closer proximity to the fingers allows stronger acoustic reflections from the fingers and enables (a) fine-grained tracking like micro-finger gestures, which is hard to capture information from the wristband[29] and (b) similar tracking performance with only half the number of the speaker and the microphone. Last, unlike EchoWrist, Ring-a-Pose does not necessitate returning to a neutral hand pose between transitioning between poses (further detailed in Sec. 6.1.1). We are the first to explore active acoustic sensing on the form factor of a ring to enable accurate continuous hand pose tracking and micro-finger gesture recognition.

2.3 Sensing Rings

Commodity smart rings today specialize in fitness&wellness [43, 48] and contactless payment [49], missing the hand-related interaction space unveiled uniquely by the ring form factor. Researchers tried to fill the gap with IMUs [13, 59, 79], proximity sensors [54], electric field sensing [9, 56, 62], capacitive sensing [4, 27], electromagnetic sensing [44], infrared sensor [42], minature cameras[7, 39], acoustic signals [18, 71], etc. Sensing rings applications span text input [17, 41, 56], health sensing [46, 80], authentication [51, 62], and gestural inputs [7, 8, 44, 60, 81]. Vatavu and Bilius [60] reviewed hand gesture inputs with rings, ring-like, and ring-ready devices. Ring-a-Pose falls in the "rings" category: finger-worn device with a ring form factor. In past literature, only multi-ring systems [79, 80] can track hand pose continuously. ssLLOR [79] tracks hand poses with a wristband and 2-5 rings, each embedded with an IMU unit. As shown in Table 1, Ring-a-Pose presents the first untethered ring that tracks 20 DoF hand poses with just a single ring across sessions and users, extending practical everyday smart ring capabilities. Note the sensing principle is not the only factor determining energy consumption. Other factors like microcontroller and communication method choices make the direct comparison unfair, but we include the overall power signature to better situate our relatively low-power system. One benefit of our active acoustic sensing approach is that the ring senses all finger movements well (unlike IMUs), no matter which finger the ring is instrumented on. Thumb-In-Motion [4] employs capacitive sensing on an index finger ring to sense thumb-to-index micro gestures. Ring-a-Pose recognizes thumb-to-index micro gestures with the ring on the middle finger and demonstrates the system's potential to detect fine-grain movements even when the moving fingers are not instrumented. Furthermore, it is important to acknowledge that the ring has a limited space for electronics (especially the battery), which makes it challenging to prototype an untethered device.

3 Ring-a-Pose Implementation

The sensing ring, a novel form factor for continuous hand pose tracking, places the sensors on the finger, inside the hand. This affords unique opportunities to simplify the hand reconstruction task: maintaining great visibility of the hand and consistency between the orientations of the wrist and the sensor.

3.1 Design Objective

Our design aims to preserve the slim and miniaturized form factor of a ring for comfortable prolonged wear while achieving promising sensing performance. The complex sensing task that predicts 3D coordinates of 20 hand joints requires information about the thumb and all other 4 fingers. Thus, working with the space constraint, we experiment with only one speaker and one microphone instead of multiple speakers and microphones [24, 29, 32, 77]. Furthermore, to minimize the width of the ring, the speaker and microphone are placed side by side horizontally, instead of vertically.

3.2 Form Factor Design

We started the prototype process with rigid PCB islands but we moved to the flexible printed circuit board (FPCB) for a slimmer ring body and a better fit to the ring curvature. The speaker and the microphone should be as close to each other as possible for accurate round-trip propagation time [66], but due to the ring curvature and thickness of the components, they are placed 5.4mm away from each other. Our ring has a width of 11 mm. Excluding the battery and the microcontroller (MCU), the ring body's thickest part, containing the speaker and microphone, is 3.58 mm thick, and the rest is 2 mm thick. Fig. 2(c) shows a physical mockup of future Ring-a-Pose with replaced arc battery and more minuturized MCU, comparable with that of the commercial Oura Ring [43]. The ring is also lightweight, weighing 4.3 grams including the battery (1.8g). Finally, to fit the finger sizes of all user study participants, we opted for a 3/4 circle ring and used yarn to adjust the ring size with a sliding knot.
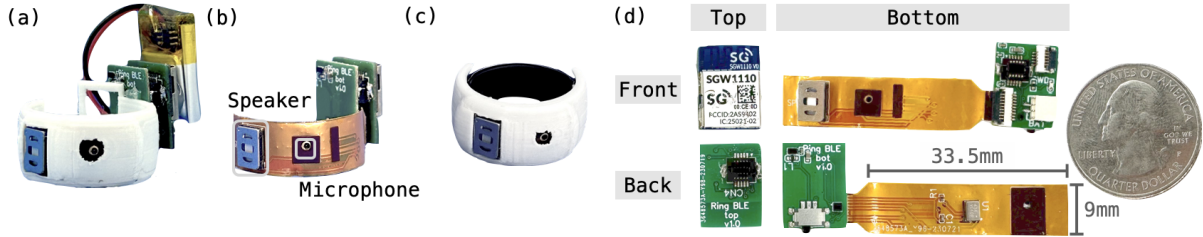
3.3 Hardware

Fig. 2. (a) The untethered Ring-a-Pose prototype used in the user studies. (b) The Ring-a-Pose prototype electronics without the case. (c) A physical mockup of future Ring-a-Pose. We replaced the battery with an arc battery and removed the MCU. (d) Details of Ring-a-Pose's PCBs.

The ring includes (a) a ring body, a customized FPCB enclosed in a 3D-printed PLA case, and (b) a customized capturing and processing PCB. Fig. 2(b) shows the assembled hardware. The top PCB (left in Fig. 2(d)) houses the SGW1110 module, featuring the nRF52840 microcontroller unit (MCU). The low-power MCU implements Bluetooth 5 and provides Bluetooth Low Energy (BLE) functions with a built-in PCB-mounted antenna. The bottom PCB, right in Fig. 2(d), contains an audio amplifier (MAX98357A), a voltage regulator (TPS62743) that provides a constant 3.3V source, a switch that turns ON/OFF the ring, a Flexible Printed Circuit (FPC) connector that connects to the ring body, and a battery connector. The ring body FPCB features a speaker (USound UT-P2019) and a microphone (TDK ICS-43434). The total cost of the prototype is about US$30 and could be decreased when mass manufactured. With the 3.7V 70mAh Lipo Battery, we measured (using a CurrentRanger) the energy consumption of the ring to be 148.0 mW. The energy consumption breaks down into the MCU consuming 24.0 mW, the speaker and the microphone consuming 120.0 mW, and the BLE transmission consuming 4.0 mW on average. We will further discuss the energy consumption in Sec. 8. Similar to [32], the speaker emits FMCW signals in the range of 20-24KHz with 600 samples per period, outside of the commonly stated range of human hearing, and the microphone samples at 50kHz. The 16-bit sampled signals are transmitted to a nearby device (smartphone or computer) using BLE UART communication at 800 Kbps.

3.4 Sensing Principle

We use active acoustic sensing as the sensing method for the ring to track continuous hand poses. The ring emits inaudible sound waves, which are then reflected and refracted by the surrounding fingers and the palm. As a result, the sound waves are received by the microphone with unique patterns. As shown in Fig. 4, different hand geometries yield distinct reflected signals (i.e., echos) which we later process to reconstruct hand poses.

3.4.1 Calculating Echo Profile using C-FMCW. Wang et al. proposed the correlation-based frequency modulated continuous wave method (C-FMCW) which on commodity audio devices, achieves higher ranging resolution, than that of traditional FMCW methods (34.3mm upper bound) [66]. For the C-FMCW method, the theoretical ranging resolution, R_{C-FMCW} is:

$$R_{C-FMCW} = \frac{C \cdot Lag}{2F_s} - vt \tag{1}$$

where C is the speed of sound, $0 \leq Lag \leq 600$ is the number of samples shifted between the transmitted and received signals, and F_s is the sampling frequency (the numerical value is specified in our hardware implementation), v is the velocity of the ranged object, and t is the time from the start of the modulation period,
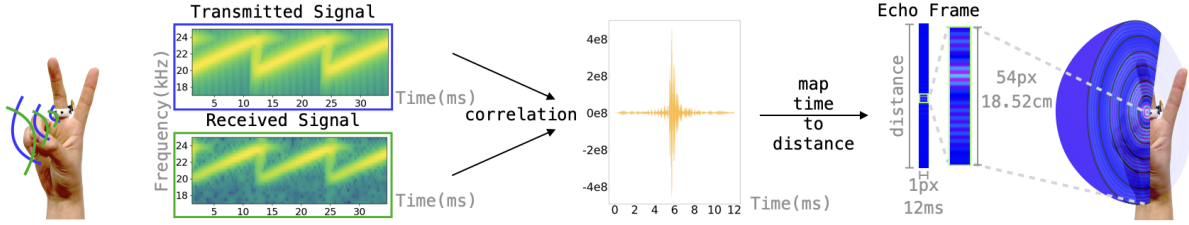
Fig. 3. Echo Frame Calculation. The cross-correlation (orange line) between the transmitted FMCW signal (blue) and the received signal (green) is mapped from the time domain to the distance domain as an echo frame. Ring-a-Pose crops 54 pixels, equivalent to 18.52cm, of the echo frame to analyze hand poses. The black lines in the 3D visualization overlayed on the hand denote a 3cm distance increment. The color of each sphere denotes the summation of reflection strengths from that radius. Note the spectrograms on the left are for visualization purposes, the signals are captured and processed with the time domain.

$0 \le t \le \frac{600}{50000} = 0.012$s. As reasoned in [66], when the ranging target is static or moving slowly R_{C-FMCW} and the theoretical resolution upper bound, δR_{C-FMCW} are:

$$R_{C-FMCW} = \frac{C \cdot Lag}{2F_s} \tag{2}$$

$$\delta R_{C-FMCW} = \frac{C}{2F_s} = \frac{343m/s}{2 \times 50000} = 3.43mm \tag{3}$$

The improved spatial resolution is critical not only for the original ranging purpose but also for capturing geometries as used in Ring-a-Pose. For the hand pose tracking task, the fingers are sometimes static and sometimes moving. When considering the ranged object's velocity (v), we derive R_{C-FMCW} from eq (1):

$$R_{C-FMCW} < \frac{300C}{F_s} - \frac{600v}{F_s} \tag{4}$$

For reference, in everyday hand movements, the human hand joints' median and max movement speed is less than 10 and 320 deg/s [23]. Male middle finger (the longest finger) length averages 10.6cm [45], yielding a v=0.018m/s and v=0.59m/s, which is significantly smaller than the speed of sound. Therefore, the impact of finger movement on the sensing resolution is very small based on this formula. Note, that these numbers are unique to ranging applications and only serve as references for understanding the sensing principle of our system.

EarIO [32], earables tracking facial movements by measuring subtle skin deformations, first utilized the C-FMCW method to generate echo profiles. Instead of identifying the length of the strongest reflection strength path which corresponds to the distance between the sensors and the ranged object, Li et al. used the correlation to capture the reflection strengths of different sound travel path lengths. The calculated echo profiles (detailed below) contain continuous spatial and temporal reflection strengths that denote the complex energy pattern of echos among the different lengths of sound travel paths. As we reasoned above, these echos are mostly created by the shape of the hand with smaller contribution from the dynamic of the movement. The echo profiles (bottom 2 rows in Fig. 4) are made up of 1-pixel wide echo frames (Fig. 3) stacked along the time axis. The original echo profile captures the static hand geometries, and the differential echo profile amplifies the hand geometry movements.

3.4.2 Echo Frame Calculation. As shown in Fig. 3, the ring's speaker emits a Frequency-Modulated Continuous-Wave (FMCW) signal. The blue box visualizes 3 transmitted frequency sweeps on the left in Fig. 3. The green box on the left in Fig. 3 visualizes an example signal received by the microphone. By applying cross-correlation [66]

between the transmitted and 5-order Butterworth band-pass filtered (20-24KHz, the same as emitted signals) received signals, we acquire the strengths of the signals at different return/reflection times (orange line in Fig. 3). The strongest correlation (which we centered at 6ms in the correlation graph, like [32]) at the timestamp t_0 denotes the direct path between the speaker and the microphone. We then map the time domain into the distance domain based on the speed of sound. The hand is small compared to the theoretical sensing range of C-FMCW (2.06m). The echo frame is 1-pixel wide, equivalent to 12ms (the duration of one frequency sweep), and 600-pixel long (the number of samples per sweep), equivalent to 2.06m. The value of each pixel represents the strength of the correlation. To remove reflections from the environment and focus on sensing the hand, we only use 54 pixels starting from t_0, equivalent to 18.52cm, large enough to cover full-hand movements.

3.4.3 Original & Differential Echo Profile Calculation. When we stack echo frames along the time (x-axis), we acquire continuous reflection strengths at different distances. The stacked echo frames are called the original echo profile, as shown in Fig. 4 middle row. To capture the changes between echo frames, we calculate the differential echo profile by subtracting the previous echo frame from the current echo frame. The bottom row in Fig. 4 shows the differential echo profile with dynamic range adjustments for visualization purposes.
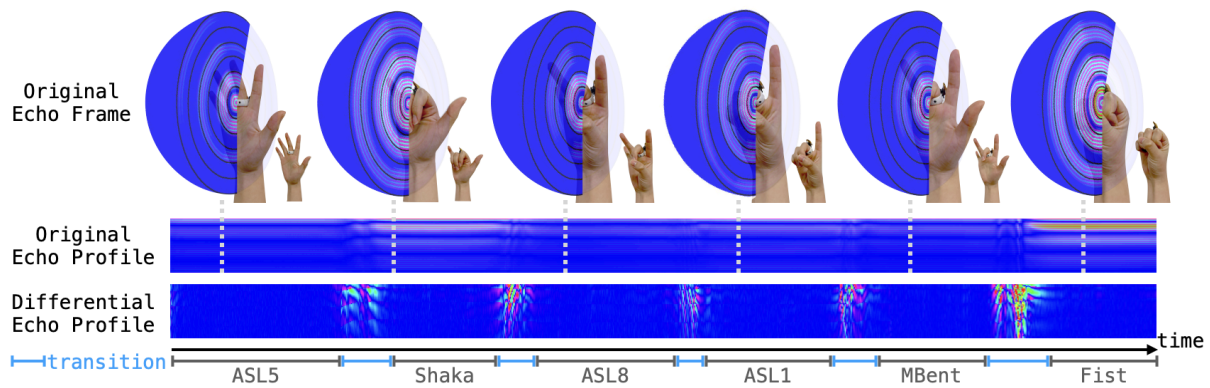
3.5 Echo Profiles for Ring-a-Pose

Fig. 4. Original and Differential Echo Profile for a sequence of Hand Poses. The black lines in the 3D echo frames visualization overlaid on the hand denote a 3cm distance increment. The color of each sphere denotes the summation of reflection strengths from the travel path lengths of the radius of the sphere.

Similar sensing approaches have been used for tracking facial movements from earables [32] and hand poses from a wristband [29], but additional challenges surface when adapting the technique to hand pose tracking with a ring: (1) the limited physical space allows fewer electronics, reducing the amount of sensed information; and (2) the moving fingers can partially block the sensors, occluding the line-of-sight. We experimented with different sizes of microphones and speakers to minimize the obstruction.

Though there are only two input arrays of information, they complement each other efficiently: the original echo profile encodes the positions of fingers related to poses, and the differential echo profile amplifies the movement of hand geometry (between poses). Every finger's position and its movements affect the original and differential echo profile, respectively. Together, they allow us to track 3D hand poses from a single point of instrumentation. With the omnidirectional speaker, we can abstract the echo frame to be half spheres of various radii centered at the ring (Fig. 4). The color of the sphere denotes the total strength of reflection from the travel

path lengths of the radius of the sphere. Note, that the echo frame only contains reflection strengths information from different distances, not exact points. In figure 4, we see that for the "ASL5" pose, the flat hand yields little reflection beyond the inner spheres. "ASL8" and "MBent" are two similar hand poses that only vary in the pose of the thumb, so they share echo frames that are visually similar but distinct to the deep learning model.

Line of sight sensors, that require a clear line between the sensor and the observed object, face the common sensor occlusion issue. Though our active acoustic sensing approach shares the obstacle, we found that the ring affords unique placement that turns hand geometrical occlusions (where fingers fully or partially block the ring) into useful and repeatable information. "Shaka" and "Fist" are two hand poses where the ring is mostly blocked, so they both have very strong reflections. However, interestingly, the slight differences in the thumb and pinky finger positions lead to echo frame differences: "Fist" has stronger reflections than "Shaka". Later in the evaluation, we investigate the effect of sensor occlusion quantitatively.

Different from many prior data-driven hand pose tracking systems[12, 22, 28], one notable strength of our system is that our sensing system is relatively independent of the user (detailed results shown in Sec. 6.3.1), because our sensing systems rely on the multipath echos reflected by the hands, which are largely determined by the anatomical structure of hands. Because human hands are similar anatomically with minor variations in size and shape, they lead to similar echo profiles for the same hand pose across users.

3.6 Deep Learning Pipeline

As described above, echo profiles encode temporal (x-axis) and spatial (y-axis) information of reflection & diffraction strengths. With the calculated echo profiles as inputs, our deep learning pipeline predicts the hand pose/gesture with data augmentation techniques.

3.6.1 Labels for Training/Testing Hand Poses: Ground Truth Acquisition & Normalization. To capture ground truth for training and evaluation, we use MediaPipe Hands [75] that acquire the 3-dimensional (3D) Cartesian coordinates of 21 key points, shown in Fig. 8. We subtract the wrist key point from the other 20 key points to acquire the relative ground truth joint position with respect to the wrist. MediaPipe Hands is also used as the ground truth method in [12, 28] to infer 3D coordinates based on RGB images. We notice the depth prediction of MediaPipe is not precise, but it is consistent under the same lighting conditions. We include the per-axis performance breakdown in the results section for comparison. Compared with a depth camera-based ground truth acquisition method, like Leap Motion [58], MediaPipe tracks occluded joints better. Though the marker-based motion-capturing system is the most accurate, the markers alter the hand geometry and reflection patterns which our acoustic sensing principle relies on.

To account for hand position and orientation differences across sessions and participants, we normalize the hand orientation. For each detected hand pose, we find the plane defined by vectors (a) starting from the wrist (joint 0 in Fig. 8) and ending at the index finger metacarpophalangeal (MCP) joint (joint 5 in Fig. 8), and (b) starting from the wrist (joint 0 in Fig. 8) and ending at the little finger MCP joint (joint 17 in Fig. 8). We then re-align the hand by rotating the palm plane to that of a reference image. Furthermore, to ensure the same hand size for the same participant, we normalize the hand size based on the physical length measured between the wrist (joint 0 in Fig. 8) and ending at the little finger MCP joint (joint 17 in Fig. 8).

The camera (built-in of Apple Macbook Air 2022) we use with MediaPipe samples at 30 fps, and our ring samples at 83 fps, a much higher frequency. We synchronize the ground truth with our sensor signals based on timestamps. To minimize real-time inference lagging, we pick the last sensor reading's corresponding hand coordinates as the ground truth for the echo profile window.

3.6.2 Model Framework. The trained deep learning model takes echo profiles as input and outputs either 3D hand joint coordinates or classification labels, depending on the sensing task. The regression and classification models share similar architectures. We present detailed comparison results of different algorithms in the Sec. 8.2.
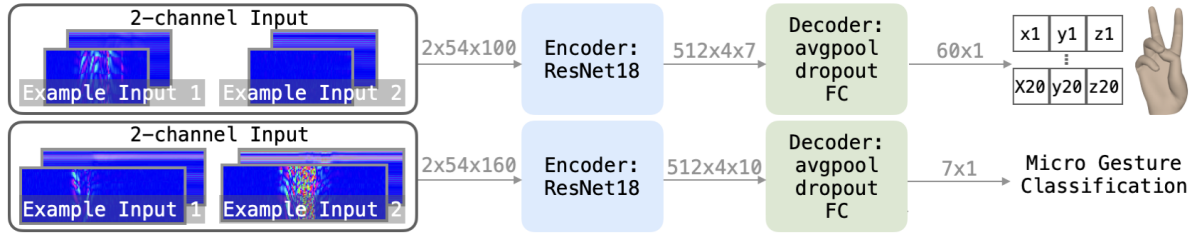
Fig. 5. Encoder-decoder Architecture for Hand Pose Tracking and Gesture Classification. Example visualized inputs have the differential echo profile channel in the front and the original echo profile channel in the back.

Input: The differential echo profile (amplifying hand movements) and original echo profile (capturing static hand poses) are stacked as 2-channel input. The dimensions of the inputs are $[2 \times 54 \times 100]$ and $[2 \times 54 \times 160]$, equivalent to a 1.2-second (100 pixels) window and a 1.92-second (160 pixels) window for the regression and classification models, respectively. Note, for the hand pose tracking study data collection, each pose lasts 2 seconds, so some input instances capture the transition between poses (pose tracking example 1 in Fig. 5), while other inputs capture static poses without movements (example 2 in Fig. 5). Thus, the pose tracking model learns both the static poses and dynamic transitions. For hand gesture classification data collection, each gesture is performed within a 2-second window, slightly longer than the training instance. Every collected gesture has its unique input instance. The input echo profiles are normalized per channel to account for inconsistent magnitudes.

Encoder-Decoder Architecture: We adopt ResNet18 [20] as the encoder backbone that outputs a feature vector of size $[512 \times 4 \times 10]$. The decoder consists of an average pooling layer (pool size=$[1, 1]$), a dropout layer ($p = 0.8$), and a fully connected layer. The final fully connected layer either outputs 20 joints x 3 coordinates = 60 coordinates for the hand pose tracking model, or outputs the predicted label for gesture classification models.

3.6.3 Training Scheme. Our models are implemented in PyTorch and trained on an NVIDIA GeForce RTX 2080 Ti. For the *hand pose tracking model*, we use the Adam Optimizer with a learning rate scheduler starting at 0.002. We train the regression model to predict the 20×3 hand joint coordinates using the mean-square error (MSE) loss. We train the user-independent model with 10 epochs and a batch size of 256. We fine-tune the model with another 15 epochs and a batch size of 32. For the *gestures classification model*, we use the Adam Optimizer with a learning rate scheduler starting at 0.0002. We train the classification model to predict the gesture label using the cross entropy loss. The batch size is 32. We train the user-independent model with 120 epochs and fine-tune the model with another 120 epochs.

3.6.4 Data Augmentation. To improve the model's robustness against worn locations, ring orientations, and hand dimension variations, we apply the following in-place data augmentation techniques, further investigated in Sec. 8.2.6:

- Randomness: Applied in both pose tracking and classification models, random $[-5\%, 5\%]$ increases at 80% chance in each pixel of the echo profile that introduces noise to hand geometry.
- Vertical Shift: Applied in both pose tracking and classification models, vertical shifts account for the different ring positions relative to the hand. All input echo profiles are randomly vertically shifted by ±3 pixels, equivalent to ±10.02mm.
- Horizontal Shift: Applied only in the classification models, horizontal shifts account for inconsistent reaction times that the participants take to start performing the gesture when requested. The input echo profiles are randomly horizontally shifted by $[-13\%, 13\%]$ pixels at 80% chance, equivalent to ±0.15s. Horizontal shifts are not needed for hand pose tracking because the movements are continuous.

Table 2. Statistical details of participants' hand dimensions.

	Middle Finger Ring Size	Hand Length (cm)	Height (cm)	Weight (kg)
mean	6.5	17.9	170.0	62.2
std	1.6	2.0	9.12	10.4
max	10.5	22.0	189.0	93.0
min	4.0	14.0	153.0	44.0

4   Evaluation Overview

To extensively evaluate Ring-a-Pose's continuous hand pose tracking and recognition performance, we conducted three user studies approved by the Institutional Review Board (IRB): (1) **the comparison on the performance of different worn fingers**, detailed in Sec. 5, (2) **continuous hand pose tracking**, detailed in Sec. 6, and (3) **micro gesture recognition**, detailed in Sec. 7.

The first user study aims to assess the **impacts of worn finger choice** on tracking performance. Because the deep learning models are finger-dependent, we selected a finger, the middle finger, for in-depth evaluation. The sensing tasks were separately evaluated to avoid uncomfortably long studies for the participants and each study had a US$15 or US$20 compensation, depending on the study lengths. The goals of the latter two studies are to assess:

- Ring-a-Pose's **stability within a participant** so that as the user removes the ring and puts it back on, no additional training data is needed: We address this by asking the participant to remount the ring before each session;
- Ring-a-Pose's **generalizability across participants** with different hand shapes and hand movement patterns so that no or little training data is needed from a new participant: we address this by evaluating user-independent and fine-tuned models performance; and
- Ring-a-Pose's **robustness to noise factors** for real-world uses: we address this by testing scenarios with various wrist & forearm orientations, sounds, movements, and nearby surfaces/objects.

At the end of each study, the participant completed a survey inquiring about their age, height, weight, and experience with the ring. The researcher also measured the lengths and sizes of their fingers. In total, we conducted a series of 3 studies with a total of 36 different participants (19 self-identified as male, 17 female, mean age=23.5, std age=3.6), including a variety of hand dimensions, detailed in Table 2. Two of the third study participants participated in the first and second studies, respectively. Note the hand length is the measurement between the tip of the middle finger and the center of the wrist.

Overall, across all 3 studies, the participants felt comfortable with wearing the ring (Median=4 on the 5-point Likert scale; 1=very uncomfortable, 5=very comfortable). 46% of the participants could hear the sounds when they performed the gestures, mostly when the hand was near the fist pose, but the sound did not bother them: Median = 4 on the 5-point Likert scale (1=very uncomfortable, 5=very comfortable).

5   User Study 1: Worn Finger Comparison

As an accessory, a ring can be worn on different fingers based on aesthetic preferences and/or to indicate marital status. Thus, when designing a sensing ring for wide adoption, it is crucial to investigate the performance of the ring sensing system on various worn fingers. It's worth noting that the majority of previous single-ring sensing systems only tested the system on one finger [60]. To evaluate Ring-a-Pose's performance on all possible worn fingers, we conducted a user study with 10 participants.
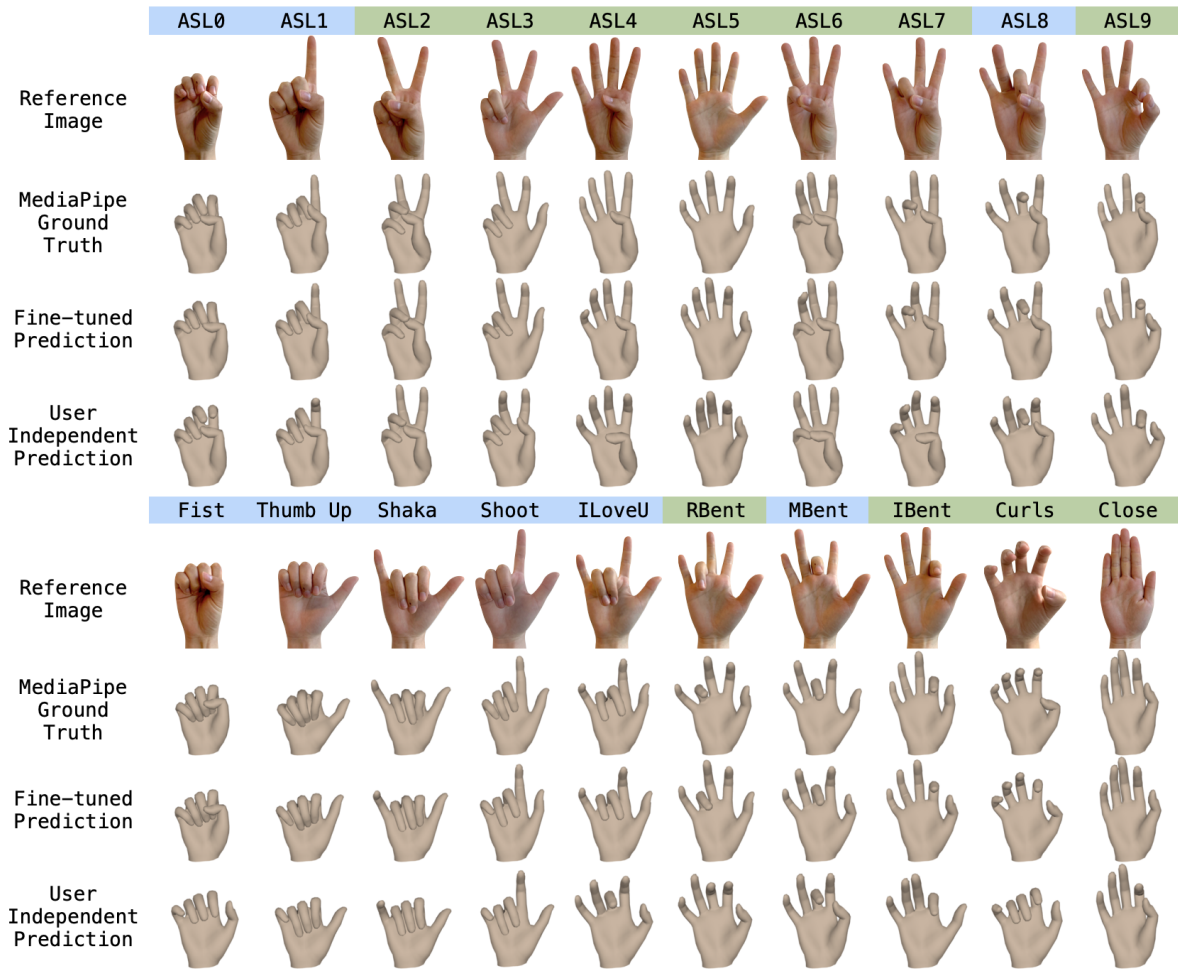
Fig. 6. The twenty terminal hand poses evaluated in our hand pose tracking study. Poses are labeled blue or green based on whether the hand geometries occlude the sensor when the ring is worn on the middle finger. The four rows show (1) reference images displayed during the user study, (2) example MediaPipe ground truths of a participant, visualized using MANO [50], (3) example predictions using the fine-tuned model, and (4) example predictions using the user-independent model. (3) and (4) share the same timestamps as (2).

5.1 Pose Set

Because we tested all five fingers on each participant, to ensure a comfortable study duration, we limited to a small pose set and transitions between poses. We selected (a) 4 isolated individual finger movements: "ASL4" (equivalent to bending the thumb), "RBent", "MBent", and "IBent" in Fig. 6; and (b) 5 compound finger movements: odd ASL digits between 0 and 9. Between each 2 poses, the hand returns to a neutral position ("ASL5").

5.2 Procedure

5.2.1 Study Setup. We conducted a user study with 10 participants. The user study was conducted in an experimental room on a university campus. At the beginning of the study, one researcher explained the study task and the user interface displayed on a laptop monitor. Participants sat in desk chairs during the study with their elbows on the table to ensure that the camera could capture their hands for ground-truth purposes.

A smartphone (Xiaomi Redmi Note 10 Pro) was used to receive and save BLE-transmitted acoustic data from the ring. Participants were instructed to independently wear the ring on their right middle finger before each session, following the experimenter's guidance for alignment. Ground truth data for hand pose was collected using the built-in laptop camera (Apple Macbook Air, 2022) positioned about 55 cm away from the hand on the table and powered by MediaPipe. The experimenter adjusted the camera angle to ensure that the palms stayed parallel to the camera for reliable ground truths. The study lasted about 90 minutes.

5.2.2 Data Collection Sessions. Each participant underwent 21 sessions: 1 practice session followed by 4 sessions for each finger in a randomized order (1 practice + 4 sessions * 5 fingers = 21 sessions). Participants were instructed to independently wear the ring on the selected finger before each session, following the experimenter's guidance for alignment: centering the inner edge of the speaker with respect to the selected finger. Within each session (2.7min), each pose is performed 9 times in a randomized order: $9 \times 9 = 81$ terminal poses. A reference pose image was displayed on the monitor for 2 seconds for each pose as the visual stimuli. Within 2 seconds, the participant's hand leaves the neutral ("ASL5") position to the referenced pose and returns. For each finger, we collect 4 sessions $\times$ 2.7 min = 10.8 min of data containing 4 sessions $\times$ 81 terminal poses = 324 terminal pose instances.

5.3 Results

Evaluation Metrics. We adopt the mean per-joint position error MPJPE to be our quantitative evaluation metric: the mean Euclidean distance errors of all 20 relative (to the wrist) joint positions. Note, MPJPE measures distance errors so it depends on the hand size. An alternative evaluation metric is the mean angular error, but due to MediaPipe Hands' unreliable depth predictions, we chose MPJPE. To account for MPJPE's dependency on the hand size in the user-independent models, we normalize the predicted hand with the participant's physical hand size, as described in Sec. 3.6.1. In real-world uses, it is an additional step a new user needs to do when receiving the device. For fine-tuned user-dependent models, such information is no longer needed as the model quickly learns the hand dimension as shown in the convergence of MPJPE with only 2.67 min of data in Fig. 9(c).

We compare Ring-a-Pose's performance on each finger with user-independent models and the fine-tuned user-dependent models. For each finger, to evaluate the user-independent performance, we use the leave-one-participant-out (LOPO) cross-validation (97.2min of training data from other participants' same finger); and to evaluate the fine-tuned user-dependent performance, we fine-tune the user-independent model with data from the first three sessions (8.1min of fine-tuning data) and test on the last session. In addition to the MPJPE of the entire hand, we further break down the results into the MPJPEs of each finger to analyze whether the worn finger affects the tracking accuracies of individual fingers. Ring-a-Pose has similar performances when worn on all five fingers. The results are detailed in Fig. 7(a). For user-independent models, the errors range from 9.2mm (thumb) to 10.7mm (index finger). For user-dependent models, the errors range from 5.6mm (thumb) to 6.1 mm (middle finger). We break down the individual tracked finger's performance (Fig. 7(b)): the worn finger generally tracks itself better than other fingers do. The multivariate analysis of variance (MANOVA) with model type and worn fingers as independent variables and individual tracked finger errors as dependent variables showed that the worn finger has a significant effect ($F_{5,93}=3.35$, $p=0.008<0.05$) on individual tracked finger errors. We then follow up the MANOVA test with individual ANOVA tests for each tracked finger, we find the thumb ($F_{4,36}=5.76$, $p=0.001<0.05$), ring finger ($F_{4,36}=3.35$, $p=0.020<0.05$), and pinky finger ($F_{4,36}=3.58$, $p=0.015<0.05$) depend on the worn finger, and the other tracked fingers do not. But in most practical cases, the hand is tracked as a whole.
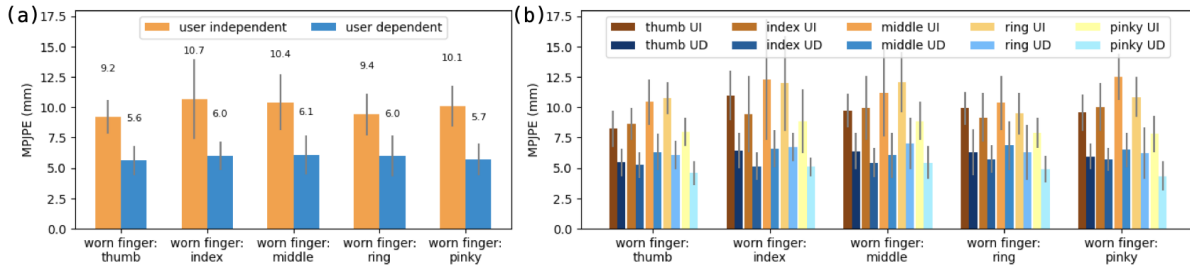
Fig. 7. Worn Finger Comparison Study Results. (a) Ring-a-Pose's tracking performance when worn on different fingers. (b) Tracking performance breakdown on each tracked finger when worn on different fingers. The legend denotes the tracked finger and model type. For example, "thumb UI" and "thumb UD" are the user-independent and the user-dependent tracking errors on the thumb, respectively. Error bars in this figure represent the standard deviation.

A two-way repeated ANOVA test on model types and worn fingers showed that the worn finger factor is not statistically significant ($F_{4,36}$=2.26, p=0.08).

In summary, while Ring-a-Pose has comparable performance across all worn fingers for full-hand tracking, it tracks individual fingers differently based on the worn finger. The result implies that for a general-purpose hand-pose tracking solution, the worn finger location does not matter: unlike most prior works [60], users are free to choose which finger to wear the ring based on their own preferences without performance limitation. However, if the application demands higher tracking accuracies on certain fingers (e.g., thumb and index finger for pinch detections), the worn finger location can alter the performance. We acknowledge that our sample size of 12 (participants) is small for statistical analysis, but in practice, designers and researchers can pick the worn location based on their needs.

6 User Study 2: Continuous Hand Pose Tracking

In this section, we detail the procedures and findings from the continuous hand pose tracking study.

6.1 Pose Set

To capture a wide range of hand poses and movements, we select a set of 20 terminal poses, shown in Fig. 6. Informed by prior works [12, 22, 28, 70], our pose set includes all 10 American Sign Language (ASL) digits evaluated in Opisthenar[70] and all 11 poses evaluated in prior work [28].

6.1.1 Natural Transition between Poses. Just like our first user study, many prior works related to continuous hand pose tracking [22, 33, 68], necessitate that participants revert to a neutral hand pose between transitioning to different poses. However, this requirement deviates from many real-world scenarios where individuals seamlessly move from one pose to another without returning to a neutral position. Tracking poses without the need to return to a neutral pose poses significantly greater challenges for data-driven pose-sensing systems, because it exponentially increases the number of potential hand shapes and poses between two target poses, demanding a much larger volume of training data. We deliberately opted for the more challenging task of instructing users to perform hand poses without reverting to the neutral pose.

6.1.2 The Impact of Sensor Occlusion by Fingers. To investigate the impact of sensor occlusion by the fingers, we divide the pose set based on whether the ring, worn on the middle finger, is partially (*e.g.* "ASL8") or even fully (*e.g.* "Fist") occluded by the hand geometry: (a) NO-poses, no-occlusion poses colored in green, and (b)

WO-poses, with-occlusion poses colored in blue in Fig. 6. Further, we purposely include terminal poses with similar occlusions (*e.g.* "Shaka" v.s. "Thumb Up" and "Shoot" v.s. "ILoveU") to probe Ring-a-Pose's performance with occlusions.

6.2 Study Procedure

6.2.1 Study Setup. We first recruited 12 participants on the university campus for the main study and then recruited another 6 participants for the follow-up study on robustness to nearby surfaces and objects (Sec. 6.3.7). The study setup is similar to that in the first study, except all data are collected with the ring worn on the middle finger.

6.2.2 Data Collection Sessions. Each participant from the main study underwent 25 sessions, and within each session, they were tasked with performing three sets of gestures: 1) all 20 poses; 2) 11 No-poses; and 3) 9 WO-poses. The order of these terminal poses, within each set of poses, was randomized. Between two terminal poses, transition poses are also recorded for evaluation. To assist participants in performing the target poses, a reference pose image was displayed on the monitor for 2 seconds for each pose, accompanied by a progress bar as a visual stimulus. We did not strictly enforce hand pose transition speeds. These 25 sessions can be categorized into three sections:

1) Practice Section (Sessions 1-2, 2.56 min each): The first two sessions were designed as practice sessions to help the participants get familiar with the target poses and the data collection interface. The data from practice sessions were not used in the evaluation.

2) Major Testing Section (Sessions 3-14, 2.56 min each): During the subsequent 12 sessions, the procedure closely mirrored that of the practice sessions. In each of these sessions, each participant performed $2\times(20+11+9)$ = 80 poses. In total, one participant performed $12\times2\times(20+11+9)$ = 960 terminal poses in the major testing section.

3) Robustness Testing Section (Sessions 15-25, 1.28 min each): In the robustness testing section, we sought to evaluate our system under 11 distinct conditions that could potentially affect its performance in real-world settings. Each session was specifically designed to assess one condition. Unlike the previous sessions, the participant only performs the three sets of gestures once per session, leading to $1\times(20+11+9)$ = 40 terminal pose instances per session. These 11 sessions represent the following scenarios:

(1) Pose-Neutral-Pose: Between each prompted pose, the hand returns to a neutral position ("ASL5").
(2) Uncontrolled Hand Movement: Participants moved their hands freely in front of the screen while performing the poses, as long as their hands stayed parallel and inside the camera view for ground truth acquisition.
(3) Environmental Acoustic Noise (Music Playing): Participants played music of their choice using their phone's speaker at a volume they typically listen to music at.
(4) Talking: Participants chose to talk either with the experimenter or monologue throughout the session.
(5) Radial Deviation: Wrist stayed in the radial deviation orientation.
(6) Ulnar Deviation: Wrist stayed in the ulnar deviation orientation.
(7) Flexion: Wrist stayed in the flexion orientation.
(8) Extension: Wrist stayed in the extension orientation.
(9) Neutral: The forearm stayed in the neutral orientation.
(10) Supination: The forearm stayed in the supination orientation.
(11) Hand-Down: The right hand pointed downwards next to the participants' legs.

For the main study, due to a hardware malfunction, 1 participant's data was damaged and this participant was invited to participate in the user study again. In total, we collected 531 mins of pose (each participant 44.25 min x 12 participants) containing 22080 (12 participants x (960+440) = 22080) terminal pose instances. For the follow-up study, there are 10 major testing sessions with the same setup. For the robustness testing sections, the
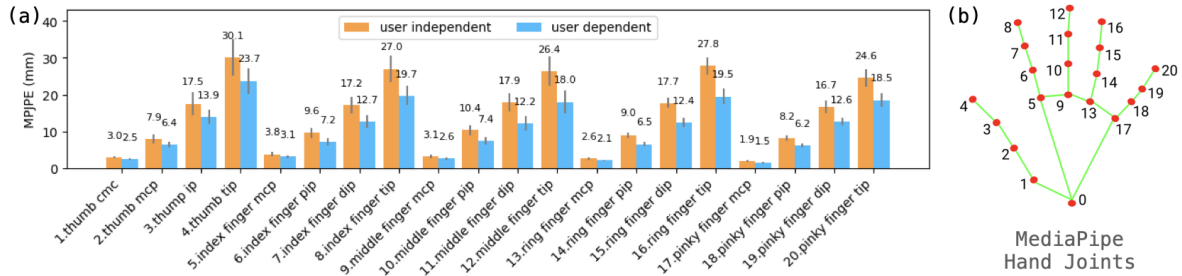
Fig. 8. Joint errors Breakdown. The label indices in (a) map to the mediapipe joint labels in (b) that show MediaPipe hand joint labels. Error bars in this figure represent the standard deviation.

scenarios (Fig. 11) are changed to the following to study the effect of nearby objects/surfaces while ensuring reliable ground truth acquisition:

(1) Acrylic sheet at 36 cm away from the hand.
(2) Acrylic sheet at 36 cm away from the hand with uncontrolled hand movements: Similar to the second scenario in the main study above, the participants moved their hands freely in front of the acrylic sheet. As a result, the distance between the hand and the acrylic sheet varies throughout the session.
(3) Acrylic sheet at 24 cm away from the hand.
(4) Acrylic sheet at 12 cm away from the hand.
(5) Plastic box to the side of the hand.
(6) Box covered with cotton fabric to the side of the hand.
(7) Box covered with copper tape to the side of the hand.
(8) Plushie to the side of the hand.

For the follow-up study, due to a hardware malfunction, 1 participant's data was damaged and we recruited an additional participant.

6.3 Results

The evaluation metric is the same as the first study: mean per-joint position error (MPJPE). For the noise factors, we first fit a linear mixed-effects model with the model type and the robustness testing condition as independent variables, and then we follow with a post-hoc Dunnett's test for comparisons with a control (i.e., the major testing condition). Note we do not evaluate within-session performance because Ring-a-Pose generalizes well with remounting.

6.3.1 User-Independent Performance. We use a leave-one-participant-out (LOPO) cross-validation to simulate when the new user does not provide any calibration data. For each participant, in addition to the data from the other 11 participants (43.97 min x 11 = 8.06h), we also add data collected from 5 researchers (9.96h). In total, 18.02 hours of training data from 16 hands are used.

Overall, we achieved a mean MPJPE of 14.1 mm (SD=5 mm) across 12 participants. The impact of sensor occlusion is shown in Fig. 9(a). The error distribution of joint distance errors is shown as the orange line in Fig. 9(b). The breakdown of the errors in different axes are: x (left/right of the palm) = 5.8 mm, y (up/down of the palm) = 8.0 mm, and z (front/back of the palm) = 6.7 mm. The individual joint error breakdown, orange bars in Fig. 8(a), show that the joints that are further away from the wrist exhibit larger errors, not surprisingly.
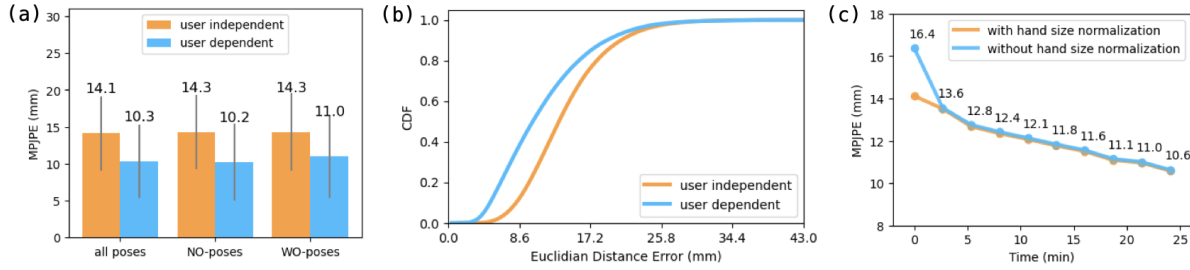
Fig. 9. User Inpendent and User-Dependent Fine-Tuned Hand Pose Tracking Results. (a): Ring-a-Pose's performs well on both poses that fully/partially occlude the sensor (WO-poses) and those that do not (NO-poses). (b): Overall Per-Joint Error distribution. (c): With more training data from a new user, the sensing performance increases. The model quickly learns the physical dimension of the hand in just 2.5 min. Error bars in this figure represent the standard deviation.

6.3.2 Fine-Tuned User-Dependent Performance. Human hands vary in size, shape, and movement patterns across different people. Based on the user-independent models trained above, we fine-tune the models with different amounts of training data from the new participant. Additional training data from a new participant improves the tracking performance, as shown in Fig. 9(c). With 26.7 min of additional fine-tuning data, the overall MPJPE decreases to 10.3 mm, 3.8 mm smaller than the user-independent model. Example hand predictions are included in the 3rd rows of Fig. 6. In Fig. 9(b), the blue line shows the error distribution of joint distance errors. The breakdown of the per-axis errors decrease to x = 4.6 mm, y = 5.8 mm, and z = 4.8 mm, which are 1.23 mm, 2.18 mm, and 1.9 mm smaller than the user-independent model: the improvements are consistent with movement ranges along each axis. The blue bars in Fig. 8(b) show the per-joint improvement: the performance increases more for the joints that are further away from the wrist. As illustrated in Fig. 9(c), the quick MPJPE convergence between with and without normalization shows that the model quickly learns the physical hand dimension with just 2.67 min of data. Further, the slopes in Fig. 9(c) decreases as more data is added, but they do not yet flat out. This indicates that with more fine-tuning data, there is still room for improvement.

6.3.3 The Impact of Sensor Occlusion by Fingers. As mentioned above, to investigate the impact of sensor occlusion by the fingers, we divide the pose sets NO-poses (No occlusion) and WO-poses (fully/partially occlude the sensor) based on whether the ring is partially or fully occluded by the hand geometry. Fig. 9(a) shows the tracking performance on the two pose sets. We performed a two-way repeated ANOVA test, with the model type and occlusion conditions as the independent variables, and did not find statistically significant effects ($F_{1,11}$=1.03, p=0.33) of the pose type on the tracking performance, and there were no interactions ($F_{1,11}$=1.36, p=0.27) between the user- independent/dependent models and NO-/WO- poses.

6.3.4 Robustness to Wrist & Forearm Orientations. A clear benefit of using the ring to track hand poses, compared with wrist-worn devices, is the minimized discrepancy between the hand orientation and the sensor orientation. The user-independent model is trained with both neutral wrist position data and all other wrist orientation data from other users. The user-dependent model is only fine-tuned with data collected in the neutral wrist orientation. We then test the models with data collected from sessions with the 6 extreme orientations as shown in Fig 10(b). The wrist orientation columns in Fig. 10(a) show that, except for supination (t_{253}=4.30, p=0.003<0.05 from the Dunnett's test), the user-independent model performs similarly across all orientations: 0.55mm decrease in performance without statistical significance (p-value range: 0.36-0.99). The user-dependent models have small decreases (mean=2.4 mm) in performance and all but the ulnar wrist orientation (t_{253}=2.52, p=0.10) have statistical significance (p-value range: 0.000-0.001). For the supination orientation, MPJPE has big increases for both the
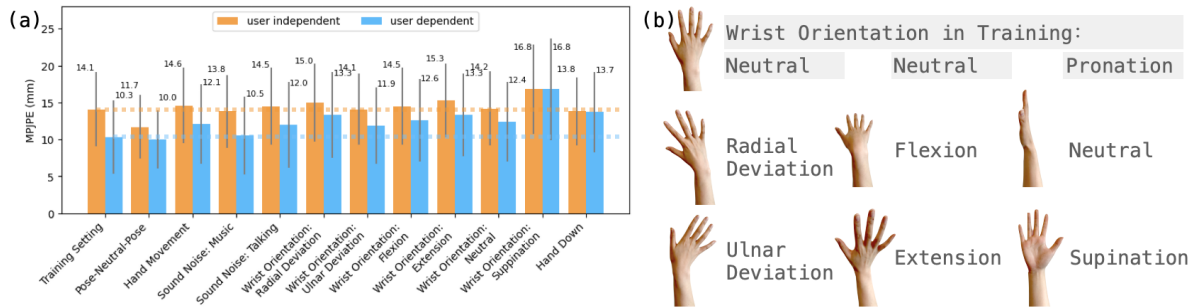
Fig. 10. Ring-a-Pose's Robustness to Noise Factors. (a) Using models trained with data collected in the same setting, the testing results on data collected in various settings show that Ring-a-Pose is robust to many noise factors. (b) Wrist Orientations. Error bars in this figure represent the standard deviation.

user-independent model and the fine-tuned models. During the user studies, this orientation was hard to perform to keep the palm parallel to the screen. In the detected ground truths, we also noticed many frames with visually incorrect hand poses. Thus, the increase in errors may not accurately depict our system's performance.

Furthermore, we tested the performance when the hand is down on the side of the body ("Hand Down" column in Fig. 10(a)), a more natural interaction site. The user-independent model performed similarly (t_{253}=-0.70, p=0.97) to that in the training orientation, but the fine-tuned model (t_{253}=5.73, p<0.001) performance degraded to that of the user-independent model. In summary, we conclude that the wrist&forearm orientations have little impact on the more generic user-independent model, but have a performance degradation for the fine-tuned model.

6.3.5 Robustness to Sounds. Because Ring-a-Pose leverages active acoustic sensing, it is necessary to evaluate the performance with environmental sounds. The two selected sound noise scenarios are visualized in the "Music" and "Talking" columns in Fig 10(a). The mean amplitude levels (in dBFS) in the audible range (20-16000Hz) calculated from the microphone recordings without sound noise, with "Music", and with "Talking" are -54.6, -53.6, and -51.0. For reference, we used a sound level meter app[1] on an iPhone which colocates with the ring. In the experiment room when with only ambient noise, we measured 38 dB(A) with the app and -55.06 dBFS with the ring; when the computer plays music at 68cm, 45cm, and 23cm away from the ring&phone, we measured 47 dB(A)/-55.02 dBFS, 53 dB(A)/-55.02 dBFS, 58 dB(A)/-52.94 dBFS. The measured dBFS change is relatively small as the noise sound source is much further away from the microphone on the ring than the speaker on the ring.

With the participant's choice of music played at their preferred volume, we do not see performance degradation for both the user-independent (t_{253}=-0.81, p=0.94) and -dependent (t_{253}=0.114, p=1) models when compared to the result in the training setting shown in Fig 10(a). However, when the participant talks as they perform the gestures, we see a small degradation without statistical significance: 0.4mm and 1.7mm increase in MPJPEs for the user-dependent (t_{253}=0.39, p=1) and user-independent (t_{253}=2.6, p=0.08) models, respectively. We are unsure about the cause of the different effects brought by music and talking, but regardless, the system still performs well with environmental sounds.

6.3.6 Robustness to Movement. Since people's hands move when they speak, gesture, and interact with spatial interfaces, we evaluate our system under these scenarios. The "Hand Movement" column in Fig. 10(a) shows that the movements slightly harm the performance by 0.5mm for a user-independent model (t_{253}=0.55, p=0.99) but yields an additional 1.8mm increase for the user-dependent model (t_{253}=2.91, p=0.03<0.05).

---

[1]https://www.cdc.gov/niosh/topics/noise/app.html

Fig. 11. Ring-a-Pose's Robustness to Nearby Surfaces and Objects. (a) Using models trained with data collected in the same setting, the testing results on data collected in various settings show Ring-a-Pose's performance with nearby surfaces and objects. (b) Training and Tested Settings. The acrylic sheets are highlighted in light blue. Error bars in this figure represent the standard deviation.

6.3.7 Robustness to Nearby Objects. In addition to finger occlusions, the sensor could be occluded by nearby surfaces and objects(*e.g.*, pants, and desks). Fig. 11(b) shows our tested scenarios with objects placed near the hand, without interfering with our vision-based ground truth acquisition method. Note that the laptop for data collection is placed about 55 cm away from the hand, and we crop the echo profiles at 18.52 cm. With data collected from the six participants in the follow-up study, we evaluate the models' robustness and show the results in Fig. 11(a).

For the user-independent model (trained with 5 other participants' data), we do not find statistical significance (from Dunnett's test) between the training setting and the testing setting: the smallest p-value is 0.94. For the user-dependent model (fine-tuned with 9 sessions of 23min data), the "acrylic@12cm" scenario is significantly different from the training setting (t_{85}=2.80, p=0.04<0.05) with a 1.72 mm increase in error. The next largest performance degradation is "hand movement with acrylic@36cm" scenario with 1.66 mm increase in error with a trend towards significance (t_{85}=2.70, p=0.05). This finding is consistent with the results from "hand movement" without the acrylic sheet in the front, from Sec. 6.3.6. For other scenarios, p-values range from 0.43 to 0.76. Objects directly in front of the hand and within the echo profile sensing range harm the user-dependent performance. For other nearby objects, they have little effect on the tracking performance. Since our sensing range is small, the limitation caused by object occlusion is less prevalent than that from vision-based solutions.

6.4 Takeaways

To situate Ring-a-Pose's continuous hand pose tracking performance with other wearable solutions, we compare its performance with that of prior works in Table 3. Note that this comparison may not be completely fair, as each system used different hardware and different datasets. Because, unlike the vision-based hand-pose tracking systems which use the same benchmark datasets, creating such benchmark datasets is nearly impossible for wearables with customized hardware and varying sensing principles. Nonetheless, the aim of this comparison is to present results and assist readers in better understanding the positioning of our system relative to prior works.

Because Ring-a-Pose is the first single-ring system that tracks continuous hand pose across sessions and users, we compare the performance with multi-ring systems [79], armbands [33] and wristbands [12, 22], in addition to single-ring systems [63]. Although we can directly compare the MPJPEs, they are heavily affected by the pose/gesture sets (detailed in Sec. 6.1.1), the evaluation condition, and the amount of training data.

Table 3. Continuous hand pose tracking performance comparison. In the "Pose/Gesture Set" column, "gestures" refer to when a neutral hand pose is required between transitioning to different poses, and "poses" refer to such neutral pose is not required between transitions. In the error columns, "SD" stands for session dependence, "UD" stands for user dependence, and "UI" stands for user independence. All errors refer to mean per-joint position error (MPJPE).

| System	Devices	Pose/Gesture Set	SD Error (mm)	UD Error (mm)	UI Error (mm)
ssLOTR, 2022 [79]	5 rings and a wristband	free movements	—	6.55	—
WR-Hand, 2021 [33]	an armband	11 gestures + 3 free motions	—	—	18.57
FingerTrak, 2020 [22]	a wristband	19 gestures	12	27.2	—
EchoWrist, 2024 [29]	a wristband	18 gestures	—	4.81	12.2
DiscoBand, 2022 [28]	a wristband	10 poses	11.69	17.87	19.98
Z-pose, 2023 [63]	a ring	10 poses	8.5	—	—
Ring-a-Pose	a ring	20 poses	—	10.3	14.1
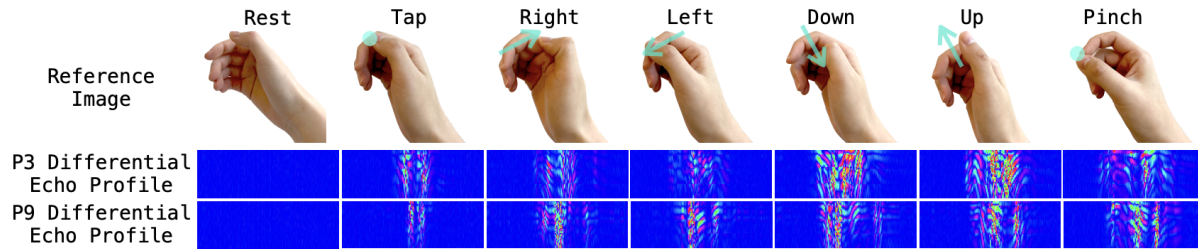
Fig. 12. The seven micro thumb-to-index gestures evaluated in our user study. Top row: reference images similar to those displayed during the user study. Middle row: Example differential echo profiles from participant 3. Bottom row: Example differential echo profiles from participant 9.

Regardless, Ring-a-Pose demonstrates promising session-independent and user-independent performances when compared with other single-device systems. Ring-a-Pose falls short of the multi-device system that contains 5 rings and a wristband but greatly alleviates the burden of wearing a sensing ring on every finger.

Our user-independent model generalizes well across different hand sizes/shapes and is robust to noise factors. The fine-tuned model indeed benefits from user-specific data. Both models show resilience to the noise factors, but compared with the user-independent model, the fine-tuned model is more vulnerable to noise factors. One reason for this is that the user-independent model is trained with some "noise" (using both 12 long sessions and 11 short sessions), but the fine-tuned model is not fine-tuned with "noise" (only using the 12 long sessions), so including more diverse training data for training fine-tuned models will be helpful. In this evaluation, we isolate the "noise" factors for controlled evaluation. Future works on in-the-wild studies will help us understand the system better, but it is challenging to acquire ground truth data in the wild.

7 User Study 3: Thumb-to-Index Micro-Gesture Recognition

The previous showed Ring-a-Pose continuously track the hand poses effectively. In this study, we evaluate Ring-a-Pose's performance on tracking fine-grained micro-gestures which are more discreet, intuitive, and natural [13].
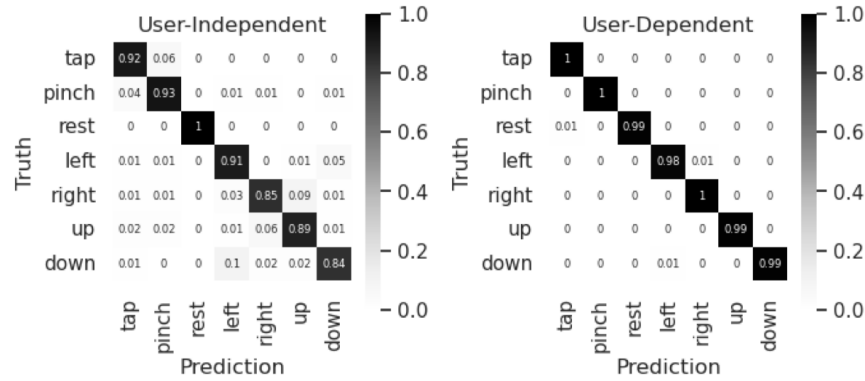
Fig. 13. Micro-Finger Gesture Recognition Study Results. For the 7-class classification task, the user-independent model achieves an accuracy of 90.60%. The user-dependent model achieves an accuracy of 99.27% with 9.3 min of training data.

7.1 Thumb-to-Index Micro-Gesture Set

Among various micro gesture sets [28, 40, 54], we chose thumb-to-index micro-gestures [4, 9, 16] due to their ease of performance and social acceptance [6]. Shown in Fig. 12, we chose 7 gestures, including a "Rest" no-gesture class. Similar to that in the previous study, each gesture started from the "Rest", moved the thumb and index finger, and returned to the "Rest" position.

7.2 Procedure

We conducted a user study with 10 participants, in which 2 participants each participated in 1 of the 2 previous studies. The study lasted about 60 minutes with a similar procedure as the third study. Each participant performed 10 sessions of gestures, and the first session was discarded as a practice session. Before each session, the participants remounted the ring by themselves. From each participant, we collected: 7 gestures ×10 repetitions ×9 sessions = 630 instances. The participants were seated throughout the study with their hands naturally placed on the table.

7.3 Results

Similar to that in the previous study, we use LOPO cross-validation (with 3.15h of training data) to evaluate the user-independent performance for the 7-class classification. Across 10 participants, Ring-a-Pose has a mean accuracy of 90.60% (confusion matrix depicted in left in Fig. 13). We then fine-tune the user-independent models with the participant's data. The mean accuracy quickly increases to 96.6% with 2.3 min of training data and 99.27% with 9.3 min of training data (confusion matrix depicted in right in Fig. 13). We further compare the results with prior ring systems that recognize thumb-index micro-gestures in Table 4. The promising results demonstrate that Ring-a-Pose detects fine-grain movements effectively.

8 Discussion, Limitations and Future Work

The user studies showcase that our system tracks hand positions continuously with resilience to varying noise factors and wrist orientations. The micro-gesture recognition study and the input to AR glasses demo in the video figure serve as example uses of the pose tracking capabilities for practical one-hand interaction. In the future, as the ring's tracking matures, more applications (*e.g.*, sign language translation) will be unlocked. As Ring-a-Pose is the first single-ring system that enables hand pose tracking across sessions and users, it still has shortcomings

Table 4. Thumb-to-index micro-gesture recognition performance comparison. "UD" stands for user dependence and "UI" stands for user independence.

System	Gesture Set	UD Accuracy	UI Accuracy	
Thumb-in-Motion, 2018 [4]	5 gestures	—	89	%F1
Z-ring, 2023 [62]	9 gestures	88 %	83.67%	
EF-ring, 2023 [9]	9 gestures	89.5 %	85.2 %	
Ring-a-Pose	7 gestures	99.27%	90.60%	90.58%F1

Table 5. Real-time Tracking Delay Time Breakdown.

Step	BLE transmission	Echo Profile Calculation	Inference	WiFi Transmission	Hand Visualization
Time (ms)	200.0	14.7	54.6-70.4	16.7	126.9-190.5

that must be resolved before large-scale deployment in real-world settings. For example, Gheran et al. 's attempt to replicate a ring-based gesture elicitation study revealed [14] real-world replicability issues. In this section, we discuss the limitations of Ring-a-Pose and the challenges and opportunities of broader smart ring systems.

8.1 Real-Time Tracking Delay & Performance

Our current system is deployed on a smartphone for real-time inference. However, as the demo video shows, we observe a noticeable delay. Though we already alleviated the delay by using the last frame in the sliding window as ground truth, the transmission and computing times are not negligible. We measured the averaged delay using (a) the ring that captures and sends the acoustic signals to a smartphone via Bluetooth; (b) a smartphone (Redmi Note 12 Pro) that receives the signals, decodes the echo profiles, makes inferences with PyTorch Mobile [2], and sends the prediction to a computer via WiFi; and (c) a computer (Apple Macbook Air, 2022) that receives and processes the predictions for visualization or control purposes.

We report in Table 5 the time for each step. For hand pose tracking, we found a total delay time of 413-492 ms depending on the processing availability of the phone and the computer, in which 126.9-190.5 ms attributed to the hand visualization. The visualization time is long because it uses an inverse kinematic solver and renders a high-fidelity hand in 3-dimensions. The visualization time can be omitted if the joint coordinates are directly used. While the echo-profile calculation and inference times, 69.3-85.1 ms, are short, the BLE transmission takes 200 ms. The selected Android phone has high Bluetooth latency as a result of the tradeoff with reliable high bandwidth. In the future, this delay time can be further reduced with direct hardware control (e.g. nRF52840 Dongle) or advanced BLE hardware/protocols, or compressing data for a much smaller transmission package.

We also observed joint distortions and inconsistent internal torques in the predictions of Ring-a-Pose, stemming from the model lacking information about kinematic constraints. A possible solution involves regressing pose parameters (e.g. MANO [50]) as ground truth, that account for such kinematic constraints. From our preliminary exploration of within-user models, the MANO representation exhibits an angular error of 11.36°, 0.6°smaller than the original MediaPipe representation. We do not compare the distance errors here as they vary in scale. Future research in wearable hand-pose tracking could consider similar approaches or investigate alternative methods for acquiring high-fidelity ground truths, such as multi-camera systems like Quest [37].

---

[2]https://pytorch.org/mobile/home/

8.2 Model, Model Input, and Data Augmentation Selection

Using the user-independent and user-dependent evaluation data collected and similar training schemes in the second user study, we compare our current approach with the alternatives.
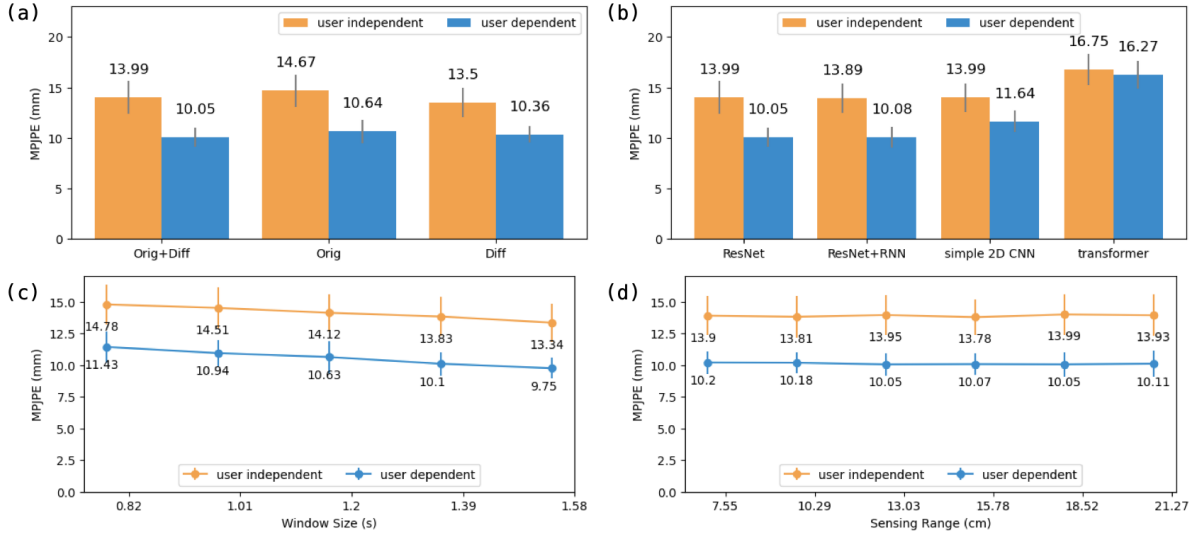

Fig. 14. Comparative studies on the deep learning model and model inputs. (a): Echo profile ablation study: "Orig" denotes the original echo profile and "Diff" denotes the differential echo profile. (b): Model comparison study: Resnet, our current approach, and ResNet+RNN have better performances than simple 2D CNN and transformer. (c) Sliding window length comparison study: a reasonable increase in the window size increases the performance. (d) Sensing range comparison study: sensing range has little impact on the sensing performance.

8.2.1 Model Comparison Study.
In addition to the adopted ResNet backbone model, we experimented with (a) more complex models, such as adding RNN (GRU/LSTM) layers after the ResNet encoder and replacing ResNet with a transformer encoder; and (b) simpler models, like a 3-layer CNN. We show the results in Fig. 8.2(a): ResNet (13.99mm for user-independent, 10.05mm for user-dependent) and ResNet+RNN (13.89mm, 10.08mm) similar performance. The simple 2D CNN has equally good user-independent performance (13.99mm) but much worse user-dependent performance (11.64mm). The transformer model has the worst performance for both independent (16.75mm) and dependent(16.17mm) evaluations. We conjecture ResNet's good performance is a result of echo profiles' efficient 2D representation of time and distance, so the additional temporal learning is unnecessary. ResNet was picked for the smallest errors and the lower computational costs, compared with ResNet+RNN.

8.2.2 Echo Profile Ablation Study.
We used the combination of original and differential echo profiles to capture absolute hand shapes and movements. In this ablation study (Fig. 14(b)), we train and test with only the original or differential echo profiles. The two-way repeated ANOVA test with model type and input types as independent variables further shows statistical significance ($F_{2,22}=65.0$, $p<0.001$). For user-independent models, only the original echo profile (14.67mm) has a larger error than the original+differential echo profiles (13.99mm), but the differential echo profile itself has even smaller errors. This is attributed to the relatively large individual hand shape differences, in comparison to finger movements. However, for the user-dependent model, both the original

Table 6. Data Augmentation and Input Normalization Ablation Study Results. "UD" stands for user dependence, and "UI" stands for user independence. All errors refer to mean per-joint position error (MPJPE).

	UI Training	UD Training	UI Robustness	UD Robustness
w/o data augmentation	14.16	10.67	15.24	12.92
w/o vertical shifts	14.37	10.62	15.25	12.94
w/o randomness	14.07	10.48	14.76	12.68
w/o input norm	14.34	10.43	15.06	12.63
w/ data augmentation w/ input norm	14.12	10.44	14.74	12.70

echo profile (10.64mm) and the differential echo profile (10.36mm) contribute heavily to the prediction. This is because the additional fine-tuned data contains hand shape and movement information tailored to the user.

8.2.3 Window Size Comparison Study. In our current system implementation, we chose a window length of 1.2 seconds. However, the window size can potentially impact the performance. The experiment results on different window sizes are shown in Fig. 14(c), his showed the performance improved as the window size increased: the user-independent and dependent errors decreased from 14.78mm and 11.43mm to 13.34mm and 9.75mm when the window size increased from 0.82s to 1.58s. The two-way repeated ANOVA test with model type and window size as independent variables further shows statistical significance ($F_{4,44}$=9.2, $p<0.001$).

8.2.4 Sensing Range Comparison Study. The selection of sensing range affects the balance between captured noise and information. In our experiment comparing sensing ranges (Fig. 14(d)), we observed minimal impact ($F_{5,55}$=0.79, p=0.56 in two-way repeated ANOVA test with model type and sensing range as independent variables) on performance based on varying sensing ranges: errors of user-independent models range from 13.78mm to 13.99mm, and errors for user-dependent models range from 10.05 and 10.20. Our choice of sensing range was determined by the average hand size.

8.2.5 Input Normalization Ablation Study. As described in Sec. 3.6.2, the original and differential echo profiles are each normalized to account for inconsistent signal magnitudes. Using the same evaluation scheme as the main pose tracking study, we compare the user-independent and -dependent performance when tested with data collected in the same setting as the training data (UI and UD Training in Table 6) and when tested with data collected from the robustness testing sessions (UI and UD Robustness in Table 6). For user-dependent models, input normalization has little effect. Input normalization has larger improvements when signals vary more: 15.06→14.74 for UI robustness with variations across users and settings, 14.34→14.12 for UI testing with variations across users.

8.2.6 Data Augmentation Study. For the pose-tracking task, we apply 2 data augmentation techniques: vertical shifts and randomness, as described in Sec. 3.6.2. We conduct the ablation study in the same manner as the above section by removing individual augmentation techniques and both augmentation techniques. Without any data augmentation, the performance drops for all testing scenarios, proving effectiveness. Similar to the observation in the input normalization ablation study, UI robustness has the largest improvement as it has the largest variations: 15.24→14.74. Introducing randomness on its own has little impact on the performance while applying vertical shifts consistently decreases the errors. Future work on systematically investigating different data augmentation techniques has the potential to improve tracking abilities and increase robustness.

8.3   Future Evaluation in Unconstrained Settings

Our current hand pose tracking study, capturing a wide range of noise factors, evaluates the system in a relatively controlled setting to demonstrate the feasibility of the proposed sensing system. Our pose set of size 20 is relatively large when compared among literature (*e.g.*, 11 in [28] and 16 in [70]), but can not exhaust all possible human hand poses as a research prototype concept. Though we included two uncontrolled hand movement sessions, one with a surface in the front, and one without, in the robustness testing section, the participants kept their hands within the camera view for ground truth acquisition, so the movement does not exactly replicate those in everyday activities. For example, fast global hand movements in walking/running and finger movements may introduce a larger Doppler effect. Our system does not directly analyze the frequency shift caused by the Doppler effect for speed or movement analysis. Instead, our system learns the holistic profile of echoes for pose estimation. A limitation of this approach is that if the distribution of the training data significantly differs from that of the testing data, performance may decline. To address this issue, as with other data-driven approaches, we need to collect additional training data in these specific scenarios to augment the dataset.

We did not constrain the transition speeds in our user studies, but we did not formally evaluate the impact of transition speeds between hand poses with a hand motion speed study [29]. Although our evaluation protocol is similar or comparable to prior work [28, 29]it remains important for future work to evaluate the system's performance in a true freestyle manner in the wild.

Another limitation is that we conducted the user study with adults only, but children and infants have smaller and thinner hands which introduces even more hand geometry variations. It is unclear how our system works with children and infants, beyond the scope of the paper as a proof-of-concept.

Furthermore, in terms of location, our indoor studies do not account for outdoor settings where there could be ultrasonic interference. Because Ring-a-Pose leverages reflection strengths information from within 18.52 cm, objects and other body parts (*e.g.*, legs and hands) within the range alter the signals, as shown in the object robustness study. This is a limitation shared by many existing wearable hand pose sensing systems. Conversely, this "noise" might be seen as information regarding adjacent objects [29] and the hand status. For instance, when an object is detected and moves closer to the hand, the reflected acoustic signal could possess a unique pattern that not only indicates the object's shape but also provides insights into the hand pose. It would be interesting to see other sensing tasks the ring can accomplish, such as object detection and activity recognition.

8.4   Integration into Smart Ring Platforms

Energy Consumption. From the hardware perspective, the simplicity of our sensor setup allows easy integration of the hand pose tracking solution into an existing smart ring. In Fig. 2(c), we show a modification of our prototype with an arc battery (Gepow GRP1507028) and without the microcontroller unit. For instance, to integrate our solution into a commercial product like the Oura Ring [43], which already features a speaker and Bluetooth capability, the addition of just a microphone would be enough. However, there are other obstacles. Though relatively low-power for a wearable hand-pose tracking solution (Table 1), active acoustic sensing, while effective, does not present the absolute best energy efficiency for a ring. The current power consumption of the device still only lasts about 1.75h with our 70mAh Lipo battery due to the (relatively) high energy consumption of our selected flat speaker. This speaker was a tradeoff between a slim form factor and a lower efficiency (we experimented with thicker speakers with much lower energy consumption). The speaker's transmission power could also be lowered to reduce energy consumption. The speaker's current transmitted signal reaches far enough to capture the whole hand's reflection, and our cropped sensing range is much smaller than the theoretical sensing range as described in Sec. 3.4. Our study on robustness to nearby objects further validates that when the object is within the sensing range in front of of the ring, the reflection indeed affects the signals. Future works on alternative hardware and reduced speaker transmission power will make the device more energy efficient.

Multimodal Sensing. In this paper, we solely employed active acoustic sensing. When integrated into a smart ring platform like [43, 80], multimodal sensing, using existing inertial measurement units (IMUs) and Photoplethysmography (PPG), can further enrich Ring-a-Pose's hand sensing capabilities [2, 80]. Other sensing principles will surely complement some of Ring-a-Pose's limitations: electric field sensing may correct the blockage interference within the ring's sensing range [9]; IMUs/capacitive proximity sensor can provide additional movement information on the instrumented finger [79] and on additional two neighboring fingers [67]. Additional sensors bring additional energy consumption and space constraints, posing challenges and opportunities for low-power and miniaturized sensor solutions.

Wearable Ecosystem. As we revisit the research question stated in the introduction, a single untethered ring succeeds in continuously tracking hand poses. However, as an accessory, it is not uncommon for one to wear multiple rings on one hand and even on one finger. When designing gestures for most prior single-ring systems [9, 27, 54], the worn finger is heavily involved in performing the gestures for accurate sensing results, but our micro-gesture study demonstrated that a single ring effectively distinguishes gestures performed by the other fingers, enabling a larger design space for smart ring gestures with fewer constraints. Current multi-ring systems [56, 79] contain a wristband in addition to multiple rings. Building on top of Ring-a-Pose, it would be interesting to explore the design, technical, and interaction space of multiple rings where users choose their number of rings and placements. Further, we envision that in the future, the ring will work together with other wearables as part of the ecosystem to provide users with the optimal always-available interaction experience.

9 Conclusion

In this paper, we present, Ring-a-Pose, the first smart ring that tracks hand poses continuously and recognizes hand postures using inaudible active acoustic sensing. A series of 3 user studies with 36 participants showed that it can achieve 14.1 mm accuracy in inferring the relative (to the wrist) positions of 20 finger joints without the need to collect training data from a new user. The joint error decreases to 10.3 mm with additional calibration data from the user. The promising results pave the way for ring-based hand pose sensing systems.

Acknowledgments

References

[1] Apple. 2023. Apple Vision Pro. https://www.apple.com/apple-vision-pro/.

[2] Apple. 2023. Apple Watch Series 9. https://www.apple.com/shop/buy-watch/apple-watch.

[3] Yin Bi, Mingsong Lv, Chen Song, Wenyao Xu, Nan Guan, and Wang Yi. 2015. AutoDietary: A wearable acoustic sensor system for food intake recognition in daily life. *IEEE Sensors Journal* 16, 3 (2015), 806–816.

[4] Roger Boldu, Alexandru Dancu, Denys JC Matthies, Pablo Gallego Cascón, Shanaka Ransir, and Suranga Nanayakkara. 2018. Thumb-In-Motion: Evaluating Thumb-to-Ring Microgestures for Athletic Activity. In *Proceedings of the 2018 ACM Symposium on Spatial User Interaction*. 150–157.

[5] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li. 2020. Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 95–108.

[6] Edwin Chan, Teddy Seyed, Wolfgang Stuerzlinger, Xing-Dong Yang, and Frank Maurer. 2016. User elicitation on single-hand microgestures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3403–3414.

[7] Liwei Chan, Yi-Ling Chen, Chi-Hao Hsieh, Rong-Hao Liang, and Bing-Yu Chen. 2015. Cyclopsring: Enabling whole-hand and context-aware interactions through a fisheye ring. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 549–556.

[8] Liwei Chan, Rong-Hao Liang, Ming-Chang Tsai, Kai-Yin Cheng, Chao-Huai Su, Mike Y Chen, Wen-Huang Cheng, and Bing-Yu Chen. 2013. FingerPad: private and subtle interaction using fingertips. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 255–260.

[9] Taizhou Chen, Tianpei Li, Xingyu Yang, and Kening Zhu. 2023. EFRing: Enabling Thumb-to-Index-Finger Microgesture Interaction through Electric Field Sensing Using Single Smart Ring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–31.

[10] James Connolly, Joan Condell, Brendan O'Flynn, Javier Torres Sanchez, and Philip Gardiner. 2017. IMU sensor-based electronic goniometric glove for clinical finger movement analysis. *IEEE Sensors Journal* 18, 3 (2017), 1273–1281.

[11] Artem Dementyev and Joseph A. Paradiso. 2014. WristFlex: Low-Power Gesture Input with Wrist-Worn Pressure Sensors. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 161–166. https://doi.org/10.1145/2642918.2647396

[12] Nathan Devrio and Chris Harrison. 2022. DiscoBand: Multiview Depth-Sensing Smartwatch Strap for Hand, Body and Environment Tracking. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[13] Bogdan-Florin Gheran, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2018. Gestures for smart rings: Empirical results, insights, and design implications. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 623–635.

[14] Bogdan-Florin Gheran, Radu-Daniel Vatavu, and Jean Vanderdonckt. 2023. New Insights into User-Defined Smart Ring Gestures with Implications for Gesture Elicitation Studies. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–8.

[15] Oliver Glauser, Shihao Wu, Daniele Panozzo, Otmar Hilliges, and Olga Sorkine-Hornung. 2019. Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Transactions on Graphics (ToG)* 38, 4 (2019), 1–15.

[16] Jun Gong, Yang Zhang, Xia Zhou, and Xing-Dong Yang. 2017. Pyro: Thumb-tip gesture recognition using pyroelectric infrared sensing. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 553–563.

[17] Yizheng Gu, Chun Yu, Zhipeng Li, Zhaoheng Li, Xiaoying Wei, and Yuanchun Shi. 2020. Qwertyring: Text entry on physical surfaces using a ring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–29.

[18] Jeremy Gummeson, Bodhi Priyantha, and Jie Liu. 2014. An energy harvesting wearable ring platform for gestureinput on surfaces. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. 162–175.

[19] Chris Harrison, Desney Tan, and Dan Morris. 2010. Skinput: appropriating the body as an input surface. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–462.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] Masoumehsadat Hosseini, Tjado Ihmels, Ziqian Chen, Marion Koelle, Heiko Müller, and Susanne Boll. 2023. Towards a Consensus Gesture Set: A Survey of Mid-Air Gestures in HCI for Maximized Agreement Across Domains. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–24.

[22] Fang Hu, Peng He, Songlin Xu, Yin Li, and Cheng Zhang. 2020. FingerTrak: Continuous 3D hand pose tracking by deep learning hand silhouettes captured by miniature thermal cameras on wrist. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–24.

[23] James N Ingram, Konrad P Körding, Ian S Howard, and Daniel M Wolpert. 2008. The statistics of natural hand movements. *Experimental brain research* 188 (2008), 223–236.

[24] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2019. BeamBand: Hand gesture sensing with ultrasonic beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–10.

[25] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: "Hearing" Your Silent Speech Commands In Ear. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 57 (jul 2022), 28 pages. https://doi.org/10.1145/3534613

[26] Lei Jing, Zixue Cheng, Yinghui Zhou, Junbo Wang, and Tongjun Huang. 2013. Magic ring: A self-contained gesture input device on finger. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*. 1–4.

[27] Wolf Kienzle, Eric Whitmire, Chris Rittaler, and Hrvoje Benko. 2021. ElectroRing: Subtle Pinch and Touch Detection with a Ring. 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 3, 12 pages. https://doi.org/10.1145/3411764.3445094

[28] Daehwa Kim and Chris Harrison. 2022. EtherPose: Continuous Hand Pose Tracking with Wrist-Worn Antenna Impedance Characteristic Sensing. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–12.

[29] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Deng, Ke Li, Mose Sakashita, Francois Guimbretiere, and Cheng Zhang. 2024. EchoWrist: Continuous Hand Pose Tracking and Hand-Object Interaction Recognition Using Low-Power Active Acoustic Sensing On a Wristband. arXiv:2401.17409 [cs.HC]

[30] Ke Li, Ruidong Zhang, Boao Chen, Siyuan Chen, Sicheng Yin, Saif Mahmud, Qikang Liang, François Guimbretière, and Cheng Zhang. 2024. GazeTrak: Exploring Acoustic-based Eye Tracking on a Glass Frame. *arXiv preprint arXiv:2402.14634* (2024).

[31] Ke Li, Ruidong Zhang, Siyuan Chen, Boao Chen Mose Sakashita, Francois Guimbretiere, and Cheng Zhang. 2024. EyeEcho: Continuous and Low-power Facial Expression Tracking on Glasses. In *To Appear in Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. NA.

[32] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24.

[33] Yang Liu, Chengdong Lin, and Zhenjiang Li. 2021. WR-Hand: Wearable armband can track user's hand. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–27.

[34] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. NeuroPose: 3D hand pose tracking using EMG wearables. In *Proceedings of the Web Conference 2021*. 1471–1482.

[35] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–28.

[36] Wenguang Mao, Jian He, and Lili Qiu. 2016. CAT: High-Precision Acoustic Motion Tracking. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking* (New York City, New York) *(MobiCom '16)*. Association for Computing Machinery, New York, NY, USA, 69–81. https://doi.org/10.1145/2973750.2973755

[37] Meta. 2023. Meta Quest VR Headsets, Accessories & Equipment: Meta Quest. https://www.meta.com/quest/.

[38] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 49–59.

[39] Suranga Nanayakkara, Roy Shilkrot, Kian Peen Yeo, and Pattie Maes. 2013. EyeRing: a finger-worn input device for seamless interactions with our surroundings. In *Proceedings of the 4th Augmented Human International Conference*. 13–20.

[40] Viet Nguyen, Siddharth Rupavatharam, Luyang Liu, Richard Howard, and Marco Gruteser. 2019. HandSense: capacitive coupling-based dynamic, micro finger gesture recognition. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 285–297.

[41] Shahriar Nirjon, Jeremy Gummeson, Dan Gelb, and Kyu-Han Kim. 2015. Typingring: A wearable ring platform for text input. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. 227–239.

[42] Masa Ogata, Yuta Sugiura, Hirotaka Osawa, and Michita Imai. 2012. iRing: intelligent ring using infrared reflection. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 131–136.

[43] OuraRing. 2023. Oura Ring, Oura Health Oy. https://ouraring.com.

[44] Farshid Salemi Parizi, Eric Whitmire, and Shwetak Patel. 2019. Auraring: Precise electromagnetic finger tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–28.

[45] Michael Peters, Kevin Mackenzie, and Pam Bryden. 2002. Finger length and distal finger extent patterns in humans. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* 117, 3 (2002), 209–217.

[46] M Poongodi, Mounir Hamdi, Mohit Malviya, Ashutosh Sharma, Gaurav Dhiman, and S Vimal. 2022. Diagnosis and combating COVID-19 using wearable Oura smart ring with deep learning methods. *Personal and ubiquitous computing* (2022), 1–11.

[47] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. 2014. Realtime and Robust Hand Tracking from Depth. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1106–1113. https://doi.org/10.1109/CVPR.2014.145

[48] Ringly. 2023. Smart Jewelry and Accessories. https://ringly.com/.

[49] RingPay. 2023. RingPay, McLEAR. https://mclear.com/product/payment-ring/.

[50] Javier Romero, Dimitrios Tzionas, and Michael J Black. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610* (2022).

[51] Sougata Sen and David Kotz. 2020. VibeRing: Using vibrations from a smart ring as an out-of-band channel for sharing secret keys. In *Proceedings of the 10th International Conference on the Internet of Things*. 1–8.

[52] Adwait Sharma, Christina Salchow-Hömmen, Vimal Suresh Mollyn, Aditya Shekhar Nittala, Michael A Hedderich, Marion Koelle, Thomas Seel, and Jürgen Steimle. 2022. SparseIMU: Computational Design of Sparse IMU Layouts for Sensing Fine-Grained Finger Microgestures. *ACM Transactions on Computer-Human Interaction* (2022).

[53] Rujia Sun, Xiaohe Zhou, Benjamin Steeper, Ruidong Zhang, Sicheng Yin, Ke Li, Shengzhang Wu, Sam Tilsen, Francois Guimbretiere, and Cheng Zhang. 2023. EchoNose: Sensing Mouth, Breathing and Tongue Gestures inside Oral Cavity using a Non-contact Nose Interface. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers*. 22–26.

[54] Wei Sun, Franklin Mingzhe Li, Congshu Huang, Zhenyu Lei, Benjamin Steeper, Songyun Tao, Feng Tian, and Cheng Zhang. 2021. Thumbtrak: Recognizing micro-finger poses using a ring with proximity sensing. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. 1–9.

[55] Wei Sun, Franklin Mingzhe Li, Benjamin Steeper, Songlin Xu, Feng Tian, and Cheng Zhang. 2021. Teethtap: Recognizing discrete teeth gestures using motion and acoustic sensing on an earpiece. In *26th International Conference on Intelligent User Interfaces*. 161–169.

[56] Ryo Takahashi, Masaaki Fukumoto, Changyo Han, Takuya Sasatani, Yoshiaki Narusue, and Yoshihiro Kawahara. 2020. TelemetRing: A Batteryless and Wireless Ring-Shaped Keyboard Using Passive Inductive Telemetry. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1161–1168. https://doi.org/10.1145/3379337.3415873

[57] Hoang Truong, Shuo Zhang, Ufuk Muncuk, Phuc Nguyen, Nam Bui, Anh Nguyen, Qin Lv, Kaushik Chowdhury, Thang Dinh, and Tam Vu. 2018. Capband: Battery-free successive capacitance sensing wristband for hand gesture recognition. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 54–67.

[58] UltraLeap. 2023. *World-leading Hand Tracking Products: Small. Fast. Accurate. | Ultraleap*. Retrieved Sep 12, 2023 from https://www.ultraleap.com/product/

[59] Radu-Daniel Vatavu. 2023. iFAD gestures: understanding users' gesture input performance with index-finger augmentation devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[60] Radu-Daniel Vatavu and Laura-Bianca Bilius. 2021. GestuRING: A web-based tool for designing gesture input with rings, ring-like, and ring-ready devices. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 710–723.

[61] Vicon. 2023. Vicon | Award Winning Motion Capture Systems. https://www.vicon.com/.

[62] Anandghan Waghmare, Youssef Ben Taleb, Ishan Chatterjee, Arjun Narendra, and Shwetak Patel. 2023. Z-Ring: Single-Point Bio-Impedance Sensing for Gesture, Touch, Object and User Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[63] Anandghan Waghmare, Ishan Chatterjee, and Shwetak Patel. 2023. Z-Pose: Continuous 3D Hand Pose Tracking Using Single-Point Bio-Impedance Sensing on a Ring. In *Proceedings of the 2nd Workshop on Smart Wearable Systems and Applications*. 1–6.

[64] Anran Wang and Shyamnath Gollakota. 2019. Millisonic: Pushing the limits of acoustic motion tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[65] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. 2016. Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 851–860.

[66] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–20.

[67] Mathias Wilhelm, Daniel Krakowczyk, and Sahin Albayrak. 2020. PeriSense: Ring-based multi-finger gesture interaction utilizing capacitive proximity sensing. *Sensors* 20, 14 (2020), 3990.

[68] Erwin Wu, Ye Yuan, Hui-Shyong Yeo, Aaron Quigley, Hideki Koike, and Kris M Kitani. 2020. Back-hand-pose: 3d hand pose estimation for a wrist-worn camera via dorsum deformation network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1147–1160.

[69] Koji Yatani and Khai N Truong. 2012. Bodyscope: a wearable acoustic sensor for activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 341–350.

[70] Hui-Shyong Yeo, Erwin Wu, Juyoung Lee, Aaron Quigley, and Hideki Koike. 2019. Opisthenar: Hand poses and finger tapping recognition by observing back of hand using embedded wrist camera. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 963–971.

[71] Boning Zhang, Yiqiang Chen, Yueliang Qian, and Xiangdong Wang. 2011. A Ring-Shaped Interactive Device for Large Remote Display and Mobile Device Control. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 473–474. https://doi.org/10.1145/2030112.2030177

[72] Cheng Zhang, Anandghan Waghmare, Pranav Kundra, Yiming Pu, Scott Gilliland, Thomas Ploetz, Thad E Starner, Omer T Inan, and Gregory D Abowd. 2017. FingerSound: Recognizing unistroke thumb gestures using a ring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–19.

[73] Cheng Zhang, Xiaoxuan Wang, Anandghan Waghmare, Sumeet Jain, Thomas Ploetz, Omer T Inan, Thad E Starner, and Gregory D Abowd. 2017. FingOrbits: interaction with wearables using synchronized thumb movements. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 62–65.

[74] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Ruichen Meng, Sumeet Jain, Yizeng Han, Xinyu Li, Kenneth Cunefare, Thomas Ploetz, Thad Starner, et al. 2018. FingerPing: Recognizing fine-grained hand poses using active acoustic on-body sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–10.

[75] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214* (2020).

[76] Ruidong Zhang, Hao Chen, Devansh Agarwal, Richard Jin, Ke Li, François Guimbretière, and Cheng Zhang. 2023. HPSpeech: Silent Speech Interface for Commodity Headphones. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers*. 60–65.

[77] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: Continuous Silent Speech Recognition on Minimally-obtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on*

Human Factors in Computing Systems. 1–18.

[78] Yang Zhang and Chris Harrison. 2015. Tomo: Wearable, low-cost electrical impedance tomography for hand gesture recognition. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology.* 167–173.

[79] Hao Zhou, Taiting Lu, Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2022. Learning on the Rings: Self-Supervised 3D Finger Motion Tracking Using Wearable Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–31.

[80] Hao Zhou, Taiting Lu, Yilin Liu, Shijia Zhang, Runze Liu, and Mahanth Gowda. 2023. One Ring to Rule Them All: An Open Source Smartring Platform for Finger Motion Analytics and Healthcare Applications. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation.* 27–38.

[81] Alexandru-Ionuț ȘIEAN. 2022. A Set of Smart Ring Gestures for Drone Control. (2022).